XIV. Magyar Számítógépes Nyelvészeti Konferencia

MSZNY 2018

Szerkesztette:

Vincze Veronika

Szeged, 2018. január 18-19.

ISBN: 978-963-306-578-5

Szerkesztette: Vincze Veronika
vinczev@inf.u-szeged.hu

Felelős kiadó: Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.

Nyomtatta: JATEPress
6722 Szeged, Petőfi Sándor sugárút 30–34.

Szeged, 2018. január

Előszó

2018. január 18-19-én tizennegyedik alkalommal rendezzük meg Szegeden a Magyar Számítógépes Nyelvészeti Konferenciát. A konferencia fő célkitűzése a kezdetek óta állandó: a nyelv- és beszédtechnológia területén végzett legújabb, illetve folyamatban levő kutatások eredményeinek ismertetése és megvitatása, ezen felül lehetőség nyílik különféle hallgatói projektek, illetve ipari alkalmazások bemutatására is.

Nagy örömet jelent számunkra, hogy a hagyományokat követve a konferencia idén is nagyfokú érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A tavaly útjára indult kezdeményezést követve idén is teljes munkák beküldésével lehetett jelentkezni a konferenciára, melyek alapos elbírálása után döntött a programbizottság a cikkek elfogadásáról. A nagy számban beérkezett tudományos cikkek közül idén a programbizottság 24 előadást, 12 poszter-, illetve 4 laptopos bemutatót fogadott el. A programban a magyar számítógépes nyelvészet rendkívül széles skálájáról találhatunk előadásokat a számítógépes szintaxistól kezdve a beszédtechnológián át a többnyelvű alkalmazásokig, mindemellett a poszteres és laptopos bemutatók témái is izgalmasnak ígérkeznek.

Örömünkre szolgál az a tény is, hogy Szarvas György, az Amazon munkatársa elfogadta meghívásunkat, és plenáris előadása is gyarapítja a konferencia résztvevőinek szakmai ismereteit.

Az idei évben is szeretnénk különdíjjal jutalmazni a konferencia legjobb cikkét, mely a legkiemelkedőbb eredményekkel járul hozzá a magyarországi nyelv- és beszédtechnológiai kutatásokhoz. A díj anyagi hátterét a Neumann János Számítógéptudományi Társaság biztosítja, amiért ezúton is hálás köszönetet mondunk. Továbbá szeretnénk megköszönni a programbizottság és a szervezőbizottság minden tagjának áldozatos munkájukat, nélkülük nem jöhetett volna létre a konferencia.

Ács Judit
Novák Attila
Simon Eszter
Sztahó Dávid
Vincze Veronika

Szeged, 2018. január

Tartalomjegyzék

I. Szintaxis

Egy egységesített magyar igei vonzatkerettár építése és felhasználása 3
Vadász Noémi, Kalivoda Ágnes, Indig Balázs

Mazsola - mindenkinek *16*
Sass Bálint

Nulla vagy semmi? Esetegyértelműsítés az ablakban 25
Ligeti-Nagy Noémi, Vadász Noémi, Dömötör Andrea, Indig Balázs

Tanulságok magyar mondatellenőrző nyelvi adatainak átvitelénél 38
Naszódi Mátyás

Conclusions from the Conversion of Linguistic Data of a Hungarian Grammar Checker 51
Naszódi Mátyás

Névutók, előre! Korpuszvezérelt elemzés a névutószerű elemekről ... 52
Ligeti-Nagy Noémi

II. Szemantika, információkinyerés

Magyar szóbeágyazási modellek kézi kiértékelése 67
Novák Attila, Novák Borbála

The World is Built with our Words to Each Other -- Basic and Fine-Tuned Intensional Profiles in Hungarian 78
Szeteli Anna, Alberti Gábor, Kleiber Judit, Dóla Mónika

Egy orosz nyelvű korpusz (NarRu) narratívaelemzése saját fejlesztésű szentiment- és emóciószótárakkal 88
Nyíri Zsófi, Szabó Martina Katalin, Ilyés Virág

Kacsa vagy nem kacsa? Magyar nyelvű álhírek automatikus azonosítása 100
Vincze Veronika

A Rákosi-éra pártjegyzőkönyveinek feldolgozása, elemzése és vizualizációja szövegalapú kapcsolatháló-elemzési módszerekkel 109
Gulyás Attila, Szabó Martina Katalin, Ifj. Boros István, Havadi Gergő

III. Korpusz, alapmodulok

Közös crawlnak is egy korpusz a vége -- Korpuszépítés a CommonCrawl .hu domainjából 125
Indig Balázs

Normo: Egy automatikus normalizáló eszköz középmagyar szövegekhez 135
Vadász Noémi, Simon Eszter

Hyphenation using deep neural networks 146
Németh Gergely Dániel, Ács Judit

Lemmi vagy nem lemmi 159
Novák Attila, Novák Borbála

IV. Beszédtechnológia

Beszédfelismerők mély neuronhálós állapotkapcsolási algoritmusainak kísérleti összehasonlítása 171
Tóth László, Grósz Tamás, Gosztolya Gábor

Televíziós feliratok írásjeleinek visszaállítása rekurrens neurális hálózatokkal 183
Tündik Máté Ákos, Tarján Balázs, Szaszák György

Kísérletek az alapfrekvencia becslésére mély neuronhálós, ultrahang-alapú némabeszéd-interfészekben 196
Grósz Tamás, Tóth László, Gosztolya Gábor, Csapó Tamás Gábor, Markó Alexandra

A diszfónia súlyosságának automatikus becslése, a szakértői értékelések szubjektív jellegének figyelembevételével 206
Tulics Miklós Gábriel, Jászai Henrietta, Vicsi Klára

Az enyhe kognitív zavar és korai Alzheimer-kór automatikus azonosítása spontán beszédből akusztikus jellemzők segítségével ... 219
Gosztolya Gábor, Hoffmann Ildikó, Tóth László, Vincze Veronika, Pákáski Magdolna, Kálmán János

Kriminalisztikai alapú beszélői profilalkotás 231
Beke András

V. Többnyelvűség

Etudes in Chinese-Hungarian Corpus-Based Lexical Acquisition 247
Ugray Gábor

Lexikai erőforrások automatikus előállítása kisebbségi finnugor nyelvekre 260
Simon Eszter, Mittelholcz Iván, Ferenczi Zsanett

Towards cross-lingual utilization of sparse word representations 272
Berend Gábor

Gépi fordítórendszerek kombinálása minőségbecslés segítségével .. 281
Laki László János, Yang Zijian Győző

VI. Poszterek

Evaluation of Universal Dependency parsers for Hungarian 295
Ács Evelin, Recski Gábor

Így írtok ti. Nem sztenderd szövegek hibatípusainak detektálása gépi tanulásos módszerrel 305
Dömötör Andrea, Yang Zijian Győző

A HuTongue spontán beszélt nyelvi korpusz leiratozásának és annotálásának minőségbiztosítási munkálatai 317
Gulyás Attila, Galántai Júlia, Szabó Martina Katalin, Szebeni Zea

"S azóta jól élnek és vidám dalokat énekelnek Leninről és Sztálinról." Szovjet propagandamesék műfaji azonosításának kísérlete 331
Horváth Csilla

Mozgást jelentő igék argumentumszerkezetének korpuszalapú vizsgálata 340
Lázár Bernadett, Szabó Martina Katalin, Vincze Veronika

A szöveg mint skálafüggetlen hálózat 352
Makrai Márton, Sass Bálint

Az érzelmi epizódok narratív konstrukciója és az érzelmi intelligencia érzelemszabályozás komponense közötti kapcsolat 360
Pólya Tibor, Puskás Éva

A magabiztosság-krízis index alkalmazása angol nyelvű megnyilatkozásoknál és rossz minőségű felvételeknél 371
Puskás László, Pólya Tibor

Igei többszavas kifejezések felismerése nyelvfüggetlen módszerekkel 381
Simkó Katalin Ilona, Kovács Viktória, Vincze Veronika

Jogosultság- és kötelezettségfogalmak a KRESZ-ben 393
Szakadát István, Markovich Réka, Hamp Gábor

Relevance Segmentation of Long Documents 405
Szántó Zsolt, Sliz-Nagy Alex, Nagy T. István, Csuma-Kovács Ádám, Vincze Veronika, Farkas Richárd

A szó elszáll, az írás megmarad? Nyelvtechnológiai eszközök a déli manysi nyelvre 413
Szilágyi Norbert, Horváth Csilla, Vincze Veronika, Nagy Ágoston

VII. Laptopos bemutatók

Shtylo: stilometriai elemzések webes támogatása.....423
Dobi Jan Sándor, Mészáros Tamás, Kiss Margit

Természetes nyelvi interfész menetrend- és utazástervező szolgáltatásokhoz 437
Kemény Boldizsár, Recski Gábor

TANIT – Magyar nyelvű szövegeket elemző eszköz összehasonlító digitális bölcsészeti feladatokhoz 450
Labádi Gergely, Farkas Richárd, Nagy Roland, Péter Róbert

Nagyfelbontású pragmaszemantikai igazságértékelés egy játékprogramban.....456
Nőthig László, Szeteli Anna

Szerzői index, névmutató 467

I. Szintaxis

Egy egységesített magyar igei vonzatkerettár építése és felhasználása

Vadász Noémi[1,3], Kalivoda Ágnes[1,3], Indig Balázs[2,3]

[1]Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar
[2]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[3]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
{vadasz.noemi,kalivoda.agnes,indig.balazs}@itk.ppke.hu

Kivonat A cikkben egy egységesített magyar vonzatkeret-adatbázist ismertetünk, amelyet az elérhető magyar vonzatkeret-gyűjtemények összefűzésével és egységesítésével építettünk. A vonzatkerettár felépítésének, használatának, valamint az erőforrásokban talált szisztematikus hibák javításának ismertetése után bemutatunk egy vonzatkeret-egyértelműsítő eljárást, amely a vonzatkerettár egy részén alapul. Végül ismertetjük a modul teljesítményének kiértékelését, összehasonlítva az irodalomból ismert más módszerekkel.
Kulcsszavak: vonzatkeret-adatbázis, vonzatkeret-egyértelműsítés, korpusznyelvészet, mondatelemző

1. Bevezetés

A mondat feldolgozása során az ember számára triviális lépés az ige vonzatainak megtalálása, míg egy számítógépes nyelvi elemző ugyanezt a feladatot egy vonzatkeret-adatbázis segítségével képes végrehajtani. A cikk első felében egy ilyen adatbázist ismertetünk, a Manócskát[1], amelyet az elérhető magyar vonzatkeret-adatbázisok összefűzésével és egységesítésével készítettünk. A cikk második felében bemutatjuk, hogyan hasznosítottuk a Manócska infinitívuszi vonzatokat tartalmazó vonzatkereteit az AnaGramma [1] nyelvi elemző rendszer vonzatkeret-egyértelműsítő moduljában.

2. A Manócska

Az általunk épített vonzatkeret-adatbázis, a Manócska a jelenleg ingyenesen hozzáférhető magyar vonzatkerettárak összekapcsolásával és harmonizálásával készült, így a legbővebb szabadon elérhető, ilyen jellegű erőforrás magyar nyelvre. A felhasznált erőforrások harmonizálásakor törekedtünk a korántsem teljes és precíz, különböző felépítésű erőforrások gyengeségeinek javítására is. A következő fejezetben a Manócska építéséhez használt erőforrásokat mutatjuk be. Ezután ismertetjük az adatbázis felépítését és használatát, végül pedig sorra vesszük az erőforrásokból származó hibás bemenetek javítási módszereit is.

[1] https://github.com/ppke-nlpg/manocska

2.1. A felhasznált erőforrások

A Manócska vonzatkerettárába integrált erőforrások mindegyike korpuszból nyert adatokon alapul és automatikus módszerekkel készült, amelyet néhányuknál kézi ellenőrzés is követett. Sass Bálint a leggyakoribb magyar vonzatokat és szókapcsolatokat tartalmazó, nyomtatásban megjelent szótára [2] a Magyar Nemzeti Szövegtár első, központozás nélkül 187 millió tokent tartalmazó verziójából (MNSZ) [3] készült egy automatikus vonzatkeret-kinyerő módszerrel [4]. A papírszótár anyagát kézzel ellenőrizték, ugyanakkor a nyersanyagául szolgáló tételek manuálisan nem ellenőrzött, bővebb gyűjteménye [5] – amely 28 millió elemzett mondatot és félmillió igei szerkezetet tartalmaz – elérhető adatbázis formájában is.

A Mazsola igei argumentum böngésző platform[2] [6] vonzatkeret-adatbázisa [5] sokáig a magyar vonzatkeretek egyetlen elérhető statisztikai adatbázisa volt. A Mazsolában megjelenő tételekből generált vonzatkeretek és példamondatok képezik a papírszótár bővebb verzióját. Maga az adatbázis csak később vált elérhetővé, és a szótár nehezebben volt hozzáférhető a platformnál, ezért mindhárom forma (platform, szótár, adatbázis) röviden Mazsola néven is használatos, a szövegkörnyezet által meghatározott módon, egyértelműen behelyettesítve.

A Tádé vonzatkeret-gyakorisági listáját a Webkorpusz [7] „4%-ából" (amely központozással együtt 589 millió token) spektrális klaszterezéssel [8] nyerték ki [9]. A gyakorisági lista az opcionális vonzatok keresésének egyik melléktermé̇ke. Méretében az eddig bemutatott erőforrásoknál jóval nagyobb, de ez köszönhető annak is, hogy sokkal nagyobb osztályt vizsgáltak, amely magában foglalta az infinitívuszt vonzó igék vonzatkereteit is.

A felhasznált erőforrások közül érdemes a Mazsolát és a Tádét összevetni, hiszen különböző módszerrel és különböző elvek mentén készültek. A Mazsola készítésekor inkább a vonzatkeret-adatbázis pontossága lehetett a fő cél, tehát az, hogy ne tartalmazzon hibás vonzatkereteket. Ezzel szemben a Tádé készítői inkább a nagyobb fedésre törekedhettek. Ennek megfelelően a Mazsola alkalmasabbnak tűnik olyan feladatokra, amelyekhez lexikai erőforrásként szükség van egy vonzatkerettárra, ugyanakkor a mérete jóval kisebb. A 4. táblázatban látható, hogy a Tádé az igék száma szerint kicsit több mint háromszorosa a félmillió igei szerkezetet tartalmazó, legbővebb Mazsola adatbázisnak, a keretek számában viszont alig van különbség a Tádé javára. Egy másik nagyon fontos eltérés a két erőforrás között az, hogy míg a Mazsola egyáltalán nem tárolja az infinitívuszi vonzatokat, a Tádé igen – igaz, nem kezeli őket megfelelően (lásd a 2.4. fejezetet).

Két további, korpuszból kinyert listát is beépítettünk a Manócska adatbázisába. Az első az MNSZ 2.0.3 [10] 785 millió tokent (írásjelek nélkül) tartalmazó anyagából automatikusan kinyert gyakorisági lista[3] [11], amely 27 091 igekötő-ige párt tartalmaz. Az igekötő-ige lista készítésénél a legnagyobb pontosság volt a cél, így a lista manuálisan ellenőrzött és javított. Ugyanakkor az

[2] `http://corpus.nytud.hu/mazsola/`

[3] `https://github.com/kagnes/hungarian_verbal_complex`

igekötő-kapcsolódás produktivitása miatt a lista fedése soha nem lehet teljes. A második az infinitívuszi vonzattal rendelkező igék gyakorisági listája[4]. Ez az MNSZ 2.0.4 verzióján [10] készült, amely 1,04 milliárd tokent tartalmaz (írásjelek nélkül). Összesen 1 507 ige szerepel benne, az infinitívuszi vonzatuk gyakoriságával, típusával, a korpuszból vett egy-egy példamondattal, valamint a potenciális lexikális vonzatokra vonatkozó információval. Akárcsak az első listánál, itt is a pontosság volt az elsődleges szempont, ezért ez a lista is manuálisan ellenőrzött.

2.2. Az erőforrások egységesítése

Az erőforrások összefűzése és egységesítése során elvégeztünk néhány hibajavító eljárást is, amelyek általunk definiált szabályok automatikus alkalmazását jelentik. Ezek a legegyszerűbb, triviális javításoktól a komplexebb megoldásokig terjednek. A módosítások – egy kivétellel – a szótár bal oldalát, tehát az igét és a hozzá tartozó igekötőt érintik. A javításokat részletes ismertetésük után a 2. táblázat foglalja össze.

Triviális hibajavításnak nevezzük a helytelen igealakok kiszűrését a vonzatkerettárból. Feltételezzük, hogy ezek a tövesítő hibájából eredően kerülhettek be a felhasznált erőforrásokba. Kiszűrésükhöz minden ige esetében a Humor számítógépes morfológiát [12] használtuk. Ha a Humor az igéhez „[IGE]” elemzést (is) adott, akkor meghagytuk, ha nem, akkor töröltük a teljes keretet az adatbázisból.

Az erőforrások vizsgálatakor azt találtuk, hogy egy-egy ige gyakran olyan igekötővel szerepel egy keretben, amely valójában nem az övé. Az ilyen hibákat fontosnak tartottuk javítani, hiszen ismert, hogy az igekötőnek nagy szerepe van a vonzatkeret-egyértelműsítésben: ha igekötős az ige, akkor a lehetséges vonzatkeretek száma jelentősen csökkenhet [13]. A hibák kiküszöbölésére egy szabályalapú megoldást javaslunk. Az igekötős igék esetében – ahol az igekötő és az ige egy *függőleges vonal* (pipe, |) karakterrel vannak elválasztva – mind az igekötőt, mind az igét megvizsgáltuk a Humor segítségével. Ha sem az igekötő, sem az ige nem kapott megfelelő elemzést, akkor a két tokent összevonva megnéztük, hogy a Humor [IGE] elemzést adott-e vissza. Ha igen, akkor az adatbázisban is összevontuk a két tokent egyetlen, igekötő nélküli igévé.

Nem mindig egyértelmű, hogy igekötős-e az ige. A döntéseink során arra támaszkodtunk, hogy míg a semleges mondatokban általában egybeírjuk az igekötőt az igével, a nemsemleges mondatokban speciális szórendben szerepelnek. Két intuitív tesztet alkalmaztunk: (1) a finit igekötős ige tagadásakor a szórend többnyire: tagadószó, finit ige, igekötő, valamint (2) az igekötős ige *is*-sel történő módosítása esetén a szórend: igekötő, *is*, ige. Az 1. táblázat illusztrálja a két tesztet.

További nehézséget okoz, hogy a lexikális vonzatok, az igemódosítók és az igekötők kategóriái között nem húzódik éles határvonal, a három kategória kontinuumot képez. Sok esetben az egybeírás/különírás sem segít, hiszen a problémás szavak helyesírása a korpuszban sem következetes. Az egybeírást/különírást így

[4] `https://github.com/kagnes/infinitival_constructions`

	igekötő	nem igekötő
szó	***ki**folyik*	***fel**vételizik*
tagadás-teszt	*nem folyik **ki***	*nem **fel**vételizik*
„is”-teszt	***ki** is folyik*	-

1. táblázat. Két teszt annak eldöntésére, hogy igekötős-e az ige. Az igekötő-gyanús szavakat **félkövérrel** szedtük.

az „igekötőség” mértékének tekintettük: minél több esetben írják egybe az igével, annál közelebb áll a kérdéses szó a prototipikus igekötőkhöz. Megszámoltuk az egybeírt és a különírt változatok gyakoriságát, és ha az egybeírás gyakorisága elérte az összes előfordulás 15%-át, akkor igekötőként kezeltük. Például az MNSZ 2.0.4-ben a *helyt* szó 1 396-szor szerepel különírva az ige előtt, és 6 548-szor egyeírva, így igekötőként kezeltük.

A 2. táblázat a triviális hibákat, a lehetséges okokat és a megoldásukat ismerteti. Az első három oszlopban az látható, hogy a Humor milyen szófaji címkét adott az igekötőre és az igére külön, valamint a két elemet egybevonva. A Humor által adott szófaji címkék alapján (1) töröltünk olyan kereteket, amelyekben az igealak semmilyen átalakítással nem volt elfogadható, (2) átalakítottuk a hibásnak talált igealakokat.

	Humor				
igekötő	ige	egyben	példa	hiba	javítás
-	-	-	*nyug*	tövesítés	törlés
-	-	[FN]	*12-15-ért*	szófaji egyértelműsítés	törlés
[IK]	[IGE]	[IGE]	*ki/abál*	szegmentálás	*kiabál*
[IK]	-	[IGE]	*alá/z*	szegmentálás	*aláz*
-	[IGE]	-	*kölcsön/sikerül*	szegmentálás	*sikerül*
[IK]	[IGE]	-	*meg/akar*	szegmentálás	*akar*
[IK]	[IGE]	[IK][IGE]	*ráró*	szegmentálás	*rá/ró*
[IK]	[IGE]	-	*össze/tud*	szegmentálás	*tud*
[IK]	[IK][IGE]	-	*abba/bele/fut*	szegmentálás	*bele/fut*

2. táblázat. Triviális hibák és esetleges okaik. A hibák kombinálódhatnak is.

A fenti hibajavító eljárások a vonzatkerettár bejegyzéseinek bal oldalát javítják, tehát azt az oldalt, amely az igét és az esetleges igekötőjét tartalmazza. Ezen felül a bejegyzések jobb oldalán is elvégeztünk egy triviális javítást. A több azonos vonzatból csak egyet őriztünk meg, azt feltételezve, hogy ezek tévesen kerültek az erőforrásokba (valószínűleg mellérendelő szerkezetek vagy felsorolások következtében, a vonzatkeretek korpuszból történő kinyerése során). A

javítóeljárásokban megfogalmazott szabályok az adatbázissal együtt, forráskód formájában elérhetőek.

2.3. A Manócska felépítése

A Manócska egy `tsv` fájl formájában használható, amely az igék vonzatkereteit tartalmazza az egyes erőforrásokban talált gyakoriságukkal együtt. A 3. táblázat sorai a vonzatkerettár egy vonzatkeretéhez tartozó információt mutatják.

***abba/hagy* [ACC]**	a keret gyakorisága	összes keret összgyakorisága
Magyar Igei Szerkezetek (szótár)	3 622	6 117 057
Félmillió igei szerkezet	3 771	18 303 463
Tádé	397	8 867 536
igekötő-ige lista	11 528	13 715 465
infinitívusz lista	0	1 507
rang	0.00168342	

3. táblázat. A Manócska egy vonzatkerete. A vonzatkerettár oszlopait itt sorokba rendeztük.

Az első oszlop maga az ige (az igekötőjével együtt). A vonzatkeret második oszlopa (a vonzatokat tartalmazó oszlop) felépítése a következő: (1) ha az igének semmilyen vonzata nincs, azt egy „@” karakter jelöli, (2) a főnévi vonzatokat az esetragjuk vagy a névutójuk képviseli. A lexikális vonzatok az esetragon vagy a névutón kívül magát a lexikális tövet is tartalmazhatják. A felhasznált erőforrások mindegyike korpuszadatok alapján készült, így feltételezhetjük, hogy a viszonylag gyakori vonzatkeretek szerepelnek bennük[5].

A következő oszlopokban az egyes kereteknek a felhasznált erőforrásokban talált gyakorisága szerepel. A legutolsó oszlopban a rang, egy összesített gyakorisági szám látható, amely a kerethez az egyes erőforrásokra kiszámított normalizált gyakoriságok összege[6]. A rang alapján rendezett vonzatkeret-szótárban a legtöbb adatbázisban megtalálható és gyakori keretek kerülnek előre – a normalizálás miatt az adatbázis méretétől függetlenül – az egyes igékhez tartozó vonzatkeretek közül.

A Manócska adatbázisa az eredeti erőforrásokból az összefűző, egységesítő és hibajavító szkript futtatásával bármikor reprodukálható[7]. A reprodukálhatóság megőrzése fontos célunk, hiszen így biztosíthatjuk a további hibák javításá-

[5] A kutatás jelenlegi fázisában nem foglalkozunk a vonzatok kötelezőségével vagy opcionalitásával, illetve a szabad határozókról sem hozunk ítéletet.

[6] A normalizált gyakoriság a vonzatkeret gyakorisága elosztva az összes vonzatkeret gyakoriságának összegével.

[7] Jelenleg a reprodukáláshoz szükség van néhány kiinduló erőforrás beszerzésére – amelyeket megfelelő formátumban a megfelelő mappákba kell elhelyezni a működés-

nak lehetőségét. A MANÓCSKA reprodukálhatóságával kiemelkedik a felhasznált erőforrások közül.

2.4. Az infinitívuszi vonzatok a MANÓCSKÁBAN

A vonzatkerettár építésekor fontosnak tartottuk az infinitívuszi vonzatok megfelelő kezelését, ezért a MANÓCSKA ezeket is tárolja. Amint a 2.1. fejezetben már utaltunk rá, a felhasznált erőforrások közül egyedül a TÁDÉ tartalmazza az infinitívuszi vonzatokat, azonban ez az erőforrás sem kezeli őket megfelelően. Ha egy ige vonzatkeretében infinitívuszi vonzat van, akkor az infinitívuszi vonzatot a tövével együtt (tehát lexikális vonzatként) sorolja fel a mondatban található többi főnévi vonzattal együtt. Ez azt jelenti, hogy a keretben összekeveredhetnek a finit és infinit ige saját vonzatai. Ráadásul a TÁDÉ leírásában[8] azt olvashatjuk, hogy ha a finit igének van infinitívuszi vonzata, akkor a vonzatkeret többi eleme az infinitívuszhoz tartozik. Ez azonban nem mindig igaz (pl. a *megtanít*, *megérez*, *megvár* tárgyesetű vonzatainál), valamint az összes infinitívuszi vonzat lexikális vonzatként történő kezelése sem megfelelő.

Az infinitívuszi vonzatok kezelésére a következő javaslatot tesszük: egyelőre nem próbáljuk meg szétválogatni az infinitívuszi vonzatot is tartalmazó keretek többi vonzatát aszerint, hogy a finit vagy az infinit igéhez tartoznak-e, hanem az infinitívuszt tartalmazó keretekből csak az infinitívuszi vonzatokat tároljuk (a többi vonzatot töröljük). Azt állítjuk, hogy az infinitívuszok töve az ő vonzat tulajdonságuk szempontjából irreleváns, ezért az infinitívuszi vonzatokhoz tartozó tövet sem tároljuk (az egyes tövek gyakoriságát természetesen megőrizzük, és a tövek gyakoriságát összegezzük az infinitívuszt vonzó ige keretének gyakoriságában). Feltételezzük, hogy minden igető, amely infinitívuszi vonzatként előfordul, szerepel finit igeként is a vonzatkeret-adatbázisban az ő többi saját vonzatával együtt. Például a TÁDÉ keretei között talált *alá/bukik INF_ meg/nedvesíteni* keretet átalakítjuk *alá/bukik INF* keretté, ami azt jelenti, hogy az *alábukik* igének egy bármely tövű infinitívuszi vonzata lehet (pl. *úszni*, *búvárkodni*, *gyöngyöt halászni*, *megnedvesíteni* stb.).

A TÁDÉ infinitívuszi vonzatai mellett használtuk a már bemutatott gyakorisági listát is, amely az infinitívuszt vonzó igéket (és igekötőjüket) tartalmazza. A MANÓCSKÁBAN az infinitívuszt tartalmazó keretek tehát két forrásból származnak, így a forrásoknak megfelelő gyakoriságuk is szerepel az adatbázisban.

Ezt az infinitívuszt vonzó ige-igekötő listát használtuk a 3. fejezetében ismertetett vonzatkeret-egyértelműsítő eljáráshoz.

2.5. A MANÓCSKA számokban

Az erőforrások eredeti méretét összevetettük a javítóeljárások alkalmazása után kapott méretükkel. Az eredményeket a 4. táblázat mutatja. Feltételezzük, hogy

hez –, mivel azok licencbeli korlátozások miatt nem szerepelhetnek a MANÓCSKA repozitóriumban.

[8] `https://hlt.bme.hu/en/resources/tade`

a 2.2. fejezetben ismertetett, a vonzatkerettár bal oldalán (az igéken) végzett szabályalapú javításaink nem eredményeztek hibás kereteket. A táblázaton látható, hogy az egyes erőforrások méretéből különböző százalékban vágott le a javítóeljárások alkalmazása. Ennek az az oka, hogy az erőforrásokat különböző korpuszokon és különböző technikákkal állították elő.

erőforrás	igék száma			keretek száma		
	eredeti	javított	(%)	eredeti	javított	(%)
Igei szerkezetek	2 226	2 185	1,84	6 266	6 203	1,00
Félmillió...	9 999	9 629	3,70	535 607	527 059	1,59
Tádé	30 263	29 269	3,28	864 281	549 753	36,39
ige-igekötő	27 091	27 091	0,00			
ige-infinitívusz	1 507	1 507	0,00			
ÖSSZESEN	31 455	30 119	4,25	1 242 354	972 524	21,72

4. táblázat. A javítóeljárások alkalmazása előtti és utáni keretszámok az egyes erőforrásokhoz, valamint a különbség mértéke %-ban feltüntetve.

Az 5. táblázat azt mutatja, hogy az egyes erőforrásokban talált keretek hány százalékban voltak megfeleltethetőek egymásnak. Feltételezzük, hogy minél több erőforrásban szerepel egy-egy vonzatkeret, annál valószínűbb, hogy helyes keretről van szó.

típus	keretszám	erőforrások száma	arány (%)
összes keret	972 524	5	100,00
összes infinitívusz	2 772	2	0,29
csak infinitívusz	729	2	26,30
csak infinitívusz	2 043	1	73,7
összes nem infinitívusz	969 752	4	99,71
csak nem infinitívusz	1 734	4	0,18
csak nem infinitívusz	47 900	3	4,93
csak nem infinitívusz	377 123	2	38,89
csak nem infinitívusz	54 2995	1	56

5. táblázat. A keretek számának eloszlása aszerint, hogy hány erőforrásban szerepelnek. Az arányszámok esetében az infinitívuszi kereteket az összes infinitívuszos keret számához, a nem infinitívuszi kereteket az összes, infinitívuszt nem tartalmazó keret számához viszonyítottuk.

3. Az infinitívuszi vonzatok szerepe a vonzatkeret-egyértelműsítésben

A fejezet a [13] által ismertetett vonzatkeret-egyértelműsítő eljárás, a VFrame alkalmazását mutatja be az AnaGramma nyelvi elemző rendszer kereteiben. Az AnaGramma mint pszicholingvisztikailag motivált elemző az emberi mondatfeldolgozás kétfázisú működését modellálja [14], amelynek első fázisában történik meg a vonzatkeret-egyértelműsítés. Ezután az igei komplexum készen áll arra, hogy a mondatelemzés második fázisában kielégíthesse az ige vonzatigényét. A kétfázisú mondatelemzés első fázisát a balról jobbra és szavanként történő elemzésben egy +2 token méretű előretekintő elemzési ablak valósítja meg.

Az igék, igekötők és infinitívuszok korpuszban mérhető viselkedése alapján előállított VFrame algoritmus képes arra, hogy az ige szűk kontextusa ismeretében a felmerülő összes lehetséges vonzatkeret közül kizárja a nem aktuális kereteket [13]. Ebben a vonzatkeret-egyértelműsítő algoritmusban az igekötők és az infinitívuszi vonzatok segítik a nem aktuális vonzatkeretek kizárását. Az ismertetett eljárást az igekötők és az infinitívuszok viselkedésének alapos megfigyelésével és korpuszmérésekkel támasztottuk alá, ugyanakkor az eljárás kiértékelése a cikkben nem történt meg.

A VFrame eljárás működésének alapfeltétele egy szótár, amelyben az igékhez rendelt igekötők tárolódnak azzal az információval együtt, hogy az ige-igekötő pár vonzatkeretei között szerepel-e olyan, amely infinitívuszi vonzatot tartalmaz. Ez a szótár a Manócska felhasználásval készült, pontosabban annak az ige-igekötő és az ige-infinitívusz listájából. Az 1. ábra két bejegyzést mutat ebből a szótárból[9].

```
utál {?, el:X, ki:X, meg:X}
felejt {?, el:?, ki:X, le:X, ott:X, rajta:X}
```

1. ábra: Két bejegyzés a VFrame által használt szótárból. Az igekötőhöz rendelt *?* azt jelenti, hogy azzal az igekötővel együtt az igének lehet infinitívuszi vonzata, az X azt jelenti, hogy nem. Az igéhez rendelt *?* azt jelenti, hogy az igének igekötő nélkül lehet infinitívuszi vonzata.

Ennek a szótárnak a segítségével ki tudjuk értékelni az eljárás teljesítményét. A VFrame algoritmusának ismertetése után kiértékeljük az eljárást[10].

[9] Az 1. ábrán szereplő *megutál* igének elvileg lehet infinitívuszi vonzata, de az MNSZ 2.0.4 korpuszban mindössze egy példa volt erre: *már megutáltam folyton hasznos lenni*. Az ebből előállított gyakorisági lista ötös gyakoriságnál kezdődik, így néhány ritka eset nem került be az erőforrásba.

[10] Az implementációnk elérhetősége: `https://github.com/ppke-nlpg/vframe`

3.1. A VFrame algoritmusa

A mondatelemzés során az igék elemzésénél[11] elindul a VFrame eljárás, amely leírható a 2. ábrán látható automata formájában. Az eljárás három keresőből áll, amelyek sorban a következők: (1) infinitívuszkereső balra, (2) infinitívusz- vagy igekötőkereső az ablakban, majd (3) igekötőkereső balra. A három keresés célja az, hogy a lehető legjobban lerövidüljön az ige elemzésekor felmerülő lehetséges vonzatkeretek listája.

Az automatának két belépési pontja van. Az olyan igék esetében, amelyekhez a szótárban – bármilyen igekötővel vagy igekötő nélkül – van infinitívuszi vonzat, az első keresőeljárás indul el. Ekkor az elemzett igét megelőző mondatszakaszban esetlegesen talált infinitívusz – feltételezve hogy az ige vonzata lesz – az igéhez esetlegesen tartozó igekötők listáját megszorítja azokra az igekötőkre, amelyekkel párosulva az igének lehet infinitívuszi vonzata. A második keresőre már ezzel a megszorított igekötőlistával lép tovább. Ha nem volt infinitívusz az elemzett igét megelőző mondatrészben, akkor a második keresés következik.

A második keresésnél az igét követő néhány elem vizsgálata történik aszerint, hogy van-e közöttük igekötő vagy infinitívusz. Találat esetén a talált elemmel megszorításra kerül az igéhez tartozó ige-infinitívusz lista.

A harmadik keresés ismét az elemzett ige előtti mondatrészt vizsgálja, amelyben igekötőt keres. Ha korábban volt infinitívusz-találat, amely megszorította az igekötőlistát, akkor csak olyan találatot fogad el, amely szerepel ebben a megszorított listában.

Mindhárom keresés után kiléphet az automatából abban az esetben, ha sikerült annyira megszorítani az igekötőlistát, amennyire csak lehetséges. Ehhez elsősorban a megfelelő igekötő megtalálása szükséges, amelyet az infinitívusz megjelenése segít. A megfelelő igekötő igéhez kapcsolása után (vagy annak tisztázása után, hogy nincs igekötő) a vonzatkeretek lehívása következik. Ezek között már csak olyanok szerepelnek, amelyeket az ige szűk környezete alapján nem tud az algoritmus kizárni.

Ha az elemzett igén eleve volt igekötő, akkor a VFrame eljárás kimarad, egyből a vonzatkeretek lehívása következik.

3.2. Kiértékelés

A VFrame teljesítményét 1 000 tesztmondaton mértük ki. A tesztmondatokat (egész pontosan a tagmondatokat, amelyekben finit ige szerepel) az MNSZ 2.0.4 szolgáltatta. A VFrame teljesítményét három jelenség kezelése teszi ki: 1) a finit ige és az igekötőjének összekapcsolása, 2) az infinitívusz és az igekötőjének összekapcsolása, valamint 3) a finit ige és az infinitívuszi vonzatának összekapcsolása. A kiértékelés során a finit igékre koncentráltunk, a többi igének (így a melléknévi és határozói igeneveknek) és az igekötőjének vagy infinitívuszi vonzatának az összekapcsolását kihagytuk a vizsgálatból.

[11] A VFrame jelenleg a finit igéket és az infinitívuszokat kezeli, a melléknévi és határozói igenevek kezelése még nem megoldott.

2. ábra: A *VFrame* keresőeljárás állapotainak véges állapotú automata reprezentációja.

A tesztmondatokat ennek megfelelően úgy válogattuk, hogy egy finit igét tartalmazzanak, ezen kívül a tagmondatban legyen legalább vagy egy igekötő, vagy egy infinitívusz. A finit ige és az infinitívusz is lehet igekötős, hiszen a VFrame ezt külön esetként kezeli. Az így leszűrt mondatok közül véletlenszerűen választott 1 000 darabot vettünk fel a teszthalmazba. A tesztmondatok, valamint az összetételükkel kapcsolatos részletes információ megtalálható a VFrame git repozitóriumban[12].

A mondatokhoz kézzel megjelöltük a bennük található ige-igekötő, infinitívusz-igekötő, ige-infinitívusz kapcsolatokat, amely referenciaadatként szolgált a kiértékeléshez. A kézi annotációt és a VFrame kimenetét automatikusan összevetettük, és a megegyező vagy különböző eredményeket a megfelelő kategóriákba soroltuk, amelyeket a 6. táblázat tartalmaz.

A kategóriák számosságát mindhárom feladatra külön megnéztük, így megvizsgálhatjuk a VFrame teljesítményét az ige-igekötő, az infinitívusz-igekötő, a finit ige/infinitívusz-igekötő, valamint az ige-infinitívusz összekapcsolására is, de a VFrame teljesítményére összességében is. A VFrame teljesítményét összevetettük két egyéb eljárásával is. Az eredményeket a 7. táblázat mutatja.

A VFrame teljesítményét összevetettük egy baseline eljárással, amely a finit igét és az infinitívuszt az igekötővel, valamint a finit igét az infinitívuszi vonzatával azok közelsége alapján kapcsolja össze. Az eljárást Recski Gábor módszere [15] alapján dolgoztuk ki úgy, hogy az minden igekötőhöz a hozzá legközelebb

[12] `https://github.com/ppke-nlpg/vframe`

kategória	finit ige/infinitívusz–igekötő	finit ige–infinitívusz
TP	van igekötő és megtalálta	van infinitívusz és megtalálta
TN	nincs igekötő és nem találta meg	nincs infinitívusz és nem találta meg
FP	rossz igekötőt talált	rossz infinitívuszt talált
FN	nem találta meg az igekötőt	nem találta meg az infinitívuszt

6. táblázat. Az egyes kategóriák, amelyek az igekötő-ige és az ige-infinitívusz összekapcsolásánál felmerülnek. **TP**: valós pozitív, **TN**: valós negatív, **FP**: álpozitív és **FN**: álnegatív

álló igét (finit igét vagy infinitívuszt) rendeli. Ez a baseline eljárás tehát nem támaszkodik arra az információra, hogy az igének lehet-e infinitívuszi vonzata, illetve hogy milyen igekötője lehet egyáltalán, csupán annyi megszorítással él, hogy bizonyos finit igéknek nem keres igekötőt (ezek a segédige-szerű igék [16] alapján az *akar, bír, fog, kell, kezd, kíván, lehet, mer, óhajt, próbál, szabad, szándékozik, szeret, szokik, talál, tetszik* és a *tud* a létigével kiegészítve). A baseline eljárás az igéhez az infinitívuszt is annak közelsége alapján kapcsolja. Mind az igekötő, mind az infinitívusz összekapcsolásának feltétele, hogy egy tagmondatban szerepeljenek a finit igével, ez a feltétel a tesztmondatainkban mindig teljesül.

A baseline módszer mellett a `magyarlanc` függőségi elemzőjének [17] eredményével is összevetettük a VFrame teljesítményét. A függőségi elemzésben megnéztük, hogy hányszor egyezett meg a kézi annotációval az ige-igekötő és az ige-infinitívusz összekapcsolása. Ez az összekapcsolás gyakran amiatt volt hibás, hogy az elemző eleve rosszul állapította meg a finit igét (összesen 40 alkalommal).

		Fin–Ik	Inf–Ik	Fin/Inf–Ik	Fin–Inf	Összesen
pontosság	VFrame	**97,57**	**94,71**	**96,82**	**97,88**	**97,21**
	baseline	92,39	90,40	91,87	96,98	93,72
	`magyarlánc`	88,22	89,36	88,53	89,93	89,08
fedés	VFrame	96,30	**94,21**	**95,76**	98,34	96,70
	baseline	**96,49**	92,75	95,50	**99,05**	**96,80**
	`magyarlánc`	79,20	86,15	80,96	89,74	84,23
F-mérték	VFrame	**96,93**	**94,46**	**96,29**	**98,11**	**96,95**
	baseline	94,40	91,56	93,65	98,00	95,24
	`magyarlánc`	83,47	87,73	84,58	89,83	86,59

7. táblázat. A különböző alfeladatok és a VFrame teljesítményének kiértékelése összevetve egy baseline eljárással és a `magyarlanc` függőségi elemző eredményével. **Vastag betűvel** szedtük a legmagasabb értékeket.

Az eredmények azt mutatják, hogy a VFrame és a baseline módszer teljesítménye között csupán kis különbség van. A baseline módszer néhány alfeladatban a fedés szempontjából valamivel jobban teljesített, míg a VFrame minden alfeladatban és összesítve is a pontosságban volt jobb. A baseline módszer a tesztmondatok korpuszból származó szófaji címkéjére támaszkodik, így előfordult, hogy nem helyesen állapította meg a finit igét (pl. a *vagy* kötőszót vette finit igének). Ebből a hibából összesen 4 darab fordult elő. A baseline módszerhez képest a VFrame a felhasznált szótárnak köszönhetően tudott jobban teljesíteni, amely segítségével kizárhatóak a helytelen igekötő-ige vagy infinitívusz-ige kapcsolatok.

A baseline módszer és a VFrame esetében a két hibatípusba (FP, FN) tartozó hibákat megvizsgálva kiderül, hogy a legtöbbjük az eleve hibás bemenetből fakad. Mindkét eljárás esetében a korpuszból vett tesztmondatokban a szófajegyértelműsítő hibát vétett (például az *elég* főnevet igekötős finit igének jelölte meg).

Egy másik, hibát okozó jelenség az, amikor a példamondatban az ige töve nem a megfelelő módon van feltüntetve – elsősorban az ikes igék esetében. Például a *mit lélegeznek ki a falak* tesztmondatban a *lélegeznek* ige töve a korpuszban a következő formában jelenik meg: *lélegezik/lélegzik*. Az ige-igekötő-infinitívusz listában azonban ez a tő nem szerepel (hiába van felsorolva a *ki* igekötő a *lélegez* tőhöz a szótárban). A VFrame esetében más, relevánsabb hibatípust nem találtunk, tehát a hibás eredmény lényegében a hibás bemenetből adódik.

A `magyarlanc` eredménye mind a baseline módszerrel, mind pedig a VFrame módszerrel szemben alulmaradt. A hibák számát jelentősen növelte, hogy a másik két módszerhez képest jóval többször rontotta el a finit ige megtalálását.

Mindent összevetve a VFrame előnye elsősorban abban áll – a legmagasabb pontosság és F-mérték mellett –, hogy a balról jobbra és szavanként történő feldolgozás miatt beépíthető az AnaGramma elemzőbe.

4. Összegzés

A cikk első felében a Manócskát ismertettük, egy egységesített vonzatkerettárat, amelyet az eddig elérhető erőforrások felhasználásával és javításával készítettünk. A vonzatkerettár a legbővebb szabadon elérhető, ilyen jellegű erőforrás magyar nyelvre. A cikk második felében az infinitívuszi vonzatok vonzatkeret-egyértelműsítő szerepe mellett érveltünk, majd bemutattunk egy eljárást, amely az infinitívuszi vonzatok és az igekötők segítségével oldja meg a vonzatkeret-egyértelműsítés feladatát. A VFrame eljárás kiértékelése azt eredményezte, hogy mind a finit ige-igekötő, infinitívusz-igekötő és infinitívusz-finit ige összekapcsolását igen magas pontossággal és fedéssel hajtja végre. Az eljárás az AnaGramma nyelvi elemzőrendszer keretébe illeszkedik.

Hivatkozások

1. Prószéky, G., Indig, B.: Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. Alkalmazott Nyelvtudomány **15**(1-2) (2015) 29–44

2. Sass, B., Váradi, T., Pajzs, J., Kiss, M.: Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára. Tinta Könyvkiadó, Budapest (2010)
3. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002) European Language Resources Association, Paris, ELRA (2002) 385–389
4. Sass, B.: Igei szerkezetek gyakorisági szótára – Egy automatikus lexikai kinyerő eljárás és alkalmazása (2011) Doktori disszertáció. Pázmány Péter Katolikus Egyetem ITK.
5. Sass, B.: 28 millió szintaktikailag elemzett mondat és 5 00000 igei szerkezet. In Tanács, A., Varga, V., Vincze, V., eds.: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015), Szeged, SZTE TTIK Informatikai Tanszékcsoport (2015) 399–403
6. Sass, B.: „Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Váradi, T., ed.: Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból, Budapest, MTA Nyelvtudományi Intézet (2009) 117–129 `http://corpus.nytud.hu/mazsola`.
7. Halácsy, P., Kornai, A., László, N., András, R., Szakadát, I., Viktor, T.: Creating open language resources for Hungarian. In N, C., ed.: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). (2004) 203–210
8. Brew, C., Schulte im Walde, S.: Spectral clustering for german verbs. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. EMNLP '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 117–124
9. Kornai, A., Nemeskey, D.M., Recski, G.: Detecting Optional Arguments of Verbs. In Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
10. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Calzolari, N., et al., eds.: Proceedings of the 9th International Conference on Language Resources and Evaluation, May 26-31, 2014, Reykjavik, Iceland, ELRA (2014) 1719–1723
11. Kalivoda, Á.: A magyar igei komplexumok vizsgálata (2016) Mesterszakos szakdolgozat. PPKE-BTK. `https://github.com/kagnes/hungarian_verbal_complex`.
12. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
13. Vadász, N., Kalivoda, Á., Indig, B.: Ablak által világosan – Vonzatkeretegyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével. In Vincze, V., ed.: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017), Szeged, Szegedi Tudományegyetem Informatikai Tanszékcsoport (2017) 3–12
14. Frazier, L., Fodor, J.D.: The Sausage Machine: A New Two-Stage Parsing Model. Cognition **6**(4) (1978) 291–325
15. Recski, G.: A sekély mondattani elemzés további lépései. In Tanács, A., Vincze, V., eds.: VIII. Magyar Számítógépes Nyelvészeti Konferencia. (2011) 113–118
16. Kálmán C., Gy., Kálmán, L., Nádasdy, Á., Prószéky, G.: A magyar segédigék rendszere. Általános nyelvészeti tanulmányok – Tanulmányok a magyar mondattan köréből **17**(1) (1989) 49–103
17. Zsibrita, J., Vincze, V., Farkas, R.: MAGYARLANC: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP 2013. Hissar, Bulgária, 2013.09.08-2013.09.13. (2013) 763–771

Mazsola – mindenkinek

Sass Bálint

MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu

Kivonat A igék bővítményszerkezetének vizsgálatára szolgáló *Mazsola* eszközt szabadon elérhetővé tesszük. Ezzel megnyílik a lehetőség, hogy a saját adatai használatával bárki egyszerűen készítsen hasonló funkcionalitású kutatóeszközt. Amint látni fogjuk, a Mazsola alkalmas nem csak magyar nyelvű, hanem más nyelvű szövegek kezelésére is; nem csak egynyelvű, hanem párhuzamos korpuszra is; sőt nem csak igék és bővítmények, hanem más struktúrák vizsgálatára is.
Kulcsszavak: korpusz, ige, bővítmény, többszavas kifejezés, korpuszlekérdező

1. Bevezetés

Az igék és bővítmények vizsgálatára szolgáló *Mazsola* [1,2,3] korpuszlekérdező eszköz felülete régóta hozzáférhető online. A szöveganyag, amin a keresés történik a Magyar Nemzeti Szövegtár első, 187,6 millió szavas változatának [4] egy korábban már közzétett speciálisan a Mazsola számára feldolgozott formája [5].

Az elmúlt évek során a Mazsola magyar szakmai körökben viszonylag ismertté vált. Használják a nyelvtechnológiában [6,7], kontrasztív nyelvészeti [8] és nyelvelméleti kutatásokban [9,10], valamint a magyar mint idegen nyelv oktatása során [11,12]. 2013-ban született egy szakdolgozat, mely lengyel szövegeket feldolgozva a lengyel igéket vizsgálja hasonló felfogásban [13], és keletkeztek olyan igei vonzatkeret- illetve bővítménykeret-adatbázisok, melyek egyik ihletőjének a Mazsola tekinthető [14,15].

Maga a szoftvereszköz azonban – általánossága révén – sokkal többre képes annál, mint hogy magyar igék és bővítmények viszonyait kutassuk vele. Felismerve a nyílt hozzáférés előnyeit most közzétesszük a forráskódot. A cikk két fő üzenete tehát a következő: 1) a Mazsola szoftvereszköz mostantól hozzáférhető, bárki által használható, így saját adataink könnyen vizsgálhatók a Mazsola módszerével; 2) megfelelő adatbázis használatával számos egyéb célra is alkalmazható azon túl, ami az eredeti online változatból közvetlenül látszik. A továbbiakban a szoftver jellemzőiről valamint a kivonatban vázlatosan már említett felhasználási lehetőségekről lesz szó számos példával illusztrálva.

2. Használata, működése

A `http://corpus.nytud.hu/mazsola` címen elérhető magyar nyelvű szövegeket tartalmazó eredeti Mazsola felület alapvetően arra szolgál, hogy feltérképezzük az

egyes magyar igék melletti adott esetraggal/névutóval járó jellegzetes névszókat, illetve fordítva: az egyes (esetragos/névutós) névszók mellett jellegzetesen előforduló igéket. (Alapvető használatáról [3] 3.2. fejezetében olvashatunk. Megjegyezzük, hogy a mögöttes adatbázis az elváló igekötőket az igékhez illesztve tartalmazza, így az eszköz lehetőséget biztosít egy ige összes előfordulásának egységes vizsgálatára függetlenül attól, hogy az igekötő éppen elválik vagy nem.)

Az első esetben – az igék bővítményeire kérdezve – általában azt az eredményt kapjuk, hogy adott ige adott bővítményeként előforduló jellegzetes névszók egy szemantikailag koherens szóosztályt alkotnak, amiben kakukktojásként jelennek meg az idiomatikus, az igével intézményesült kifejezést, komplex igét, szólást alkotó szavak, mint például az *emelkedik -rA*-ragos bővítményei közül kilógó *jogerő*. A skála egyik szélén azok az esetek vannak, mikor szinte egységes a kijövő szóosztály (pl. a *fogyaszt* tárgyai különféle erőforrások); a másik szélén pedig azok, mikor az ige szinte minden egyes jellegzetes névszóval különálló komplex igét alkot (pl. a *hagy* legjellemzőbb tárgyaiból adódó kifejezések: *(sok) kivánnivalót hagy maga után*, *(mély) nyomot hagy (vhol)*, *(nem) hagy kétséget vmi felől*, *hagyja magát*, *üzenetet hagy*). A fentiek különféle variációi is előfordulnak, a *kitör* alanyai például nagyjából két osztályra bomlanak: a konkrét (*vihar*, *vulkán*) és a metaforikus jelentés mentén (*pánik*, *háború*, *botrány*, *forradalom*). Speciális eset, mikor adott ige mellett adott esetraggal csak néhány, szélső esetben egyetlen szó szerepel (pl.: a *csóvál* tárgyaként a *fej*, *farok*, *kutya* vagy a *huny* tárgyaként a *szem*).

Fontos kiemelni a második esetet (a fenti eset duálisát), amikor bizonyos bővítmények megadásával az igére kérdezünk rá. Erről kevesebb szó esett korábban. Általában egyszerűen mindenféle jellegzetesen az adott bővítménnyel használatos igét kapunk eredményül egy ilyen lekérdezésre. A bővítményt megadhatjuk esetrag nélkül is, így esetragtól függetlenül, értsd tetszőleges esetraggal rákérdez a rendszer. A *ló* bővítményre legjellegzetesebbként kapjuk: *ül lovon*, *megül lovat*, *húz ló*, *leszáll lóról*. Bizonyos esetekben itt is szemantikailag koherens szóosztályt kapunk: nagyon hasonló jelentésű igéket. Erre talán legszebb példa a *hideg + hátán* bővítményekre kapott *végigfut*, *futkos*, *futkározik* igeosztály (1. ábra). A *kígyót* bővítmény két legjellegzetesebb igéje viszont a *kiabál* és a *melenget*, melyekkel ez a szó külön-külön önálló szerkezetet, szólást alkot. És itt is előfordul az az eset, hogy kizárólag egy ige szeret az adott bővítménnyel társulni (pl.: *vérszemet* → *kap*, vagy *füst alatt* → *elintéz*).

A Mazsolánál jóval összetettebb, de mégis hozzá sok szempontból hasonló eszköz a közismert SketchEngine [16,17]. A SketchEngine-ben – a benne alkalmazott szkeccs-nyelvtanok (sketch grammar) segítségével – explicite megmondjuk, hogy mely elemek kapcsolataira vagyunk kíváncsiak, a Mazsolában ezzel szemben implicite minden mindenhez kapcsolódik, és az összes ilyen kapcsolat vizsgálható. A SketchEngine alapvetően mindig *két* elem viszonyát nézi. Van benne kiegészítő eszköz, amivel meg lehet valósítani 3 elem összekapcsolását, de ez az, amiben a Mazsola jobb: *rekurzívan* képes a korábban összekapcsolt elemekhez tetszőleges számú újat kapcsolni, és egy kiválasztott elemnek a jellegzetességét vizsgálni a kialakult szerkezet többi tagja összességéhez képest, amint ezt a fenti példák-

1. ábra: A *hideg* és *hátán* bővítmények megadásának módja, és a hozzájuk tartozó jellegzetes igék lekérdezése. Fontos, hogy az „Eloszlás" kapcsoló az ige sorára van állítva.

ban láttuk is. Megvizsgálhatjuk tehát, hogy a *tesz*-nek milyen jellegzetes tárgyai vannak (*elég*, *szer*, *feljelentés*, *ígéret* stb.), majd hogy az *tesz eleget*-nek milyen *-nAk* ragos bővítményei (*kötelezettség*, *követelmény*, *kérés*, *meghívás* stb.), aztán hogy a *tesz eleget kötelezettségnek* szerkezetnek milyen alanyai (*önkormányzat*, *kormány*, *országgyűlés*, *főváros*), és így tovább. Fontos különbség még, hogy a Mazsola immár szabadon és ingyenesen hozzáférhető.

3. A szoftver jellemzői, elérhetősége, installálása

A Mazsola letölthető a `https://github.com/sassbalint/mazsola` repozitóriumból. Perl nyelven van megvalósítva, linux operációs rendszeren működik.

Az installálás leírását a github-on található `README.md`-ben olvashatjuk. Alapvetően annyit kell tenni, hogy `make` segítségével külön-külön installálni kell magát a szoftvert, illetve az adatbázisokat egy webszerverre. Az installálás után a Mazsola böngészőből használható. A repozitórium tartalmazza a fő szkriptet (`mazsola_skel.pl`), a többnyelvűsítést megvalósító szkriptet, különbözö config fájlokat, valamint az installáláshoz szükséges `Makefile`-t.

Ezen kívül itt találjuk még a `data` könyvtárt, ebben helyezhetők el a korpuszok, speciális, a Mazsola számára alkalmas formátumban. A Mazsola-adatbázisok formátumának leírása [5] 2. fejezetében található.

A Mazsola repozitóriuma eleve tartalmaz több kisméretű - néhány ezer tagmondatból álló - minta-adatbázist (1. táblázat). Egyrészt, hogy ezek segítségével ki lehessen próbálni a rendszer működését, másrészt, hogy alkalmas példákon tanulmányozhassuk az ilyen adatbázisok szerkezetét, kinézetét.

Az egyes minta-korpuszokhoz az installálás után a Mazsola felületén megfelelő példákat is találunk, melyekre rákattintva és a kérést lefuttatva kipróbálhatjuk az eszközt. A dán korpusz hasonló részletességgel van elemezve, mint a magyar. A szerb korpuszban egy primitívebb megoldás van megvalósítva. Mivel a korpuszban nem volt triviális meghatározni, hogy melyik szó alany és melyik tárgy, az elöljáró nélküli főneveket egy kalap alá vettük, és ezek mind `ANYCASE`-ként vannak kódolva. A példák kipróbálásakor láthatjuk, hogy ebből az egyszerűbb kódolással ellátott korpuszból is a korábbiakhoz hasonló hasznos eredményeket

1. táblázat. A Mazsolához adott minta-adatbázisok forrásai.

korpusz	nyelv
MNSZ [4]	magyar
Danish Dependency Treebank [18]	dán
Intera korpusz [19]	szerb
Dutch Parallel Corpus [20]	holland-francia

tud előhozni a Mazsola. Az *imati* (birtokol) ige melletti tárgyakat (az `ANYCASE` segítségével) nézve jellegzetesként megkapjuk a *pravo* (jog) szót, majd a *na* elöljáróval kiegészítve a keresést, hogy mire lehet jogunk.

Látjuk azt, hogy nyugodtan alkalmazhatjuk más nyelvekre az eszközt, ugyanolyan jellegű működést fogunk tapasztalni, mint amit a magyar esetén leírtunk. Természetesen a Mazsola mindig azon információk alapján tud dolgozni, melyek az adatbázisban kódolva vannak, másképp fogalmazva az, hogy adott Mazsola-példány milyen eredményeket ad, attól függ, hogy mit kódoltunk bele az adatbázisba. Alkalmasan választott kódolással szemantikai összefüggéseket vagy párhuzamos korpuszt is kezelni tudunk, utóbbira még visszatérünk.

Ha saját adatbázist akarunk használni, akkor el kell helyeznünk a `data` könyvtárban, és készítenünk kell hozzá egy megfelelő config fájlt (amit – ahogyan ezt a repozitóriumban most is látjuk – `mazsola_config_skel.pl` néven kell belinkelnünk). A config fájlban be kell állítanunk a korpuszt: meg kell adnunk 1) az adatbázisfájl nevét; 2) a szótőgyakoriságokat tartalmazó kiegészítő adatbázisfájl nevét (amit a `lemmafreq` könyvtárban kell elhelyeznünk); 3) a korpusz nevét; 4) és azt a küszöbértéket, aminél gyakrabban előforduló legjellegzetesebb szavakat az eredményoldalak felső részén meg szeretnénk jeleníteni. Ha rendelkezésre áll a szótőgyakoriságokat tartalmazó kiegészítő adatbázis, akkor tud a rendszer *jellegzetesség*et (*salience*-et) számolni, egyébként csak az egyszerű gyakoriságokat jeleníti meg. A fentieken kívül meg kell adnunk a korpusz karakterkódolását (jelenleg csak egybájtos karakterkódolások támogatottak), valamint megfelelő formátumban a felületen megjelenítendő, korpuszhoz tartozó kattintható példákat. Végül el kell végeznünk az installálás lépéseit. Config fájlra is találunk néhány mintát a repozitóriumban (`mazsola_config_*.pl`).

4. Létező Mazsolák

A minta-adatbázisoknak megfelelő teljes korpuszok nagy része elérhető a Mazsolán keresztül online a 2. táblázatban szereplő internetes címeken.

Ezeken a felületeken is találunk példákat, illetve a minta-korpuszokhoz tartozó példák kipróbálhatók ezeken az URL-eken is jóval nagyobb méretű szöveganyagon.

2. táblázat. A Mazsolához adott minta-adatbázisok forrásai.

korpusz	internetes cím
MNSZ	`http://corpus.nytud.hu/mazsola`
Danish Dependency Treebank	`http://corpus.nytud.hu/vabd` [21]
Intera korpusz	`http://corpus.nytud.hu/vabs`
Dutch Parallel Corpus	`http://corpus.nytud.hu/vabnlfr` (terv) [22]

5. Összetettebb példák, trükkök, kiterjesztések

5.1. Egy korábbi példa

Korábban, egy az elemzetlen Magyar történei szövegtáron végzett egyszerű vizsgálat kapcsán azt írtam, hogy az *erőt vesz rajta* alanyai általában „érzések, azon belül is inkább negatív tartalmú érzések, hangulatok” [23]. Nézzük most ezt meg a Mazsolán, ami kifejezetten az effajta kérdések megválaszolására szolgál. Azt kapjuk, hogy az *erőt vesz* kifejezés *-n* ragos jellegzetes bővítményei: *maga*, *ő*, *ők*, *mi*, *ember*. A *maga*-t – más kifejezés lévén – érdemes a mostani vizsgálatból kihagyni, a többi relevánsnak tűnik: megadjuk tehát esetnek a *-n* ragot, és vonzattövek közül letiltjuk („Nem” jelölőnégyzet) a *maga*-t. Ennek a kifejezésnek (*erőt vesz vmin* (kivéve: *magán*)) kérdezzük le a jellegzetes alanyait a Mazsola fentebb leírt rekurzív vizsgálati módszerét alkalmazva (2. ábra).

2. ábra: Az *erőt vesz vmin* (kivéve: *magán*) szerkezet jellegzetes alanyai. „Eloszlás” kapcsoló az alany során.

Azt kapjuk, hogy a legjellegzetesebb alany éppen az *érzés*, és a további találatok között megtaláljuk a *bágyadtság*, *bénaság*, *betegség*, *borzongás*, *búskomorság*, *bűntudat* stb. szavakat, azaz az eredeti állítás megállja a helyét. A második legjellegzetesebb alany viszont a *kapu*, itt egy (bibliai) szólás bukkan fel kakukktojásként, a konkrét vizsgálatból a *maga*-hoz hasonlóan ezt is érdemes kihagyni.

5.2. Reguláris kifejezések

Megjegyezzük, hogy a lekérdezőfelület mezőiben bizonyos reguláris kifejezéseket is használhatunk.

Ha például az a feladat, hogy *ezik/szik* alternációt keressünk az *emlékezik/emlékszik* példájára, akkor az alapján a sejtés alapján, hogy az ilyen igéknek *ék*-végű főnévből kell származnia, megpróbálkozhatunk az

Igető=`.*ékezik`

lekérdezéssel. Az eredmény: *emlékezik*, *verejtékezik*. Ezek valóban *ék*-végű főnevekből származnak, de *verejtékszik* alak ennek ellenére nincsen. Úgy tűnik, hogy az *emlékezik* ebből a szempontból egyedi. Hasonlóan az

Igető=`.*g[ae]t`

lekérdezés segítségével megragadhatjuk a *-gAt*-képzős igéket.

A „Vonzattő” és a „Szó” mező esetén több szóalakot is megadhatunk, ez úgy értelmeződik, hogy a szólista bármelyik tagja előfordulhat. Azaz ha például férfiakra jellemző igéket keresünk, akkor megadhatjuk a alanyként a *férfi*, *fiú*, *kamasz*, *kisfiú*, *öregember*, *férj*, *após*, *úr* stb. szólistát. Annak ellenére, hogy a szólisták a fentiek szerint diszjunktívan értelmeződnek, meg tudunk adni a „Szó” mezőben többszavas kifejezést is. Reguláris kifejezés cseles használatával, ha a szavakat ponttal kötjük össze, az azt fogja eredményezni, hogy közvetlenül egymás után kell következniük. A `rázza.az.ág` lekérdezés esetén csak azokat a találatokat kapjuk meg, amiben a *rázza az ág* sztring szerepel, nem pedig *rázza*, *az*, *ág* szavak bármelyike.

5.3. Feladatmegoldás Mazsolával

Nézzük néhány konkrét problémát, feladattípust, melyeket a Mazsola segítségével oldhatunk meg.

A *macska* és a *cica* közeli szinonimáknak tűnnek. A Mazsola szerint az előbbi *dorombol*, *nyávog*, *ül*, *megesz*, *leugrik*, *eltűnik*, *nyújtózik*, az utóbbi viszont *elvisz*.

Mit jelent az, hogy *kooptál*? Az ige tárgyainak vizsgálatából levonhatjuk azt a következtetést, hogy a jelentés nagyjából az, hogy bizottságba új tagot vesz fel.

A fordítói munka során arra lehet hasznos a Mazsola, hogy bemutatja a bevett kollokációkat. Ha a *részletes magyarázatot kíván* megfogalmazás nem tűnik ideálisnak, akkor élhetünk a Mazsola javaslatával, amely ez esetben: *részletes magyarázatot igényel*.

A *kiküszöböli a csorbát* kifejezés kicsit furcsának hat. Biztos, hogy ezt az igét szoktuk itt használni? Milyen igéket ad a Mazsola a *csorbát* tárgyhoz? Az eredmény a *szenved* és az *ejt* mellett több *kiköszörül* és jóval kevesebb *kiküszöböl*. Persze, a csorbát inkább kiköszörüljük. Mit történhetett? Az igék felől vizsgálódva azt látjuk, hogy a *kiköszörül* szinte kizárólag a *csorbát* tárggyal fordul elő, a két nagyságrenddel gyakoribb *kiküszöböl* viszont számos különféle tárggyal,

melyeknek gyakori jelentéseleme a valamiféle hiányosság (*hiba*, *hiányosság*, *probléma*, *anomália* stb.). A gyakoribb és nagyon hasonló hangzású kifejezés magába olvasztotta a ritkább, hasonló jelentésű kifejezést. Az állítás alátámasztására persze még érdemes történeti korpuszban utánajárni, hogy melyik kifejezés mikortól adatolható. Hasonlóan vizsgálhatók a *számításon kívül hagy* körüli fejlemények.

Kapcsolatot fedezhetünk fel az *össze* igekötő és a *-vAl* ragos vonzatok között. Az *össze* igekötős igék jelentős része *-vAl* ragos vonzattal jár, másik nagy része viszont e nélkül. Milyen eltéréseket találunk a két csoport között? Az első csoport (jellemzően *-vAl* ragos bővítménnyel): *összefügg*, *összeköt*, *összehasonlít*, *összekever*, *összevet* stb. A második csoport (jellemzően *-vAl* ragos bővítmény nélkül): *összegyűlik*, *összetörik*, *összeállít*, *összehív*, *összeszed* stb. Azt látjuk, hogy az első esetben két dolgot állítunk relációba, a második esetben pedig sokat. Az *össze* jelentése az első esetben valamiféle „szembe”, a második esetben pedig valamiféle „egybe”. E felismerésnek által két csoportra osztva strukturálhatjuk az *össze* igekötő jelentéseit (az alapigétől függetlenül!), és mondhatjuk, hogy a *-vAl* az első esetben az igekötőnek a vonzata.

5.4. Párhuzamos korpusz kódolása

Térjünk vissza arra a kérdésre, hogy hogyan kezelhető a Mazsolában párhuzamos korpusz. Az adatbázis kódolásában rejlik a megoldás, a szoftver maga változatlan módon használható. Hogyan valósul meg a már említett holland-francia párhuzamos korpusz esetében a megfelelő Mazsola-adatbázis kódolása? Az egynyelvű magyar `hnc_sample.mazsoladb` mintafájlban ezt látjuk:

```
Ha a város a tárcáktól pénzt nyer , stem@@nyer ABL@@tárca
  ACC@@pénz NOM@@város
```

A sor elején maga a tagmondat látható, `stem@@` után következik az igető, majd sorra az esetek a megfelelő szóval együtt. A holland-francia párhuzamos `data/tltnlfr_sample.mazsoladb` mintafájlban viszont a következő sort találjuk:

```
nl_deze te vertalen in wiskundige formules . / fr_les traduire
  en expressions mathématiques . stem@@nl_vertalen+fr_traduire
  fr_en@@expression nl_in@@formule
```

Amint látjuk, azt a megoldást választottuk, hogy a perrel elválasztott, `nl_` és `fr_` címkével azonosított kétnyelvű tagmondat után az ige helyén a holland-francia igepár van (szintén nyelvi címkével ellátva és `+` jellel összekapcsolva), a bővítmények között pedig a hollandok és a franciák „vegyesen” egy halmazban szerepelnek. Érdemes élesben kipróbálni a TLT minta-korpuszhoz mellékelt példákat.

5.5. Nem csak igére

Vegyük észre, hogy az eszköz tetszőleges olyan struktúrájú objektumok kutatására alkalmas, ahol egy egységen belül van egy központi elem (esetünkben az ige),

és mellette vannak hozzá tartozó elemek (esetünkben a bővítményi névszók), amiket valamiféle kapcsolat (esetünkben az esetragok/névutók) fűz a központi elemhez.

Ha ilyen struktúrájúak az adataink, akkor tudunk hozzá Mazsolát készíteni, bármilyen konkrét tartalmúak is legyenek az adatok.

Vegyük például a jelző+főnév párokat. Ebben az esetben a főnév lesz a központi elem, a jelző a hozzá tartozó elem, és a kapcsolat csupán annak rögzítése lesz, hogy a jelző a főnévhez tartozik. (Az eredeti ige-központú esetben mindig annyiféle kapcsolat volt, ahány esetrag ill. névutó van, itt viszont mindössze egyfajta kapcsolat van.) Ilyen felépítésű adatbázis használatával a Mazsola működésének megfelelően meg fogjuk kapni a főnevek jellegzetes jelzőit, illetve a jelzőkhöz tartozó jellegzetes főneveket.

6. Konklúzió

Szabadon elérhető, használható a Mazsola eszköz. A saját adatait, korpuszát megfelelően előkészítve, és a rendszerbe beleillesztve bárki készíthet igék és bővítmények vizsgálatára alkalmas kutatóeszközt. A [5]-ben közzétett adatbázis használatával előállítható az eredeti magyar Mazsola felület. Az eredeti rendszer reprodukálásánál azonban sokkal izgalmasabb, ha a saját adatainkra építve a fenti ötletek és kiterjesztések alapján készítünk különféle más nyelvekre és/vagy más struktúrákra vonatkozó korpuszlekérdezőket a Mazsola segítségével.

Köszönetnyilvánítás

A munkálatokat az MTA Bolyai János Kutatási Ösztöndíja támogatta (ügyszám: BO/00064/17/1; időtartam: 2017-2020).

Hivatkozások

1. Sass, B.: „Mazsola” – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: Váradi Tamás (szerk.): Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból, MTA Nyelvtudományi Intézet, Budapest (2009) 117–129
2. Sass, B.: Korpusznyelvészeti eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Sinkovics, B., ed.: LingDok 8. – Nyelvész-doktoranduszok dolgozatai. JATEPress, Szeged (2009) 143–155
3. Sass, B.: Igei szerkezetek gyakorisági szótára - egy automatikus lexikai kinyerő eljárás és alkalmazása. PhD thesis, PPKE ITK (2011)
4. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain (2002) 385–389
5. Sass, B.: 28 millió szintaktikailag elemzett mondat es 500000 igei szerkezet. In: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2015), Szeged: JATEPress (2015) 303–308
6. Prószéky, G.: A nyelvtechnológia és a magyar nyelvtudomány. Magyar Nyelv **108**(1) (2012) 1–18

7. Indig, B., Laki, L., Gábor, P.: Mozaik nyelvmodell az anagramma elemzőhöz. In: XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2016), SZTE, Szeged (2016) 260–270
8. Forintos, É.: New Zealand-Hungarian language contact phenomena. In Bátyi, S., Vígh-Szabó, M., eds.: A nyelv: rendszer, használat, alkalmazás. Budapest: Tinta Könyvkiadó (2015) 199–211
9. Kálmán, L.: Bővítménykeretek mint konstrukciók. In Kas, B., ed.: Szavad ne feledd! Tanulmányok Bánréti Zoltán tiszteletére. Budapest: MTA Nyelvtudományi Intézet (2016) 61–72
10. Kenesei, I.: A szófaj fogalmáról és meghatározásáról. In Kas, B., ed.: Szavad ne feledd! Tanulmányok Bánréti Zoltán tiszteletére. Budapest: MTA Nyelvtudományi Intézet (2016) 81–92
11. Joachim, L.: Morfológiai és szintaktikai információk a magyar mint idegen nyelv tanulói számára készített szótárakban és szószedetekben (egy tanulói szótár előmunkálatainak szempontjai, tapasztalatai). In Szűcs, T., Nádor, O., eds.: Hungarológiai Évkönyv 14. Pécs: Dialóg Campus Kiadó (2013) 40–62
12. Wéber, K.: Elvárt és használt szintaktikai szerkezetek megoszlása b1, b2 és c1 szintű magyar mint idegen nyelvi fogalmazásfeladatokban. In Szűcs, T., Nádor, O., eds.: Hungarológiai Évkönyv 14. Pécs: Dialóg Campus Kiadó (2013) 63–73
13. Wołosz, M.: A lengyel igei szerkezetek számítógépes kinyerése és gyakorisági vizsgálatuk. Master's thesis, ELTE IK (2013)
14. Kornai, A., Nemeskey, D.M., Recski, G.: Detecting optional arguments of verbs. In Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
15. Vadász, N., Kalivoda, A., Indig, B.: Egy egységesített magyar igei vonzatkerettár építése és felhasználása. In: XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2018), SZTE, Szeged (2018) jelen kötetben
16. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: The Sketch Engine. In: Proceedings of EURALEX 2004, Lorient, France (2004) 105–116
17. Kilgarriff, A.: The Sketch Engine: Ten years on. Lexicography **1**(1) (2014) 7–36
18. Trautner Kromann, M.: The Danish Dependency Treebank and the DTAG treebank tool. In: Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT 2003), Växjö, Sweden (2003)
19. Gavrilidou, M., Labropoulou, P., Piperidis, S., Giouli, V., Calzolari, N., Monachini, M., Soria, C., Choukri, K.: Language resources production models: the case of the INTERA multilingual corpus and terminology. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). (2006)
20. Macken, L., Trushkina, J., Paulussen, H., Rura, L., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus. A multilingual annotated corpus. In: Proceedings of Corpus Linguistics 2007, Birmingham, United Kingdom (2007)
21. Sass, B.: Verb Argument Browser for Danish. In: Proceedings of the 17th Nordic Conference of Computational Linguistics, NoDaLiDa 2009, Odense, Denmark (2009) 263–266
22. Sass, B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2010), SZTE, Szeged (2010) 102–110
23. Sass, B.: Keresés korpuszban: a kibővített Magyar történeti szövegtár új keresőfelülete. In Forgács, T., Németh, M., Sinkovics, B., eds.: A nyelvtörténeti kutatások újabb eredményei IX., Szeged: SZTE Magyar Nyelvészeti Tanszék (2017) 267–277

Nulla vagy semmi? Esetegyértelműsítés az ablakban

Ligeti-Nagy Noémi[1,3], Vadász Noémi[1,3], Dömötör Andrea[1,3], Indig Balázs[2,3]

[1]Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar
2087 Piliscsaba, Egyetem u. 1.
[2]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter u. 50/A
[3]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/A
VEZETEKNEV.KERESZTNEV@itk.ppke.hu

Kivonat A testes esetragot magukon nem viselő névszók mondatbeli szerepének azonosítása egy gépi mondatelemző számára nem triviális feladat. A magyarban az ilyen elem többféle szerepet is betölthet: lehet a mondat alanya; egy jelöletlen birtokos szerkezet birtokosa; egy névutós szerkezetben a névutó vonzata; egy, a főnévi fejet módosító elem; vagy névszói állítmány. A cikkben egy eljárást ismertetünk, amely pusztán a kérdéses névszót követő két tokenből nagy százalékban képes megoldani az esetegyértelműsítést.
Kulcsszavak: esetegyértelműsítés, korpusznyelvészet, mondatelemző

1. Bevezetés

A cikk az AnaGramma nyelvi elemzőrendszer [1,2] számára kínál egy esetegyértelműsítő eljárást, amely a testes esetragot nem viselő névszók mondatbeli szerepét képes tisztázni a névszó szűk környezete alapján. A nyelvi elemzőrendszer működési alapelveinek ismertetése után a testes esetrag nélküli névszók lehetséges mondatbeli szerepeit vesszük sorra, majd bemutatjuk magát az esetegyértelműsítő algoritmust.

Rögtön az elején szükséges tisztázni két terminológiai kérdést. A főnevek (köznevek és tulajdonnevek), a melléknevek, a számnevek és a melléknévi igenevek alkotják azt a csoportot, amelynek elemeit mi most tárgyalni kívánjuk, ezekre a továbbiakban az egyszerűség kedvéért együttesen névszóként fogunk hivatkozni.

Mondatbeli szerep alatt jelen tanulmányban egyrészt a hagyományos értelemben vett mondatbeli szerepet értjük, azaz azt, hogy az adott szó vonzat vagy határozó, másrészt mindazokat a szerepeket, amelyek főnévi csoporton belüli viszonyokban tölthetők be. Tehát a mondatbeli szerep azonosítása annak meghatározását jelenti, hogy az adott névszó vonzat (vagy határozó)-e a mondatban, és ha nem, akkor milyen egyéb funkciója van egy főnévi csoporton belül.

2. Előzmények

A testes esetrag nélküli névszók mondatbeli szerepét tisztázó algoritmusunk az AnaGramma nyelvi elemzőrendszer keretrendszerébe illeszkedik, amely az emberi szövegfeldolgozást modellálja, ezért az elemzendő szöveget balról jobbra és szavanként elemzi. Az aktuálisan vizsgált szó elemzését az azt közvetlenül követő néhány szó is befolyásolhatja, amihez az AnaGramma egy +2 token méretű előretekintő elemzési ablakot használ, melyben a kétfázisú mondatelemzés [3] első fázisa valósul meg. Az elemzés első fázisában a mondatot alkotó összetevők előkészítése történik, amelyek a második fázisban megkapják a mondatbeli szerepüket. A névszók esetében az első fázisban történik a testes esetrag nélküli elemek lehetséges mondatbeli szerepének tisztázása[1].

2.1. A testes esetrag nélküli névszók

A magukon testes esetragot nem viselő névszók szerepe a mondatban megfigyeléseink alapján a következő lehet: lehet a mondat alanya, jelöletlen birtokos szerkezet birtokosa, egy névutós szerkezetben a névutó vonzata[2], főnevet módosító elem vagy névszói állítmány (része). A különböző mondatbeli szerepek alapján megállapíthatjuk, hogy a testes esetragot nem viselő névszó végén vagy egy testetlen esetrag van (ha a névszó alanyesetben vagy jelöletlen birtokos esetben van), vagy nincs semmi (ha főnévi fejet módosít vagy egy névutó vonzata). A névszói állítmány szerepében álló névszók esetében (elsősorban technikai megfontolásból) szintén egy testetlen esetragot feltételezünk, amely a névszó „állítmány-esetét” jelöli. A cikkben a testetlen esetragokat és az esetragnélküliséget következetesen így jelöljük:

- az alanyeset és a jelöletlen birtokos testetlen esetragja: α
- az esetragnélküliség jele: 0
- az állítmány esetének testetlen esetragja: β

A 3. fejezetben ismertetésre kerülő korpuszvizsgálataink során szembesültünk azzal, hogy a fentieken kívül még két további olyan eset van, amikor a testetlenség jelöl egy esetet vagy egy főnévi csoporton belüli viszonyt. Az egyik a vokatívusz (ld. az 1-es példában a *báróné* tokent, amely egy testes esetrag nélküli főnév, melynek a mondatbeli szerepe itt a „megszólítás-eset”). A másik esetet nehéz a nevén nevezni, a 2a és a 2b alatti példákkal igyekszünk megvilágítani (a mondatok az MNSZ 2.0.4-es verziójából [5] származnak). A példákban félkövérrel szedtük a kérdéses elemet. Minden esetben többtagú nevekről van szó; vagy egy tulajdonnév és egy köznév kapcsolatáról, ahol a köznév a tulajdonnév birtokosának foglalkozását, társadalmi szerepét, nem élő dolgok esetében például

[1] Az igék esetében pedig a lehetséges vonzatkeretek megszorítása történik [4].

[2] Az, hogy a testes esetrag nélküli, névutó előtti névszót a névutó vonzatának hívjuk, csupán egy praktikus megfontolásból alkalmazott egyszerűsítés, nem elméleti nyelvészeti megalapozottságú terminológiahasználat. A névutó és az esetrag nélküli névszó együttjárását kívánjuk vele érzékeltetni: nincs névutó névszó nélkül.

funkcióját (*Opel személygépjármű*) nevezi meg, vagy több köznév kapcsolatáról (*elnök úr*), ahol hasonló az elemek közötti viszony, mint a tulajdonnév-köznév kapcsolat esetén. Ezekben a helyzetekben az „előtagon", a félkövérrel szedett elemeken nem feltételezhetjük egyik eddig említett esetnek a jelenlétét sem. Ezt az esetet mi most *többtagúnév-esetnek* hívjuk, de elemzése mélyebb kutatást igényel, így ezen tanulmány keretein belül nem vizsgáljuk. Sem a vokatívusz, sem a többtagúnév-eset egyértelműsítése nem célja a 3. fejezetben bemutatott algoritmusunknak, ezekkel később, a kutatás egy következő fázisában kell foglalkozni.

(1) A hintóm rendelkezésére áll, **báróné**!

(2) a. **Tóth** kisasszony a lövést követően állítólag a park fái mögé bújt.

b. Ölelgetnétek ti még engem, hogy így, aranyos **elnök** elvtárs, meg úgy aranyos **elnök** úr ...

A jelen cikkben ismertetett `Nom-or-What` eljárás az AnaGramma kereteiben működő `Nom-or-Gen` eljárás [6] továbbfejlesztett és kiegészített változata, melynek a célkitűzése az, hogy egy testes esetrag nélküli elem esetében annak szűk kontextusa alapján eldöntse, hogy egy jelöletlen birtokos szerkezet birtokosa-e. A `Nom-or-Gen` eljárás a főnevet módosító elemeket egységesen `NPMod` kategóriával látta el. Emellett a főnevet módosító elemek rendelkeznek csak a saját szófaji kategóriájukra jellemző jegyekkel is (pl. a számnevek a csak a számnevekre jellemző jegyekkel). Mivel a korpuszokban a testes esetrag nélküli névszók esetcímkéje kivétel nélkül `NOM`, és az esetegyértelműsítő eljárás feladata tisztázni ezeknek az elemeknek az aktuális szerepét, ezért az eljárás a `Nom-or-What` nevet kapta.

A `Nom-or-Gen` eljárást tehát úgy feljesztettük tovább, hogy immár az alanyesettel, a főnevet módosító elemekkel és a névutók vonzataival kapcsolatban is képes legyen döntést hozni. Az állítmányi szerepű névszók esetében csak akkor tudna dönteni, amikor a névszó szűk környezetében szerepel olyan információ, amely ezt lehetővé teszi. Ez azonban ritkán fordul elő, az állítmányi szerep felismerésében ugyanis jellemzően inkább a névszót megelőző mondatrész nyújt nagyobb segítséget. Az ablak alapján a névszói állítmányi eset csak akkor azonosítható egyértelműen, ha az ablakban létige található, sőt, még ezt is le kell szűkíteni 1-2. személyű létigére, egyébként nem lehetünk biztosak benne, hogy a létige valóban kopula. Míg a 3a példában az ablak alapján egyértelműsíthető a névszói állítmányi eset, addig a 3b és a 3c esetragjainak elkülönítéséhez az előzmények ismeretére is szükség van. Még nehezebb a helyzet, ha egyáltalán nincs a mondatban testes kopula, ilyenkor két fő ismérvre támaszkodhatunk: 1) a mondatban nincs finit ige, 2) a mondatban már van nominatívusz. Mindkét jellemző csak a nagyobb kontextusból, jellemzően az állítmányi szerepű névszót megelőző mondatrészből állapítható meg.

(3) a. *Negyedik* ***gyerek*** *voltam a családban.*
gyerek/gyerek/FN.β

b. *Erdélyi Dániel maga is iskolás* ***gyerek*** *volt a film ábrázolta korszakban.*
gyerek/gyerek/FN.β

c. *Kevés zsidó* ***gyerek*** *volt a falumban.*
gyerek/gyerek/FN.NOM

Mindezek alapján azt feltételezzük, hogy az alanyesetet, a birtokos esetet, a főnévi módosító szerepet és a névutó vonzatát a kétfázisú mondatelemzés első fázisában tisztázzuk, amikor a mondat elemeit előkészítjük arra, hogy megkapják szerepüket a mondatban. A névszói állítmány megtalálása a kétfázisú mondatelemzés második fázisában történik, hiszen felismeréséhez szélesebb kontextus szükséges. Ebben a cikkben csak az elemzés első fázisával foglalkozunk.

3. Módszer

Kutatásunk első lépéseként a Szeged Treebank 2.0 [7] főnévi csoportként jelölt egységein belüli, testes esetrag nélküli névszói (és melléknévi igenévi) tokenek viselkedését vizsgáltuk. Már a kutatás korai fázisában egyértelművé vált, hogy külön szabályrendszert kell alkotni a főnevekre, a melléknevekre, a számnevekre és a melléknévi igenevekre, mivel jellemzően máshogy viselkednek, ami a testes esetrag nélküli tokeneiket illeti. Mindegyik szófaji kategóriához két-két listát készítettünk: egyet azokról az esetekről, amikor az adott szófaji kategóriához tartozó, NOM-nak címkézett elem egy NP belsejében található, és azokról az esetekről, amikor az adott szófaji kategóriához tartozó, NOM-nak címkézett elem egy NP legutolsó eleme. Ezen listák alapján állapítottuk meg, melyek azok az elemek, amelyek előtt biztosan eldől egy NOM jelentése, és melyek azok, amelyek előtt nem. Megfigyeléseinket, és az algoritmus működését döntési fák segítségével szemléltetjük (1., 2. és 3. ábrák).

Mivel az algoritmus az ablakban szereplő tokenekre támaszkodva dönt, értelemszerűen lesznek olyan esetek, amikor nem egyértelműsítheti a testetlenséget. Ezekre az esetekre *alapértelmezett (default)* értéket kellett meghatároznunk: a főnevek (mind a köznevek, mind a tulajdonnevek) esetében ez az *alapértelmezett* érték az α, az alanyeset és a jelöletlen birtokos eset testetlen esetragja; a melléknevek, számnevek és a melléknévi igenevek esetében pedig a *default_*0, ami abban különbözik a 0-tól, hogy nem mond egyértelmű ítéletet, az adott NPMod kategóriájú elem még NOM vagy GEN címkét is kaphat később, a mondat egészének ismerete alapján. A 4a példában az *ember* token mondatbeli szerepe az ablakban látható elemek alapján nem állapítható meg biztosan, ezért α esetragot kap. A 4b példában az *egyik* esete nem egyértelműsíthető az ablak alapján, ezért *default_*0 esetet kap.

(4) a. *Elindul reggel hazulról az* ***ember*** *ingujjasan vagy pulóverben.*
ember/ember/FN.α

b. *A Nagy_ Szent_ Bazil-rendnek két kolostora is működött, az* ***egyik*** *Veszprémben, a másik Dunapentelén.*
egyik/egyik/MN_NM.default_0

Az 1. ábrán látható a főnevekre (köznevekre és tulajdonnevekre), illetve többesszámú melléknévnek, számnévnek, melléknévi igenévnek annotált elemekre vonatkozó szabályok összefoglalása döntési fában. A fa gyökere az aktuális elem szófaji címkéje. A fa első szintjének élein az ablakban látható első elemen található információk szerepelnek. Például az ablak első elemén látható NU címke következtében az algoritmus a 0 esetet ítéli meg az aktuálisan vizsgált token NOM esetragjának helyére. A fa második szintjén lévő élek az ablak második elemén látható információkat tartalmazzák: ezek csak akkor aktivizálódnak, ha az első elem alapján az algoritmus nem tudott dönteni, és a *default* α esetet illesztette a NOM helyére. Ekkor az ablak második elemén látható bizonyos címkeelemek alapján még van lehetősége egyértelműsíteni az esetet. Az ábrán nem tüntettük fel, de a 0 egyértelműsítése után (a fa második ágának végrehajtása előtt) történik egy lépés, amely független az ablak első elemétől. Ha az aktuálisan vizsgált token semmilyen esetben sem lehet jelöletlen birtokos szerkezet birtokosa, akkor az algoritmus a NOM esetet ítéli meg a szónak, és megtörtént az egyértelműsítés. Ilyen tokenek az *az, ez, mindaz, mindez, aki, ami.* Fontos kitétel, hogy az ablak második elemén látható birtokos személyjel (PS) csak abban az esetben vezet az aktuális elem NOM címkéjének GEN esetragra cseréléséhez, ha az aktuális elemen nincsen birtokos személyjel. Ezzel a *Magyarország* ***kormánya*** *mostani megbízásából* szerkezetek előfordulását zárjuk ki. A *Magyarország* ***kormánya*** *megbízásából*-típusúak az algoritmus számára is megítélhetőek, az ablak első elemén látható birtokos személyjel alapján. A példákban félkövérrel szedett tokenek az aktuálisan vizsgált elemek.

A legfelső élen, a NU társaságában található a NU_MN kategória. Bár ilyen címkét az MNSz2-ben nem találunk, mi fontosnak tartjuk a megnevezését: az *alatti, általi, mögötti* stb. szavak tartoznak ide. Ezek, a névutókhoz hasonlóan, azonnal egyértelműsítik az őket megelőző névszói elem végén a 0 esetet. Hozzájuk hasonló a *című, nevű* tokenek szerepe, ezért azokat is feltüntettük itt[3].

A 2. ábrán látható a(z egyesszámú) melléknevekre és melléknévi igenevekre vonatkozó szabályok összefoglalása döntési fában. Ennél a szabálycsoportnál is fontos kitétel, hogy az ablak második elemén látható birtokos személyjel (PS) csak abban az esetben vezet az aktuális elem NOM címkéjének GEN esetragra cseréléséhez, ha az aktuális elemen nincsen birtokos személyjel. Végül a 3. ábra a számnevekre vonatkozó szabályok összefoglalása.

[3] Megjegyzendő, hogy a *című* és a *nevű* előtti eset inkább a többtagú nevek esetével azonos, de ez további kutatást igényel.

1. ábra: A főnevekre (köznevekre és tulajdonnevekre), illetve többesszámú melléknévnek, számnévnek, melléknévi igenévnek annotált elemekre vonatkozó szabályok összefoglalása döntési fában. A fa gyökere az aktuális elem szófaji címkéje. A fa első szintjének élein az ablakban látható első elemen található információk szerepelnek. A fa második szintjén lévő élek az ablak második elemén látható információkat tartalmazzák.

4. Eredmények

Az algoritmus teljesítményét 1 000 darab tesztmondaton mértük ki. A tesztmondatokat az MNSZ2 szolgáltatta, és azzal a megszorítással éltünk, hogy a mondatban legyen legalább egy finit ige. A kapott 1 000 darab véletlenszerűen kiválasztott mondaton három változtatást eszközöltünk:

– Mivel a tulajdonneveknek jól körülhatárolható, fontos funkciója van az őket megelőző elem esetének egyértelműsítése során, ezért a tesztmondatainkon kézzel annotáltunk minden tulajdonnevet. A többelemű tulajdonneveket _ jellel összekapcsoltuk, és `FN` címkéjűket `TULN` címkére cseréltük.

2. ábra: A(z egyesszámú) melléknevekre és melléknévi igenevekre vonatkozó szabályok összefoglalása döntési fában. A fa gyökere az aktuális elem szófaji címkéje. A fa első szintjének élein az ablakban látható első elemen található információk szerepelnek. A fa második szintjén lévő élek az ablak második elemén látható információkat tartalmazzák.

- A kopulás, vagy finit igét nem – legfeljebb létigét – tartalmazó tagmondatokat töröltük. Ha egy egész mondatot kellett a kopula miatt törölni, akkor ahelyett újat kértünk a korpuszból.
- Az esetlegesen bennragadt, „szemétnek” minősülő elemeket, úgy mint a sorszám a mondat elején, vagy a végén, töröltük a mondattokenekből.

Az így megtisztított 1 000 darab mondat minden egyes, eredetileg NOM címkével rendelkező eleme három elemzést kapott: egyet az algoritmustól, egyet a kézi annotáció során az ablak elemeire támaszkodva, és egyet a kézi annotáció során mint végleges elemzés.

A kétféle kézi annotációval a célunk az, hogy összetettebben mondhassunk ítéletet mind az algoritmus megvalósításáról, mind a mögötte húzódó elméleti

3. ábra: A számnevekre vonatkozó szabályok összefoglalása döntési fában. A fa gyökere az aktuális elem szófaji címkéje. A fa első szintjének élein az ablakban látható első elemen található információk szerepelnek.

megfontolásokról. Az ablak alapján hozott manuális ítélet (az 5-ös példában az α) lényege, hogy kiértékelhessük, mennyiben valósítja meg az algoritmus a tőle elvárt viselkedést. Ha az algoritmus jól teljesít, az azt jelenti hogy az ablakban látható elemek alapján a lehető legpontosabban meghatározza, hogy az adott testes esetrag nélküli elem milyen szerep(ek)et tölthet be ebben a mondatban, és mindezt helyesen végzi. A teljes mondat alapján definiált annotáció (az 5-ös példában a NOM pedig az adott tokenek végső szerepét mondja meg (jelen esetben a *patak* a mondat alanya). Ezzel az algoritmus mögött meghúzódó elméleti elgondolásokról szeretnénk ítéletet mondani: ha ez a kézi elemzés egyezik azzal, amit a kézi annotáció során az ablak alapján mondtunk, akkor jól használható és pontos az eljárásunk; az ablak alapján nagy bizonyossággal meg lehet határozni egy névszó mondatbeli szerepét, anélkül, hogy alaposabban körül kéne néznünk (más szavakkal: a kétfázisú mondatelemzés első fázisában ezek a szerepek jól eldönthetők).

(5) *A **patak** tőlük keletre húzódott.*
az ablak: *patak tőlük keletre*
az algoritmus ítélete: *patak FN.**nom***
kézi annotáció, az ablak alapján: *patak FN.*α
kézi annotáció, a teljes mondat alapján: *patak FN.**nom***

A kézi annotáció során, első körben, tehát az ablak alapján való döntésnél, a következő címkéket kaphatták a testes esetrag nélküli tokenek:

- NOM: nominatívusz
- GEN: birtokos
- 0: esetrag nélküli (névutó előtti névszó[4], vagy más névszó módosítója)

[4] A névutó előtti névszó végén lévő testetlenség pontos jelentése kérdéses: jogos elképzelés lenne egy NOM jelenlétét feltételezni, az esetragos névszót vonzó névutók

- α: nem eldönthető; a főnevek *alapértelmezett* értékét kapja (NOM-má vagy GEN-né egyértelműsödhet később)
- *default_0*: nem eldönthető; az `NPMod` elemek *default* értékét kapja
- VOK: vokatívusz esetű
- *postag_hiba*: valamelyik vizsgált elem hibás szófaji címkét kapott, ezért rossz az elemzés (például melléknévi igenévnek címkézett ige esetében)
- többtagú nevek esete (ld. pl. 2a)

A kézi annotáció második körében, a teljes mondat alapján hozott ítéleteknél a következő címkét kaphatták az esetegyértelműsítésre váró névszók:

- NOM
- GEN
- 0
- VOK
- többelemű nevek esete
- α vagy *default_0*, abban az esetben, ha a teljes mondat kétértelmű

A tesztmondatokban összesen 125 olyan token volt, amelyeknél vagy az adott token annotációja volt hibás (például `NOM` esetragú melléknévi igenévnek volt címkézve egy ige), vagy az ablakban lévő egyik vagy másik szóalak (például az őt követő melléknévi igenév volt igének címkézve). Ezeket az eseteket nem javítottuk, nem számítottuk az értékelésnél. Szintén nem került bele a kiértékelésbe a 34 vokatívuszi esetű névszó, illetve a 45 darab *többtagúnév-esetű* token.

A kiértékelés szabályait és a kategóriákat az 1. táblázat tartalmazza. *True positive* (TP), *false positive* (FP) és *false negative* (FN) kategóriákat állapítunk meg. Az egyes oszlopokat a következőképpen kell értelmezni: ha a „kiértékelendő eredmény" oszlopban látható értékre a „sztenderd" oszlop adott értéke illeszkedik, akkor ez egy TP, FP vagy FN találat, attól függően, melyik sorban található ez a párosítás. Ez a megfeleltetés igaz akkor is, mikor az algoritmus eredményét hasonlítjuk a kézi, csak az ablakot figyelembe vevő annotációhoz, illetve akkor is, mikor a kézi, csak az ablakot figyelembe vevő annotációt a teljes mondatot figyelembe vevő kézi annotációhoz hasonlítjuk. A TP eredmények a teljes egyezések. FP eredménynek a túlspecifikálást tekintjük: ha például az algoritmus egy elemről azt állítja, hogy nominatívusz, de a kézi annotáció szerint az ablak alapján még nem mondhatna ilyet, csak egy *default* α-t, akkor az FP eredmény. Viszont ha alulspecifikál, tehát *default* értéket ad egy elemnek, pedig az ablakból eldönthető lenne pontosabban is, az FN.

4.1. Az ablak kiértékelése

A 2. táblázatban látható eredmények azt mutatják, hogy elméleti alapvetésünk, azaz a kételemű ablak alapján történő egyértelműsítés milyen pontosságot és fe-

---

példáinak analógiájára, pl. *a kerítésen kívül – a kerítés mellett*; ugyanakkor a névutós és esetragos névszók analógiája ennek ellenkezőjét sejteti, azaz hogy a névszó végén valóban nincsen semmi, ahogy az esetrag előtt sem feltételezünk semmit a szótőn, pl. *az asztalon – az asztal alatt*. Mi most ez utóbbi analógiát tekintjük mérvadónak, és a névutó előtti névszók végén nem feltételezünk NOM vagy más esetragot.

kategória	a kiértékelendő eredmény	a sztenderd
TP	NOM GEN α *default_0*	NOM GEN α *default_0*
FP	NOM GEN 0	α α *default_0*
FN	α α *default_0*	NOM GEN 0

1. táblázat. A kiértékelés szabályai.

dést eredményez, ha a testes esetrag nélküli névszói elemek teljes mondat alapján történő esetegyértelműsítéséhez hasonlítjuk. Jól látszik, hogy a pontosság 97,73%-kal igen magas. Alapvető célunk az volt, hogy precízen döntsünk, ne kelljen a mondatelemzés egy későbbi fázisában korrigálni korábbi, később tévesnek bizonyuló ítéleteinket. Ennek a feltételnek sikerült megfelelnünk, ezt a magas pontosság szemlélteti.

TP	FP	FN	pontosság	fedés	F-mérték
1 590	37	761	97,73%	67,63%	79,94%

2. táblázat. Az ablak alapján történő kézi annotálás eredményeinek összehasonlítása a teljes mondatot figyelembe vevő kézi annotálás eredményeivel.

A fedés ugyanakkor csak 67,63%: ez legfőképpen a *false negative* találatok magas száma miatt van, ami bővebb magyarázatot igényel. Azt az esetet tekintettük FN találatnak, ha az ablak szerinti kézi annotációnál az *alapértelmezett* értéket kapta egy elem, a teljes mondat alapján azonban már specifikusabb eredményre jutottunk. Tehát az alulspecifikáltságot tekintjük fals negatív eredménynek. A 3. táblázatban látható a FN találatok részletezése.

Ugyanakkor látni kell azt is, hogy ezek nem feltétlenül hibák - a főnevek *alapértelmezett* esetének számító α pontosan a NOM és a GEN esetragot fogja össze: ezekben az esetekben az történik, hogy az ablak alapján még nem állapítható meg egyértelműen, hogy az α esetragú névszó alanya vagy egy birtokos szerepű tagja a mondatnak, ezt csak a tágabb kontextus segítségével lehet eldönteni. Izgalmasabb azonban a *default_0*-k esete: kiugróan magas azoknak a melléknevek és melléknévi igeneveknek a száma, melyeknél az *alapértelmezett* esetet ítéltük meg az ablak alapján, de az esetük a mondat alapján 0. Sőt, összesen 6

hibatípus	hibaszám
FN	761
NOM helyett α	186
GEN helyett α	56
0 helyett *default_*0	519

3. táblázat. A kétféle kézi annotáció összehasonlításakor megfigyelt *false negative* eredmények. Az egyes sorok azt mutatják, milyen eset helyett milyen esetet egyértelműsített (alulspecifikálva) a csak az ablak alapján történő kézi annotáció.

olyan eset fordult elő, hogy az ablak alapján *default_*0-nak ítélt esetet a mondat teljes egésze NOM-ként egyértelműsítette, minden más alkalommal 0 lett ezekből. Mit értünk ez alatt? A 6a példában szemléltetjük, hogy az ablak alapján még elképzelhetőnek tartjuk, hogy az adott token a mondat alanya lesz, vagy egy jelöletlen birtokos szerkezetben a birtokos, de a teljes kontextus (6b) egyértelművé teszi, hogy ez egy főnevet módosító elem.

(6) a. *telepített mintegy negyven*
ablak alapján: telepített telepít IGE._MIB.*default_*0

b. *Kétéves koromban elvesztettem anyai nagyszüleimet, s velük együtt a szülőfalumból Magyarországra* ***telepített*** *mintegy negyven családban szinte minden rokonomat.*
mondat alapján: telepített telepít IGE._MIB.0

Ezeknek az eredményeknek az összefoglalásaképpen elmondhatjuk, hogy a magas pontosság megfelel az eljárással kapcsolatosan támasztott elvárásainknak. A fedés javítható lenne (67,63%-ról 89,7%-ra) azzal, ha a melléknevek és melléknévi igenevek az *alapértelmezett* érték helyett automatikusan a 0 esetet kapnák meg az ablak alapján.

4.2. Az algoritmus kiértékelése

A 4. táblázatban az algoritmus teljesítményének kiértékelése látható. Ebben az esetben a gépi esetegyértelműsítést hasonlítottuk össze a csupán az ablakot figyelembe vevő kézi annotációval.

TP	FP	FN	pontosság	fedés	F-mérték
2 220	63	125	97,24%	94,67%	95,94%

4. táblázat. Az algoritmus teljesítményének kiértékelése a csak az ablakot figyelembe vevő kézi annotáción.

Mind a pontosság (97,24%), mind a fedés (94,67%) kiemelkedően magas: az algoritmus teljesítménye megbízható, amit az ablak alapján egyértelműsíteni lehet, azt egyértelműsíti, de nem specifikál olyan eseteket, amiket még nem lehetne. Az egyetlen elemzésre számot tartó érték talán az 5. táblázat utolsó sora, ahol az algoritmus túlzottan alulspecifikál; *default* értéket illeszt a melléknevekre, melléknévi igenevekre, holott 0-t kéne. Ez az aleset párhuzamba vonható a 2. táblázat kapcsán már említett, és a 6a és a 6b példákban bemutatott jelenséggel. Az 5. táblázatban szereplő 112 eset tovább erősíti azt a sejtést, hogy érdemes lenne az `NPMod` kategóriájú elemekre *default_*0 helyett mindig 0-t illeszteni.

hibatípus	hibaszám
FN	125
NOM helyett α	9
GEN helyett α	4
0helyett *default_*0	112

5. táblázat. Az algoritmus teljesítményének kiértékelésekor megfigyelt *false negative* eredmények. Az egyes sorok azt mutatják, milyen eset helyett milyen esetet egyértelműsített (alulspecifikálva) az algoritmus.

5. Összegzés

A cikkben bemutattunk egy algoritmust, mely a testes esetragot magukon nem viselő, ezért automatikusan `NOM` címkével ellátott főnevek, melléknevek, számnevek és melléknévi igenevek esetegyértelműsítését végzi kizárólag az aktuálisan vizsgált token és az azt követő két szó, az ablak információi alapján[5]. A korpuszmegfigyeléseink alapján körvonalazódó szabályrendszert implementáltuk, majd kiértékeltük. Eredményeink szerint igen magas pontossággal és fedéssel teljesített. További feladataink közé tartozik a vokatívusz eset feltérképezése és egyértelműsítő algoritmusának megírása, illetve a többtagú nevek belsejében szereplő elemek esetének megvizsgálása. Ezeket az ablakban szereplő információk mellett a korábban már feldolgozott elemekre támaszkodva szükséges egyértelműsíteni. Mindezt pedig követheti az állítmány-esetű névszók detektálása a tágabb kontextus alapján.

Hivatkozások

1. Prószéky, G., Indig, B.: Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. Alkalmazott Nyelvtudomány **15**(1-2) (2015) 29–44

[5] Az eljárás implementciója, a tesztfájl és az annotált fájl elérhető itt: `https://github.com/ppke-nlpg/nom-or-what`.

2. Prószéky, G., Indig, B., Vadász, N.: Performanciaalapú elemző magyar szövegek számítógépes megértéséhez. In Kas, B., ed.: „Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére. MTA NYTI, Budapest (2016) 223–232
3. Frazier, L., Fodor, J.D.: The Sausage Machine: A New Two-Stage Parsing Model. Cognition **6**(4) (1978) 291–325
4. Vadász, N., Kalivoda, Á., Indig, B.: Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével. In Vincze, V., ed.: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017), Szeged, Szegedi Tudományegyetem Informatikai Tanszékcsoport (2017) 3–12
5. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Calzolari, N., et al., eds.: Proceedings of the 9th International Conference on Language Resources and Evaluation, May 26–31, 2014, Reykjavik, Iceland, ELRA (2014) 1719–1723
6. Vadász, N., Indig, B.: A birtokos esete az ablakkal. In Scheibl, G., ed.: LingDok: nyelvész-doktoranduszok dolgozatai. Szegedi Tudományegyetem (2017) Megjelenés alatt.
7. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In Matoušek, V., Mautner, P., Pavelka, T., eds.: Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings, Berlin, Heidelberg, Springer Berlin Heidelberg (2005) 123–131

Tanulságok magyar mondatellenőrző nyelvi adatainak átvitelénél

Naszódi Mátyás, e-mail: naszodim@morphologic.hu

MorphoLogic, 1122 Ráth György utca 36.

Kivonat A mondatellenőrzők a mondatok nyelvtani helyességét nem képesek eldönteni, csupán hibás eseteket keresnek a mondatban, melyekre nem derül fény a szóellenőrzésből. Jóságát abban mérik, hogy a valós hibák hány százalékét ismeri fel, illetve hányszor riaszt feleslegesen. Az elérhető minőség függ az előállítás módjától. A leíró formalizmusok és meghajtó programok lehetőségei korlátosak. Jelen cikk kitér két eszköz képességeire, és bemutatja, hogyan (nem) lehet portolni az egyik formalizmusból a másikba a nyelvi tudást. A próbálkozás egyéb nyelvészeti tanulságokkal is szolgál. Rámutat morfológiai leírások minőségi jellemzőire és nyelvi jelenségekre. Végül keresi a további fejlesztés irányát.

Kulcsszavak: mondatellenőrzés, statisztika, nyelvtan, nyelvminőség

1. Bevezető

A mondatellenőrzők a szóellenőrzők után nem sokkal, a 80-as években megjelentek. Angol, majd francia, spanyol nyelvekhez készítettek – először a Grammatik, később más műhelyek is – grammar checker-eket. Magyarra a 90-es évek közepén jelent meg, először a MS Office részeként. A késést nyelvünk összetettsége okozta. Nem csak a szótan okoz nehézséget, hanem nyelvtanunk is lényegesen elüt az indoeurópai nyelvek szerkezetétől.

A mondatellenőrzők nem azt ellenőrzik, hogy a leírt mondat megfelel-e a nyelv formálisan megadott nyelvtanának. Formálisan megadott szabályokba belefér olyan mondat, mely szemmel láthatóan (humán olvasatra) helytelen. Az angol nyelvnél talán felismerhető lehet a hibák mérhető hányada ezen a módon, de az angolnak egyrészt kötött szórendű a nyelvtana, másrészt az angol mondatok nehezen tűrik a hiányt. A szabad szórend és az erősen hiányos mondatok megengedése reménytelenné teszik a globális nyelvtan módszerét. Inkább negatív szabályokat adnak meg szép számmal, ahelyett, hogy azt írnák le, mi a helyes.

1.1. Létező programok

Magyar nyelvre három ilyen célú programot ismerek. Az egyik a MorphoLogic 95-ben termékké váló Helyesebb programja[1] - ezt használta korábban a Microsoft az Office-ban, a másik a Németh László és társai által 2009-ben készített Lightproof[2], mely a LibreOffice-ba integrálható, és harmadik az utóbbi időben megjelent Microsoft Office 365-be beépített nyelvhelyesség-ellenőrző. Más nyelvekre több is létezik, de csak kevés üti meg a hasznos szintet.

A programok minőségét nem vetem összehasonlító teszt alá, mint ahogy a helyesírás-ellenőrzőknél szokásos, mert a rangsorban objektív okokból egyértelmű: a Helyesebb a legjobb, ezt követi a Lightproof a nyelvi információ átvétele után is, majd használhatatlanul a sor végén a Microsoft saját fejlesztése.

1.2. Mondatellenőrző mint a szóellenőrző kiterjesztése

Ezeket a programokat egyesek a helyesírás-ellenőrzők kiterjesztésének tekintik. Amikor a hiba szószinten nem fedezhető fel, a szövegkörnyezet rámutathat az elírásra. Különösen a magyar nyelv szempontjából fontos, ahol a nagy szóalakszám miatt a kellő lefedettség érdekében ritkább szóalakok is bekerülnek a helyesnek ítéltek közé. Ráadásul nyelvünk igen sűrű: a szavak közt az átlagos távolság kicsi.[3] Emiatt nem árt az utólagos szövegkörnyezeti ellenőrzés. De sok más célra is felhasználható.

2. Technikai áttekintés

2.1. A hibák lefedettsége

A mondatellenőrző nem a helyességet ellenőrzi, hanem a helytelenséget fedi fel nyelvtani, még gyakrabban statisztikával felderített szabályok alapján. Hogy egy szó helyes-e, arról a helyesírás-ellenőrző is hasonlóan dönt. A helyes szóalakokat morfológiai szabályokkal adjuk meg, de sokszor ezen felül beleszól a szóstatisztika is. Így lehetséges, hogy szabályos szavakat, szóalakokat elutasít.

Míg szószinten nagy lefedettséget kis mellélövéssel el lehet érni, mondatszinten ezek az arányok eltérnek, mert a lehetséges esetek száma több nagyságrenddel nagyobb. Míg szóalakból egy nyelven több milliárd helyes van (nem ragozó nyelveknél ez lehet csupán több százezer), az ezekből alkotott helyes mondatokból ennek exponenciálisszorosa. A várható helytelenek számánál az arány jobban eltér. Ahhoz, hogy a (mondatszintű) hibák 90 százalékát jelezzük, sokkal nagyobb adatbázisra lenne szükség, mint a szóellenőrzőnél, ahol a kilencvenszázalékos követelmény természetes.[4] Ekkora adatbázist előállítani képtelenség. A minőségi követelmény szintjét máshol kell meghúzni. Ha a mondatszintű hibák több mint felét felderíti egy ellenőrző, akkor már hasznos.

Ha tökéletes a szótan adatbázisa, akkor a szóellenőrző helytelen alakokat nem fogad el. Ezt még felülbírálhatja a statisztika. Egy ellenőrző sem tartja helyesnek a képzett *tanit* szót, és több szóellenőrző megugatja az *egyenlőre* alakot. Az első esetben teljesen jogos a döntés. A *tanit* alak egyetlen természetes szövegkörnyezetben sem helyes. Az utóbbi előfordulása az esetek kicsi, de nem elhanyagolható részében kifejezetten megfelelő: *Nem szabták egyenlőre a nadrág két szárát.*

Egy nyelvi ellenőrző jóságát durván két paraméterrel lehet mérni.

- Hibás esetek hány százalékában jelez
- Hibátlan esetek hány százalékában jelez feleslegesen.

Tegyük fel, hogy egy kérdéses szóalak száz előfordulása közül statisztikailag egyszer helyes és kilencvenkilencszer hibás a szövegben, tehát a szó ugyan helyes, de nem a szövegkörnyezetben. Ekkor 1000 előfordulás közül 990 hibás. (Persze egy literátusabb szerzőnél ez a szám kisebb, míg egy nyelvérzékkel kevésbé megáldott írónál lehet nagyobb.)

Ha a szóellenőrző adatbázisából kivesszük a szót, akkor ezzel nő a hibák lefedettsége, de vele nő a tévedések száma is. Ekkor 1000 előfordulás mindegyikét jelzi, ebből 990 esetben jogosan, míg a 10 helyes alak esetében téves a jelzés. Ha viszont a szóellenőrző elfogadja a szót, de környezeti ellenőrzéssel az esetleges helyes előfordulás 95 százalékban kiszűrhető, akkor jobb eredményhez jutunk, ha a szóellenőrző elfogadja az alakot, és csak a mondatellenőrző jelez hibát a környezet függvényében. Maradva az *egyenlőre* szónál, ha a környezetben találunk *hoz*, *vág*, *szab*, *tesz* igét, akkor nagyobb a valószínűsége, hogy helyes volt a szó, sőt, a statisztikák alapján a helyes előfordulások nagyobb százalékát le is fedjük. Így az 1000 előfordulásnál 990 riasztásból legfeljebb egy a téves riasztás, míg a 10 helyes esetből alig lesz hibásnak minősített. Pár hibás esetet nem veszünk észre, de legfeljebb 1-nél kapunk felesleges jelzést.

2.2. Egyszerű hibák kimutatása

Lokális ellenőrzéssel sok stiláris hibát deríthetünk fel. Például durva, obszcén kifejezéseket, vagy olyanokat, melyeket az amerikaiak inkorrekt szóhasználatnak neveznek. Magyarországon az idegen szavak szűrésére lehet igény. Ezek a szűrések nem csak arra jók, hogy stilárisan tegyük szebbé írásunkat, de időnként arra is, hogy nem kívánt elütések következtében keletkező hibákat is felderítsünk. A magyar nyelv sűrű, emiatt elütéssel könnyen keletkezhet helyes szó. Gondoljunk arra, ha a *szár* szóról lemarad az ékezet. Ettől a mondat nyelvtani szempontból még helyes marad, csak stiláris hiba keletkezik, de előfordulhat, hogy épp ettől válik nyelvileg hibássá a mondat.

Ezért is érdemes nyelvtani következetlenségeket is szűrni, bár ilyen ritkán fordul elő. Leginkább akkor, ha a szövegszerkesztővel copy-paste segítségével szabjuk-varrjuk az írást. Viszont lehetséges – és ez a gyakoribb eset –, hogy egy elütés miatt nyelvtanilag hibás (rész)szerkezet keletkezik: *A patak partján három lapút találtam.*

2.3. A lokális szintaxis alkalmazásai

A hibák egy része felderíthető lokális szintaxis segítségével[5], de időnként ez nem elég. Főként a szabad szórend miatt, de más esetekben is távol eső szavak összefüggéséből lehet rájönni a hibára.

Lokális szintaxist használnak szó-egyértelműsítésre[6][7]. Itt viszont statisztikai alapon döntenek, ami esetünkben nem jó módszer – elfedheti a hibát, ahelyett, hogy rámutatna. Magas találati aránnyal lokális szintaxist a névfelismerésnél alkalmaznak.
Hasonlóan szavak előfordulásával, illetve lokális összefüggések felismerésével operálnak szövegek minősítésekor, például a szövegek pszichológiai kiértékelésénél[8].

A nagy különbség ott van, hogy a szövegkiértékelés eredménye statisztika. Tehát a jelenségek előfordulásának mennyisége a lényeges, nem az egyedi jelenség. Emiatt ha egyszer-egyszer, akár tíz százalékban téved az eseti elemző, attól a teljes szövegstatisztika lényegesen nem módosul. A szövegkorrekciónál viszont a felismerésben elkövetett ily mértékű tévedés használhatatlanná teszi az eszközt.

2.4. A globális szintaxis alkalmatlansága

A mondatok globális formális leírása alapján hibák lokalizálását csupán a formális, például a programnyelveknél alkalmazzák, mert a jól tervezett programnyelvek egyrészt determinisztikusan elemezhetők, másrészt szintaktikus hiba esetén jól behatárolható a hiba helye. Több értelmű nyelveknél – a természetes nyelvek ide tartoznak – ez az út nem járható. Még a könnyebben elemezhető angolnál sem a teljes mondat szintaxisából indulnak ki. Helyette hibás szituációkat adnak meg, általában szűk látókörű, részszintaxis segítségével. Az egyes szabályok, szabálycsoportok függetlenek egymástól, és ha valamelyiknek eleget tesz egy részszöveg, azt hibásnak illetik, illetve az elemzésnek megfelelően javítási tanáccsal is szolgálnak.

Teljes mondatelemzőt használnak a szabályalapú fordítóprogramok. Ott viszont feltételezik, hogy a mondat helyes, illetve helytelen mondatokat is igyekszenek megelemezni, hogy lefordítsák. Emiatt az ott használt szintaxis nem pontos, alkalmatlan akár a hiba felismerésére, de a hiba pontos helyének meghatározására végképp használhatatlan.

2.5. A kompromisszum – parciális szintaxis

Miután teljes mondatelemzésről szó sem lehet, ezért lokális szintaxist, vagy ami hasznosabbnak bizonyult, a Helyesebbnél bevezetett parciális szintaxist alkalmaznak. Ebben az esetben a lokális vizsgálaton kívül lehetőség van távolabbi szavakra is rákérdezni, a mondat közbünső részét mintegy átugorva. A leírónyelv interpretálása itt is gyors, lehetőleg determinisztikusan interpretálható kell, hogy legyen. A lehetséges megoldás szavak feletti reguláris attribútumnyelvtan, mely legfeljebb sekély szintaxist használ.[1]

3. Mit ellenőrizünk?

3.1. A cél határozza meg az eszközt vagy az eszköz a célt?

Ha a cél határozza meg az eszközt, az igény az, hogy minél több – főleg gyakran elkövetett – hibát derítsünk fel. Helyesírásunknak két nehezen emészthető szabálycsoportja van, melyet sokan elvétenek. Az egyik a szóösszetétel, a másik a vesszőhasználat. A szabályok az első esetben erősen szemantikafüggőek, a másik esetben a szabályok (majdnem) egyértelműek ugyan, csak az nem világos, mitől lesz egy rész tagmondat, és mitől csak egy részkifejezés. Szóval mindkét eset nehéz dió.

Ha az eszköz határozza meg a célt, akkor az egyszerűen, de biztosan felderíthető hibákat érdemes feltárni. Az egyedi hibás esetek lehetnek ritkák, de tömegükben gyakoribbak, mint a nehezen felismerhetőek.

3.2. Mit ellenőrizhetünk?

A különböző nyelvekre megvalósított mondatellenőrzők téma szerint a következő osztályozásokat alkalmazzák:

- Nyelvtani hibák
  - Egyeztetések (alany-állítmány; névelő, szám, személy, nem; névelő hangtani egyeztetése; vonzat ...)
  - Koordinációs szavak, kötőszavak helyes használata
  - Vesszőhasználat (mondat és mondatrész vesszőhiány, -többlet)
  - Kisbetű vagy nagybetű
- Összetett szavak (amikor egybe és külön is helyesek a szavak)
  - Egybeírás (szemantikai és nyelvtani összetételek)
  - Különírás
  - Kötőjeles összetételek (akár egybe, akár külön írták)
- Gyakori elírások
  - Elütés következtében nem odaillő szó (A szóellenőrző kiterjesztése)
  - Hibás szóhasználat (A közhasználatban elterjedt hibás változat)
- Stiláris szűrés
  - Vulgáris, obszcén szavak, kifejezések
  - Inkorrekt, sértő szavak, kifejezések (Ez az, amit az USA-ban szigorúan vesznek, de magyarban is lehet találni szűrendőt)
  - Idegen szavak, kifejezések (Ha nem elterjedt szó, akkor illik átírni. Szakszövegekben nem érdemes szűrni)
  - Köznapi szöveg. (Egyes nyelveken eltér a beszélt és a szép írott nyelv.)
  - Vidékies szöveg, zsargon, rétegnyelv ...
  - Nem preferált szerkezetek - pl. angolban passzív szerkezet
- Formátumok
  - Dátum, idő és egyéb mértékek (a dátumnál a pontok használata)
  - Megszólítás
  - Címzés
- Egyebek
  - Szóközök, tabulátorok (felesleges és hiányzó szóközök szűrése)
  - Halmozott írásjelek (Ha több írásjel követi egymást, ezek nem minden szekvenciája helyes)
  - Zárójelek, idézőjelek (Egymásba skatulyázás, helyes jelek, jelpárok használata.)
  - Kötőjelek, gondolatjelek. (Nem közismert, mikor melyik jelet ildomos használni)
  - Internet formátum
  - Számformátum (pl. tizedespont, vagy -vessző)
  - Ligatúrák, alternatív kódolások (Többek között a hármaspont, lebegő ékezetek... melyek nem feltétlen látszanak az írott szövegben)

A csoportosítás és a tartalom nyelvenként és meghajtónként eltér. Egyes típusok szűrése az egyik nyelven könnyű, míg a másikon majdnem lehetetlen.

4. Technikai kérdések

4.1. A nyelvi adatbázis nyelve

A technikai feltételek lényegesek a mondatellenőrzőnél. A bővíthető eszközök nem nélkülözhetnek egy jól definiált formális nyelvet, melyen megfogalmazhatók a szabályok. A leírás nyelve nem feltétlen közvetlen vezérli az ellenőrzőt. Általában egy előzetes fordítási procedúrával keletkezett közbülső (ember által nehezen emészthető) kódban interpretálódik a végrehajtandó ellenőrzés. Ha a szótan és a nyelv szintaxisa egyszerű szerkezetű, akkor a leíró szabályokat közvetlen karakterekre épülő reguláris kifejezéssel is lehet interpretálni. A legtöbb esetben nem ezt teszik. A magyarban a leírás és a futó interpreter hivatkozik egy szóelemzőre, mely a mondatellenőrzőn belül egy külső modult képez.

A formalizmusban nem csak a hibát kell leírni. Meg kell jelölni a hiba pontos helyét és lehetőség szerint a korrekció (csere) módját is. Van, mikor a hiba felismeréséhez nagyobb területet kell átvizsgálni, mint amit érdemes a felhasználó elé tálalni. Erre is lehetőséget kell biztosítania a leírás nyelvében.

A leíró formalizmusban gyakran hivatkoznak szóalakra, szótőre, morfémákra, szótani kategóriákra. Ha elegendő a felszíni alakok felismerése, akkor általában reguláris kifejezéssel adják meg a szabályokat. Ez a ritkább eset. A nyelvi kategóriákat felhasználó szabályok is csak sekély (lapos) szintaxist használnak, bár néha jó lenne ennél összetettebb, környezetfüggetlen nyelvtan is.

Ha hiba felismeréséhez javítási ajánlat is szerepel, márpedig általában erre igény van, akkor nemcsak szótani elemzőre, hanem generálóra is szükség van. (Kivételt képeznek a gyengén vagy nem toldalékoló nyelvek.)

A szabályok kis része nem hivatkozik szóelemzésre. Ilyenkor a szabályból pontosan követhető a felismerés mikéntje. Az esetek többségénél praktikus a szóelemzés és az abból nyert tulajdonságokra való hivatkozás. A magyarban nem felsorolhatóak az egy kategóriának megfelelő alakok. A szóelemzőre való hivatkozás növeli a mellélövések esélyét, mivel a morfológiai modulban is lehetnek hibák. Az ilyen hibák elkerülése érdekében fontos a szóelemző pontossága. Hasonló minőségi követelmény lép fel a szógenerátorral szemben is.

A leírás nyelve olvashatónak, karban tarthatónak kell, hogy legyen, mert a szabályok száma nagy, ráadásul hathatnak is egymásra, emiatt mindig igény van javításra, bővítésre.

Tehát az egyes szabályoknak a következő információkat kell tartalmaznia:

- A hiba leírása, mely szóalakra, morfológiai egységekre, ezek szekvenciáira hivatkozik.
- A hiba hatásköre – mit kell kijelölni hibásnak a megjelenítéshez.
- Javítási javaslat: mire cseréljük a megjelenített résztartomány – ebben a felismert rész egyes elemei és az ebből generálandó szavak is szerepelhetnek.
- A hiba típusa: ez csak arra kell, hogy bizonyos osztályok ellenőrzését engedélyezhessük, letilthassuk.
- Magyarázat: mellékes információ, mely nem mellékes a felhasználónak, hogy jobban megértse, mire gondol a nyelvész.

4.2. A végrehajtó motor

Mint említettem, általában a formalizmusban leírt nyelvi információ egy fordítás után lesz használható ellenőrző. Ez vagy egy közbülső nyelv interpretálása, de nem ritka eset, amikor a nyelvtan egy része adatként interpretálódik, míg más részét a fordító közvetlenül programnak generálja. Ez utóbbira példa a csehek grammar-je[9], amit a Microsoft számára készítettek. A Lightproof is alkalmaz vegyes megvalósítást. Gyakoribb viszont az, amikor a teljes anyag interpretálható közbülső formáját használják.

4.3. Sebességi követelmény

A végrehajtásnak kellően gyorsnak kell lennie. Egy mondat ellenőrzése, legalábbis a hiba felderítése egy másodpercnél nem vehet többet igénybe. Mivel a szabályok száma ezrekben mérhető, és egy hosszabb mondatban sok helyen kell próbálkozni a lokális – parciális szabályok érvényesülésével, még a mai gyorsabb gépeken is lehet kritikus az időigény. A tapasztalat szerint jól megfogalmazott szabályok esetén a program a szótani elemzéssel tölti a legtöbb időt. Emiatt itt kell spórolni. Az elemzések gyorsítótárba való gyűjtése általában meghozza a kívánt sebességet, de nem mindig.

A morfológia megkerülése is sokat segíthet. Ha valamit tisztán alaki módon is meg lehet fogalmazni, akkor ez segít. Az is segíthet, ha a morfológiai elemzés cash-elése alatt olyan attribútumokat is letárolunk, melyek az elemzésnél gyakran kellhetnek (szófaj, eset, igerag. . .), hogy ezeket ne kelljen a morfológiai elemzésből újra meg újra kihámozni.

Ha a felismerés két részre bontható – egy durva gyorsra és egy részletesebb kiértékelésre –, akkor érdemes kettébontani. Ezzel megint csak csökkenthető a terhelés, hisz a hiba nagy valószínűséggel kizáródik az első lépésben, és ilyenkor a másodikra már nincs szükség.

Magyarban gyakori a többértelmű szó. Próbálkoztam szabályalapú egyértelműsítéssel, azt várva, hogy ha pár helyen csökkentem az alternatívákat, akkor nő a globális sebesség. Az egyértelműsítés mechanizmusa hasonló volt, mint a hibák felismerése. Többnyire lokális szabályok alapján döntöttem. A statisztika alkalmazása itt azért veszélyes, mert félreviheti a későbbi hibakeresést. A sebességnövekedés nem volt érzékelhető. Ahol szükség volt rá, inkább a hiba felismerésébe öntöttem az egyértelműsítéseket.

5. A portolás

Miután a Microsoft úgy döntött, hogy az összes nyelvi modult házon belül valósítja meg, ezen az úton a nyelvi tudásunk felhasználása lényegesen csökkent. A felhalmozott nyelvi tudást viszont nem akartam, hogy kárba vesszen. Szerencsére ekkorra már megjelent a Lightproof, melybe reményeim szerint sok mindent át lehet menteni. Mindkettő támaszkodik szóelemzőre, és leírási formalizmusukban sok a hasonlóság. Például a magyarázat rész szinte megegyezik. Ezzel nem is foglalkozom a cikkben.

5.1. A két formalizmus összevetése

A szabályok mindkét megvalósításban osztályokba sorolódnak. Az osztályok a Helyesebbnél nem módosíthatóak, a Lightproofnál adatvezérelt – viszont vigyázni kell, hogy a szabályban jelzett osztály szerepeljen a deklarációban.

A szabályok a Helyesebbnél csoportokba vannak összegyűjtve – minden csoport, mint egy fejezet először a csoport fejében deklarálja a hiba típusát, a hallgatólagos hibamagyarázatot, és típusonként tér el a hiba szkópja (kijelölendő területe) és a javításnak a módja. Például a különírási osztályban a kijelölendő mindig a hiba feje.

A Lightproof esetén minden szabály maga tartalmazza ezeket az információkat, tehát a szabály vezérli, hogy mit kell aláhúzni, és mi legyen a magyarázó szöveg.

5.2. A Helyesebb hibaosztályai

- Névelőegyeztetés – *a, e* vagy *az, ezen*
- Vesszőhiány – felesleges vagy hiányzó mondat és mondarész vessző
- Lehetséges szóösszetétel – nyelvi szabályok alapján felismert egybeírás
- Létező szóösszetétel – szótár alapján felismert egybeírás
- Kötőjeles szóösszetétel – ha kötőjellel kell egybe írni
- Hibás szóösszetétel - egybe írt, de külön szóba írandó
- Helyi szerkezeti hiba, hiány – legtöbbször valamilyen alany esetű rész hiánya
- Téves szóhasználat – amit sokan rosszul tudnak
- Trágár szavak szűrése
- Durva szavak, kifejezések szűrése
- Idegen szavak szűrése
- Általános szerkezeti hiba – ilyet is felismerek, de ritkán
- Nagy kezdőbetűvel írandó – mondat eleje, tulajdonnevek
- Kis kezdőbetűvel írandó
- Szóközök, írásjelek ellenőrzése – halmozott szóköz, halmozott írásjelek, tapadó írásjelek

Ezeket az osztályokat minden gond nélkül deklarálhattam a Lightproofban. Van még két rejtett osztály. Az egyik a hagyományos helyesírás-ellenőrző – ezt a Microsoft nem használta – és egy jelzés túl hosszú mondat esetére.

5.3. Lényegi különbségek a formalizmusban

A szabályok szintaxisa hasonló. Egy szabály négy részből áll: gyors felismerési minta, egyéb feltételek, javítási opciók, magyarázat. Azonban funkcióban a részek eltérnek.

rész	Helyesebb	Lightproof
fej, primer ellenőrzés	egyetlen szó: szóalak, morféma(sorozat), kategóriafeltétel	akár több szavas karakteres reguláris kifejezés
további, másodlagos feltételek	szóalakot, HUMOR kategóriákat használó (szóalak, nyelvi tulajdonságok, morfémaszekvenciák) több irányú reguláris kifejezés	az osztályra, a fej reguláris részkifejezéseire és Hunmorph elemzésekre vonatkozó kifejezések a környezeten belül, ezek logikai kifejezései
feltétel elemeinek szekvenciája	a szintaxisban tetszőleges irány és szekvencia kijelölhető	csak jobbról balra szabályok, illetve közvetlen környezeti elemek
távolhatás	irányított, tetszőleges	irányítatlan, reguláris
hiba kijelölése	osztálytól függő, program vezérli a másodlagos feltételek szekvenciája alapján	kijelölhető a fej reguláris kifejezése alapján
javítási ajánlatok	kulcsszavak, hivatkozás a megtalált részekre, generálás a fej elemzése alapján	kulcsszavak, hivatkozás a megtalált részekre, generálás minta alapján
generátor	képzőket is	csak ragokat, jeleket
magyarázat	közvetlen magyarázat és opcionálisan részletes magyarázatra való hivatkozás, mely nem itt, hanem ettől függetlenül tárolódik – lehet hivatkozás a MHSZ-ra	közvetlen magyarázat és opcionálisan részletes magyarázat – hivatkozás weboldalra, legtöbbször a MHSZ-ra
formális hiba esetén	teljes konzisztencia ellenőrzés fordításkor	nem minden hiba szűrhető ki fordításkor

5.4. Különbségek a motorban

Helyesebb	Lightproof
8 bites kód	UNICODE
C-ben megírt interpreter	Python-alapú interpreter
Felhasználó nem bővítheti rutinokkal	Felhasználó bővítheti P rutinokkal
Prioritás az osztályok között	Nem ismert a prioritás vezérlése
Szabálykölcsönhatás kiküszöbölése	Nincs rá lehetőség

5.5. A portolás menete

Kiindulva a működő Lightproof szabályokból, először módosítottam a hibaosztályokat és a hozzá tartozó dialógust, a Helyesebb szabályokat megkíséreltem konvertálni Lightproof szabályokra. Először az egyszerűbbeket, majd az egyre összetettebbeket.

Példa kézi kódolásra, mikor a reguláris kifejezéssel biztosan megoldható a találat:

Helyesebb – a négy sort kézi kódolással eggyé lehet alakítani

```
hülyéskedik[IGE]
hülye[MN]
hülyít[IGE]
hülyül[IGE]
```

Lightproof – reguláris kifejezés, nem olvasmányos, nem használ morfológiát

```
[-\w]*hüly[eéüí][-\w]*<-option("offensive")->_
```

A további szabályokat PERL-ben megírt programmal konvertáltam. Néhány példa:

Helyesebb – itt nem kell morfológia (az összetételt a motor generálja)

```
már < mint pedig
```

Lightproof – itt mindig meg kell adni a javaslatot is

```
már (mint|pedig) <- option("compound") -> már\1
```

Helyesebb – egy egyszerű sorból

```
csesz[IGE] &-Csesznek
```

Lightproof – nehezen, de olvasható kifejezés keletkezik

```
[\w]*cses+z[\w]*<-option("offensive")and morph(\0,r"st:csesz(ik)?\b",False)->_
```

Helyesebb – a képzett alakokra is működik

```
zsáner[FN](osszetett=0) = kedvenc[FN] ízlés[FN] modor[FN]
```

Lightproof – sajnos a visszatoldalékolás nem megy képzőkre

```
zsáner\w*<-option("foreignwords")and morph(\0,r"st:zsáner\b",False)
and not morph(\0,r"st:zsáner po:noun ts:NOM pa:",False) ->=
generate(u"kedvenc",\0)[0] + "|" + generate(u"ízlés",\0)[0] + "|" +
generate(u"modor",\0)[0]
```

Helyesebb – ez már kicsit összetettebb

```
csatlakozó[MN](osszetett)> { alig eléggé apránként gyorsan lassan hamar
nem } > (nevszoi)(IGEMIN+rag=ALL) = csatlakozó_
```

Lightproof – a hatás pontosan megegyezik

```
((alig|eléggé|apránként|gyorsan|lassan|hamar|nem) )*((csatlakozó)(\w+))
<- option("devide") and morph(\3, r"po:vrb ts:PRES_INDIC_INDEF_SG_ 3
ds:Ő_PRESPART_adj ts:NOM pa:",False) and word(-1) and (morph(word(-1),
r"is:ALL", False))) -3> \4 \5
```

Helyesebb – itt tipikus távolhatás van

```
mintha <{(Felt=0) (igei)(IGEMIN=0)}<(igei)(IGEMIN+Felt=IGEMIN) = mint_ha
```

Lightproof – a távolhatást itt is majdnem tökéletesen meg lehetett oldani

```
mintha (?![\w]+n([ae]|á(|[mdk]|na?k|tok)|é(|[mdk]|nek|tek)),? )*
<- option("devide")and not(word(1)and morph(word(1),
r"PRES_COND_[SP][GL]_[123]", False)) -> mint ha
```

Helyesebb – itt egyszerű távolhatás van

```
produkál[IGE] >< zajt ricsajt lármát = csap[IGE]
produkál[IGE] -><< zajt ricsajt lármát = létrehoz[IGE] felmutat[IGE]
megteremt[IGE] alkot[IGE] termel[IGE] gyárt[IGE]
```

Lightproof – a távolhatás tökéletesen azonos

```
produkál(\w*) <- option("foreignwords") and stem(\0) == ["produkál"]
and (zajt|ricsajt|lármát) in TEXT ->= generate(u"csap", \0)[0]
[-\w]*(produkál\w*) <- option("foreignwords")and morph(\0,r"st:produkál
po:vrb",False)and not morph(\1,r"pa:",False)and not(zajt|ricsajt|lármát)
in TEXT -1>= generate(u"létrehoz", \1)[0] + "|" +
generate(u"felmutat",\1)[0] +"|"+ generate(u"megteremt" \1)[0] +"|"+
generate(u"alkot",\1)[0] +"|"+ generate(u"termel",\1)[0] +"|"+
generate(u"gyárt",\1)[0]
(produkál)\w+<-option("foreignwords")and morph(\0,r"st:produkál po:vrb",False)
and morph(\0,r"pa:",False)and not(zajt|ricsajt|lármát)in TEXT -1> létrehoz
|felmutat|megteremt|alkot|termel|gyárt
```

Mint a példákból is látható, van olyan szabály, melyet több szabállyá fordítottam, de általában egy sorból egy sor keletkezett. Ahol lehetett, kerültem az elemzést és a generálást. Mintegy háromezer-ötszáz sort sikerült lefordítanom. Ezek többsége lokális szintaxissal megfogalmazható volt. Ilyen a stiláris hibák többsége. Szám szerint ezek a szabályok adják a zömét, de nem mindet.

5.6. Amit nem tudtam átvinni

A két kritikus szabálycsoport a magyarban: az egybe- és különírás, valamint a vesszőgondok. Mindkettőnél komoly akadályba ütköztem.

A szóösszetétel szabályainak többségét sikerült átvinni, de a több esetet lefedő általános szabályoknál nem működött kielégítően. Ennek fő oka a szóelemző adatbázisa volt.

Az összetételek általános szabályaiból: két főnév – ha szerepel ige is a mondatban –, általában egybe írjuk, ha az első ragozatlan, jelöletlen, a másodiknak meg nincs birtokragja.

A Helyesebb algoritmusában ezt akkor alkalmaztam, ha az összetétel szerepel szótári tételként a morfológia szerint. Bár a Hunmorf is tartalmaz ilyen tételeket, de nem olyan jó válogatást, mint a MUMOR szótár. További korlát, hogy az általam használt HUMOR szótárban finomabb szófaji beosztások vannak. A főnév és melléknév közti szófajok (népnév, foglalkozásnév...) melyek módosítják az összetételi szabályt, szerepelnek a szótárban. Ha a szótan nem pontos, akkor bizonytalan a szóösszetételek ellenőrzése.

Máshol az egyéb feltételek mellett csupán az ellenőriztem, elfogadja-e a Hunspell az összetételt. A Hunspell a HUMOR-hoz képest nagyon megengedő.

A másik nehézség a vesszőknél támadt, Itt sok szabályban szerepelt olyan feltétel, melyet nem vagy csak korlátosan lehetett megvalósítani a Lightproofban. Tipikus eset, a ragozott ige meglétének keresése.

Például, ha két ragozott ige szerepel távol egymástól, és nincs köztük írásjel, akkor biztos kifelejtették a vesszőt. Erre számos szabály van a Helyesebb-ben. Azért számos, hogy a hiba valószínű helyét is felderítsem. Ezek egyike sem vihető át korlátozás nélkül, mert távolhatást legfeljebb reguláris kifejezéssel adhatok meg, de míg a Helyesebbnél ebben absztrak tulajdonságok is szerepelhetnek, a Lightproof csak formai (karakteres) reguláris kifejezéseket enged meg.

5.7. A Hunspell hibái

Gondot okozott a morfológia pontatlansága is. A HUMOR adatai közt is van hiba – hibátlan nyelvi adatbázis nem létezik –, de a tapasztalatom szerint a Hunspell adatai közt nagyságrenddel több tévedés található. Bár a szótár sok fésülésen ment keresztül, de visszajelzések eddig csak a helyesírás-ellenőrző használatából keletkeztek. A melléelemzések ezt nem zavarták, de a további elemzésnél már zavaró volt. Hát még a hibás generátumok.

Gyors viselkedésbeli vizsgálatom alapján arra a következtetésre jutottam, hogy a szótár építésénél túl sok valószínűségi alapon születő döntést hoztak emberi beavatkozás nélkül. A forrásanyaggal sajnos eddig nem volt szerencsém találkozni, de világossá vált, hogy gyakran dönthettek formai alapon. Így az *az* szónak – lett légyen névelő vagy mutató névmás – allomorfja az *azt, annak, abból...*, de sohasem az *arra, arról...*. Bár az allomorfokkal nem foglalkoztam, de nem szabad ilyen gyakori esetekben tévedni.

Hasonlóan érdekes, hogy a *mag* szónak a többes száma lehet *magak*. Feltehetően a *mag* főnév alakjait keresve szép számmal találtak olyan alakokat, mint *magas, magam, magad...*, emiatt ha alaki statisztikát alkalmaztak, bejött az ajakkerekítés nélküli toldalékolás is.

Az összetételeknél nem elég kifinomultak a szabályok. Itt is nagy előnyt jelent a HUMOR-nál, hogy sok célra használták már, és ezektől sok minden finomodott a morfoszintaxisban.

5.8. A kölcsönhatások kezelése

Kevés szabálynál ez nem volt lényeges, de nagy számú szabályok esetén elengedhetetlen, hogy kezeljük azok kölcsönhatását. A Helyesebb-ben ez a szóösszetételnél vált nélkülözhetetlenné. Ezzel kapcsolatos, hogy irányíthatóvá kell tenni a szabályok prioritását is. (Szabályok kölcsönhatását erősen szabályozza a MetaMorpho fordító. Ott a szabályok száma több nagyságrenddel felülmúlja a mondatellenőrzőét[10].)

6. Tervek a jövőre

A megszületett ellenőrző minden hibája ellenére sokkal jobb, mint a korábbi. Az eredeti Lightproofban pár száz gondosan megfogalmazott szabály szerepel. Ezek a hibák kis százalékát derítették fel némi mellélövéssel. Most közel 5000 szabály van, mely nem fedi fel a hibák felét, de kisebb a mellélövések aránya. Biztató kezdet, de nem elég.

A szűk keresztmetszet a Hunspell adatbázisa. Nem a mérete, hanem módszeres javításra szorul. Addig nem érdemes tovább lépni.

A másik szűk keresztmetszet, hogy lehetőség legyen a szavak tulajdonságait is bevenni a reguláris kifejezésekbe. Ehhez Python-bővítésre van szükség, amit a Lightproof megenged, de ezt eddig nem alkalmaztam.

Ha fejlődő rendszerben gondolkodunk, akkor mindenképpen szükség van egy olyan formalizmusra, ami olvasmányos. Ebből egyszerű konvertálással kell megkapni a Lightproofnak megfelelő formalizmust. A jelenlegi szintaxis nem alkalmas az adatbázis fejlesztésére, mert a szabályok nehezen olvashatóak, és formai hiba esetét nem deríti fel a fejlesztő rendszer, futás időben akár fenn is akadhat. A változtatással viszont várható, hogy olyan dinamikusan fejlődő eszközt kapunk, mint a franciák eszköze[11], amely a Lightproofból indult ki, de sokat bővült, és nap mint nap fejlődik a közösség segítségével.

Hivatkozások

1. Naszódi, M.: Nyelvhelyesség-ellenőrzés számítógéppel: Parciális szintaxis[1]. In: VII. Országos Alkalmazott Nyelvészeti Konferencia., Budapest (1997)
2. Németh, L.: A Lightproof lehet a nemzetközi OpenOffice.org nyelvi ellenőrzője[2] (2009)
3. Naszódi, M.: Statisztika megbízhatósága a nyelvészetben[3]. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem (2015) 34–45
4. Naszódi, M.: A magyar helyesírás-ellenőrzők mai állása[4]. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem (2017) 347–354
5. Verberne, S.: Context-sensitive spell checking based on word trigram probabilities[5]. Master's thesis, Taal, Spraak & Informatica University of Nijmegen (2002)
6. Hanák, P.: Magyar nyelvi szófaji egyértelműsítő módszer fejlesztése gépi tanulási algoritmusok felhasználásával[6]. (2002)
7. NyI: Magyar Nemzeti Szövegtár[7] (2002)
8. Ehmann, B., Garami, V., Naszódi, M., Kis, B., László, J.: Subjective Time Experience: Identifying Psychological Correlates by Narrative Psychological Content Analysis.[8]. (2007) 14–25
9. Kuboň, V., Holan, T., Plátek, M.: A Grammar-Checker for Czech[9]. In: ANLC '97 Proceedings of the fifth conference on Applied natural language processing. (1997) 147–154
10. Prószéky, G., Tihanyi, L., Ugray, G.: Moose: a robust high-performance parser and generator.[10]. In: Proceedings of the 9th Workshop of the European Association for Machine Translation, La Valletta, Malta (2004) 138–142
11. WWW: GRAMMALECTE correcteur grammatical open source[11] (2012)

[1] http://www.mek.iif.hu/porta/szint/tarsad/nyelvtud/nylvhely/nylvhely.htm
[2] http://libreoffice.hu/a-lightproof-lehet-az-openoffice-org-alapertelmezett-nyelvi-ellenorzoje/
[3] http://rgai.inf.u-szeged.hu/mszny2015/files/MSZNY2015_press_B5_PQ.pdf
[4] http://www.cs.bme.hu/~naso/langeng/SpellsSate20016.pdf
[5] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.324.7146&rep=rep1&type=pdf
[6] http://members.iif.hu/hanak/prj/i3/i027/ds/hu/i027-dshu.pdf
[7] http://mnsz.nytud.hu/
[8] http://www.cs.bme.hu/~naso/langeng/Ehmann_etal_Empirical_2007_3.pdf
[9] http://www.aclweb.org/anthology/A97-1022
[10] http://www.morphologic.hu/downloads/publications/tl/eamt_2004_malta_pg-tl-ug.pdf
[11] https://www.dicollecte.org/

Conclusions from the Conversion of Linguistic Data of a Hungarian Grammar Checker

Mátyás Naszódi, e-mail: naszodim@morphologic.hu

MorphoLogic, 1122 Ráth György utca 36. Hungary

Extended abstract

Grammar checkers are unable to decide whether a sentence is grammatically correct or not. They only look for faulty parts in the sentence that are hidden from spell checking. In most cases, proofreaders use local syntax (like in context sensitive spelling) based on probability of sequences of words.

The quality of proofing tools depends on the applied formalism for describing the linguistic database. Its goodness is measured by the percentage of recognized real errors and by the rate of false alarms. The applied formalisms and drivers have limited capabilities.

Microsoft decided to develop its own language support for Office. The last externally developed tool, the grammar checker *Helyesebb* (by MorphoLogic) was replaced by their own engine and data a year ago. The quality of the new proofing tool is below the acceptable level. I have tried to port the linguistic knowledge of Helyesebb to LibreOffice, as an open source code. By that time, the grammar engine *Lightproof* (by László Németh) had been developed however, it contained only a small amount of linguistic knowledge.

This article focuses on the capabilities of Helyesebb and Lightproof, and it demonstrates how it is (im)possible to convert linguistic knowledge from one formalism to the other one. The experiment also provides other linguistic conclusions. It points to the quality problems of morphological descriptions and shows some linguistic phenomena concerning proofing modules.

The formalisms of the linguistic database of the two tools are similar. More than 4000 rules which use only neighborhood checking have been transformed with the help of a PERL program. Handling far relations however needs *partial syntax* instead *local*.

Another conclusion is that the morphological system (in open source systems it is *Hunspell*) is enough for spelling, but it may not satisfy the needs of analysis and word generation requirements for grammar checking.

The obtained quality is much better than that of the original Lightproof, but is far from the desired. The bottlenecks of the development are the weakness of morphology, the lack of possibility of usage of far relations, and the low level formalism of the linguistic description.

Névutók, előre! Korpuszvezérelt elemzés a névutószerű elemekről

Ligeti-Nagy Noémi[1,2]

[1]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport,
1083 Budapest, Práter u. 50/A
[2]Pázmány Péter Katolikus Egyetem Bölcsészet- és Társadalomtudományi Kar
2087 Piliscsaba, Egyetem u. 1.
ligeti-nagy.noemi@itk.ppke.hu

Kivonat Jelen tanulmányban a magyar névutók csoportját vizsgálja. A korpuszvezérelt elemzés során hat – bináris – tulajdonságra fókuszálunk, a névutók kategóriájára zárt és merev osztály helyett kontinuumként tekintve. Eredményeink egy újszerű megközelítést jelentenek, amellyel hatékonyabban elemezhetők a névutószerű elemek a mondatban.
Kulcsszavak: névutók, korpusznyelvészet, mondatelemző

1. Bevezetés

Ebben a cikkben egy átfogó, korpuszvezérelt elemzést készítünk a magyar névutók folyamatosan változó csoportjáról. A cél az, hogy ezekre a szavakra ne egy merev és homogén kategória elemeiként tekintsünk, hanem inkább kontinuumként, amelyben több különböző tulajdonság alapján helyezhetők el az egyes szavak. Távolabbi célunk egyfajta függőségi nyelvtan keretein belül elemezni őket, lehetőség szerint mindvégig szem előtt tartva egy pszicholingvisztikailag motivált elemzőrendszer (AnaGramma, [1,2]) igényeit.

A legfontosabb tisztázandó kérdés az elején, hogy mit értünk névutó alatt. Számítógépes nyelvészeti szempontból a névutók azok az – önálló – elemek, amelyek megkérdőjelezhetetlenül jelzik egy főnévi csoport végét a mondatban (ld. például az 1-es példában). Ugyanakkor elkerülhetetlen egy mélyreható irodalmi áttekintés arról, hogy pontosan mely elemek tartoznak ebbe a csoportba. A 2.1. fejezetben bemutatjuk, hogy a jelentősebb nyelvészeti iskolák hogyan vélekedtek a névutókról, noha a mi célunk ezeknek a szavaknak az elemzése függetlenül attól, hogy minden iskola szerint névutók-e vagy sem. A 2.2. fejezetben a kutatás hátterét jelentő elemzőrendszert mutatjuk be röviden. A 3. fejezetben bemutatjuk a saját kritériumainkat és osztályozási feltételeinket. A 4. fejezetben összefoglaljuk mindazt, ami vizsgálataink során a korpuszból nyilvánvaló vált, illetve bemutatjuk, hogyan jellemezhetők az egyes névutószerű elemek a szófaj jellemző tulajdonságait képviselő bináris jegyek segítségével. Cikkünket az 5. fejezetben összefoglalással zárjuk.

(1) *A főnév* **után** *jön a névutó.*

2. Bevezetés

2.1. Kapcsolódó irodalom

Aligha találunk még egy olyan szóosztályt, amely viszonylag kevés számú eleme ellenére olyan hosszan tartó vitát eredményezne, mint a magyar névutók. Szinte lehetetlen két különböző tanulmányt találni, amelyek ugyanolyan klasszifikáció mellett érvelnének. Ebben a fejezetben megpróbálunk áttekintést adni arról, amit az elméleti nyelvészek eddig mondtak a magyar névutókról. A hagyományos magyar nyelvtanokban [3] a névutók a funkciószavak egy csoportját alkotják, amelyek morfológiai jellegű egységet alkotnak az őket megelőző főnévvel (vagy főnévi kifejezéssel), így esetragként működnek. Ezek a nyelvtanok a következőképpen csoportosítják a névutókat:

(2) a. **Valódi névutók**, a névutók, amelyek esetrag nélküli főnevet követnek: *által*, *alá*, *alatt*, *alól*, *elé*, *előtt*, *elől*, *ellen*, *felé*, *felől*, *fölé* , *fölött*, *helyett*, *iránt*, *köré* , *körül*, *közé*, *között*, *közül*, *mellé*, *mellett*, *mellől*, *miatt*, *mögé*, *mögött*, *mögül*, *nélkül*, *szerint*, *után*.

b. **Ragvonzó névutók**: *fogva*, *kezdve*, *nézve*, *keresztül*, *túl*, *együtt* stb.

c. **Birtokos személyjelet tartalmazó névutók**: *alapján*, *céljából*, *ellenére*, *érdekében*, *értelmében*, *esetében*, *folyamán*, *javára*, *következtében*, *nyomán*, *részéről*, *számára*, *táján* stb.

Ezt a kategorizációt erősen vitatják a strukturalista vagy generatív hátterű tanulmányok. A *Strukturális magyar nyelvtan* 3. kötetében [4] a névutókat három (és fél) csoportba sorolják:

(3) a. **Esetszerű névutók**: a csoport tagjai esetragként viselkednek. Nem adnak esetet az NP-nek, esetrag nélküli főnévi csoportot vonzanak. Például: *által*, *alá*, *alatt*, *alól*, *elé*, *előtt*, *elől*, *ellen*, *felé*, *felől*, *fölé*, *fölött*, *helyett*, *iránt*, *köré*, *körül*, *közé*, *között*, *közül*, *mellé*, *mellett*, *mellől*, *miatt*, *mögé*, *mögött*, *mögül*, *nélkül*, *szerint*, *után*.

b. **Valódi névutók**: azok az elemek, amelyek lexikai esettel rendelkező névszót vonzanak: *alul*, *át*, *belül*, *kívül*, *együtt* stb.

c. **Egyéb névutók**: azok a névutók, amelyek sosem fordulnak elő személyes névmással: *végett*, *közben*, *óta*, *gyanánt*, *hosszat*.

d. (Névutószerű elemek): egy átmeneti osztály, birtokos személyjeles alakokat tartalmaz.

Felsorolásunkból még hiányzik egy tisztán generatív megközelítés. Ezt biztosítja például É. Kiss a *The Syntax of Hungarian*-ben [5]. Itt az eddig említett szavak a következő kategorizációban jelennek meg:

(4) a. **Vonzatos határozószók** (lexikai esettel rendelkező főnévi csoportot vonzó „névutók”): *együtt*, *alul* stb.

b. **Idiomatikus melléknévi vagy határozói igenevek**, lexikai esettel rendelkező főnévi csoportot vonzanak: *nézve*, *kezdve* stb.

c. **Esetrag nélküli főnévi csoportot vonzó névutók**: *alá*, *alatt*, *alól*, *elé*, *előtt*, *elől*, *felé*, *felől*, *fölé*, *fölött*, *fölül*, *köré*, *körül*, *közé*, *között*, *közül*, *által*, *ellen*, *ellenére*, *helyett*, *szerint*, *iránt*, *miatt*, *nélkül*, *után*, *végett* és *révén*.

Ez a három különböző kategorizáció jól példázza, milyen bizonytalan a névutók meghatározása a magyar nyelvészeti irodalomban. Ugyanakkor szükséges egy másik, erre a kérdésre vonatkozó irodalmi tételre is hivatkoznunk, nevezetesen Dékány Éva doktori disszertációjára, amely részletes listát tartalmaz az összes magyar névutóról [6]. Dékány Éva csoportosítását a *naked* és a *dressed* névutók fogalmára alapozta, amely meghatározások eredetileg Marácz Lászlótól származnak [7]. *Naked* névutó alatt az esetragos, *dressed* névutó alatt pedig az esetrag nélküli névszóval társuló névutókat értik. Ez a kategorizáció is szerepel az 1. táblázatban, amely egy kimerítő, és összehasonlító jellegű felsorolása az irodalomban valaha névutóként említett szavaknak (a fent említett négy tanulmány alapján).

Az 1. táblázatban dőlttel szedett szavak a névutók. Az ezeket követő négy-négy oszlop a jelen tanulmányban említett négy hivatkozást jelöli (MG mint Magyar grammatika [3], Str mint Strukturális magyar nyelvtan [4], SoH mint The Syntax of Hungarian [5] és D.É. mint Dékány Éva disszertációja [6]). Az oszlopokban található jelek az adott tanulmány névutókkal kapcsolatos véleményét jelölik: egy szám azt jelenti, hogy az adott szót az adott tanulmányban névutónak tekintik; egy betű azt jelenti, hogy az adott szót említik a tanulmányban, de mint nem-névutót. A különböző számok és betűk különböző alkategorizációra utalnak. Ha zárójel szerepel az adott jel körül, azt jelenti, hogy ez a szó nem szerepel az adott tanulmányban, és nem vagyunk biztosak a(z adott tanulmány szerinti) kategorizációjában. A - azt jelenti, hogy az adott irodalom nem állít semmit az adott szóról.

A névutók kategorizációjának fő problémája azonnal világossá válik a táblázat alapján: van a szavaknak egy csoportja (hívjuk őket „tiszta névutóknak”), amelyek minden helyzetben névutóként viselkednek, a névutók minden fontos tulajdonságával rendelkeznek, és nem utolsósorban minden jelentős tanulmányban névutóként határozzák meg őket. Más szavakkal: a táblázat minden oszlopában egy szám tartozik hozzájuk. Ugyanakkor vannak olyan szavak is, amelyek – és pontosan erre kívánunk rávilágítani az 1. táblázattal, – bizonyos szempontból névutók, de más szempontból valami egészen mások. A célunk az, hogy ezt a folyamatosságot ragadjuk meg: meghatározzuk a prototipikus névutók csoportját és jellemzőit, illetve körülírjuk az ezt a magot körülvelő rétegeket, amelyek a nem prototipikus névutókat tartalmazzák.

1. táblázat. A névutóként tárgyalt szavak listája és csoportosításuk összefoglalása. Rövidítések: MG: Magyar grammatika [3], Str: Strukturális magyar nyelvtan [4], SoH: The Syntax of Hungarian [5], D.É.: Dékány Éva disszertációja [6]. Számok: a szó az adott helyen névutó; betűk: nem névutó (eltérő szám / betű: az adott tanulmány alkategoriái). –: nem meghatározható a besorolás. Zárójel: nem egyértelmű kategorizáció.

	MG	Str	SoH	D.É.		MG	Str	SoH	D.É.
alá	1	1	1	1	*körül*	1	1	1	1
alapján	3	a	-	(b)	*következtében*	3	a	-	(b)
alatt	1	1	1	1	*közben*	1	3	-	-
alól	1	1	1	1	*közé*	1	1	1	1
által	1	1	1	1	*közel*	2	-	(a)	2
alul	2	2	a	2	*között*	1	1	1	1
át	2	2	a	2	*közül*	1	1	1	1
belül	2	2	a	2	*létére*	1	a	-	b
céljából	3	a	-	(b)	*mellé*	1	1	1	1
együtt	2	2	a	2	*mellett*	1	1	1	1
elé	1	1	1	1	*mellől*	1	1	1	1
ellen	1	1	1	1	*miatt*	1	1	1	1
ellenére	3	a	-	1	*módjára*	1	a	-	-
elől	1	1	1	1	*módra*	1	-	-	-
előtt	1	1	1	1	*mögé*	1	1	1	1
esetén	3	a	-	1	*mögött*	1	1	1	1
felé	1	1	1	1	*mögül*	1	1	1	1
felől	1	1	1	1	*múltán*	1	-	-	-
felül	2	-	(a)	2	*múlva*	1	-	-	1
fogva	2	2	b	a	*nélkül*	1	1	1	1
fölé	1	1	1	1	*nézve*	2	2	b	a
fölött	1	1	1	1	*nyomán*	3	a	-	1
fölül	2	-	(a)	1	*óta*	1	3	(1)	1
folytán	3	-	-	1	*részére*	3	a	-	b
gyanánt	1	3	(1)	1	*révén*	3	a	-	b
hasonlóan	-	-	-	2	*során*	3	a	-	1
helyett	1	1	1	1	*számára*	3	a	-	b
hosszat	-	3	-	-	*szembe*	(2)	-	(a)	2
innen	2	-	-	2	*szemben*	2		(a)	2
iránt	1	1	1	1	*szemből*	(2)	-	(a)	2
ízben	1	-	-	-	*szemközt*	(2)	-	(a)	2
javára	3	a	-	b	*szerint*	1	1	1	1
kedvéért	3	a	-	b	*tájban/tájt*	1	-	-	1
képest	-	-	(a)	2	*túl*	2	2	a	2
keresztül	2	2	a	2	*túlra*	(2)	-	(a)	2
kezdve	2	2	b	a	*túlról*	(2)	-	(a)	2
kivéve	(2)	-	(b)	a	*után*	1	1	1	1
kívül	2	2	a	2	*útján*	3	a	-	1
kívülre	(2)	-	-	2	*végett*	1	3	(3)	1
kívülről	(2)	-	-	2	*végig*	2	-	(a)	2
köré	1	1	1	1					

Az említett csoportosítások mind egyetértenek a névutók egy aspektusában: az esetragoktól való elkülönítésükben. Éppen ezért itt most nem kívánunk szólni az esetragok és a névutók kapcsolatáról. Ami igazán érdekel minket, azok az utóbbiak jellemző tulajdonságai: ezek okozzák ugyanis a kategorizációk eltéréseit. A korpuszt vizsgálva ezekre a jellemzőkre koncentrálunk, próbálva egy nagyobb képet megragadni, a névutókra mint kontinuumra tekintve: a prototipikus névutók minden tulajdonsággal rendelkezni fognak, míg a perifériás névutók csak néhánnyal.

2.2. Az elemzőrendszer

Az AnaGramma nyelvi elemzőrendszer az emberi szövegfeldolgozást igyekszik modellálni, ezért az elemzendő szöveget balról jobbra, szavanként elemzi. Az aktuálisan vizsgált szó mellett azonban az azt követő néhány tokenre is „kitekint": ezt egy két szó méretű elemzési ablakkal modelláljuk. Az AnaGramma elemzésének kimenete egy függőségi fa. Az elemző egyik fontos működési alapelve a *kereslet-kínálat* keretrendszer, amelyben az épülő függőségi fa kapcsolatai a mondatban szereplő szavak keresletei és kínálatai alapján valósulnak meg. Minden egyes token több tulajdonsággal rendelkezik, amely tulajdonságok keresletként vagy kínálatként jelennek meg a mondatban. Ha egy kereslet összetalálkozik egy megfelelő kínálattal, akkor megszületik a függőségi kapcsolat közöttük.

3. Módszer

Az 1. fejezetben azt állítottuk, hogy a névutók jelentősége – számítógépes nyelvészeti szempontból – abban rejlik, hogy az esetragokhoz hasonlóan minden kétséget kizáróan jelzik egy főnévi csoport végét. Ugyanakkor az 1. táblázat számos eleme képes megjelenni a főnévi frázis előtt és után is. Jelen tanulmányunkban – tekintve, hogy távolabbi célunk minden névutószerű elem viselkedésének algoritmikus leírása – egyet tartunk csak meg a kritériumok közül: a névutó-jelöltnek minden esetben együtt kell járnia egy főnévi csoporttal, mindegy, hogy maga előtt vagy maga után veszi, esetraggal vagy esetrag nélkül. Ezt a különbséget szemlélteti az 5a és az 5b alatti példa. Az 5a-ban szereplő *kívülről* beletartozik a vizsgált szavak körébe, de az 5b-ben szereplő nem.

(5) a. *A hang a szobán* ***kívülről*** *jött.*

b. *A hang* ***kívülről*** *jött.*

A vizsgálat során használt korpusz az MNSz2 2.0.4-es verziója [8]. A vizsgálat egyik szakasza motiválta a másikat; kereséseinket az 1. táblázat elemeiből indítottuk, de a keresések során előbukkanó újabb névutó-jelöltekkel bővítettük azt.

Az irodalom, valamint korpuszlekérdezéseink eredményei alapján a következő tulajdonságokat határoztuk meg:

- **pozíció**: a névutó a névszói csoport előtt vagy után szerepel? Ennek értéke független attól, hogy azok szigorúan követik-e egymást, vagy kerülhet közéjük más elem. A pozíció tulajdonságának értéke 1, ha a névutó minden esetben a névszó után következik, 0 egyébként.
- a vonzat **esetrag**ja: első lépésként, a tulajdonságok binaritását megtartandó, az esetrag nélküli és az esetragos főneveket vonzó névutók elkülönítését célozzuk meg. Ugyanakkor a későbbiekben szükséges lehet a névutók megkülönböztetése az általuk vonzott esetragok alapján is. Ennek a tulajdonságnak az értéke 1, ha a névutó mindig esetrag nélküli főnévhez járul, 0 egyébként.
- **szomszédosság**: értéke 1, ha a névutó és a névszó minden esetben közvetlen szomszédai egymásnak, 0 egyébként.
- pozíciójuk a **kérdőszavas mondatokban**: minden esetben követi a névutó a kérdőszót, vagy hátul maradhat? Az előbbi esetben az érték 1, egyébként 0.

Ezek jellemzően szintaktikai attribútumok, főként a – jelen tanulmányban alkalmazott – számítógépes nyelvészeti nézőpontunk motiválja őket. Ezen kívül azonban szükséges lehet néhány morfológiai jellemző számbavétele is, amelyek ugyan nem befolyásolják jelentősen ezeknek a szavaknak a számítógépes elemzését, de kellenek a komplex képhez.

- **mutató névmások**: a névutó is mindig másolódik a mutató névmásra, vagy csak az esetrag? Az első esetben az érték 1, egyébként 0 (lásd 6a és 6b).
- személy-szám **egyeztetés**: ha a névutó egy személyes névmáshoz járul, hol jelenik meg az egyeztetés? Értéke 1, ha az egyeztetés a névutón jelenik meg, 0 egyébként (ld. 7a és 7b).

(6) a. *a rét alól*
*az **alól** a rét **alól***

b. *a réten át*
*azon a réten **át***

(7) a. *alól**am***

b. *rajt**am** keresztül*

4. Eredmények

A 2. táblázatban látható a névutók bináris értékű tulajdonságokra lebontott elemzése. A „poz” oszlop a névutónak a névszóhoz képest vett pozícióját értékeli: ha a névutó mindig követi a névszót, akkor ennek értéke 1, különben 0. A „NOM” oszlop a névutó által megkívánt esetraggal kapcsolatos: ha a névszó mindig esetrag nélküli, akkor ennek értéke 1, különben 0. A „szom” oszlop a szomszédosságot vizsgálja: ha a névszó és a névutó mindig közvetlenül egymás

mellett találhatóak, akkor ennek értéke 1, különben 0. Fontos kitétel, hogy a névszók előtt is megjelenni képes névutóknál a névszó előtti (és a névutó utáni) névelőt nem számítottuk, csak egyéb módosító elemek jelenlétét. A „wh" oszlop a kérdőszavas kifejezésekben vizsgálja a névutókat: ha azok a kérdőszó névutóiként jelennek meg, akkor ennek értéke 1, különben 0. Az „az_a" oszlop a mutató névmási determinánsú főnévi csoportokkal kapcsolatos: ha a névutó a mutató névmáson is ismétlődik, akkor ennek értéke 1, különben 0. Végül a „pers" oszlop azt értékeli, hogy ha személyes névmáshoz járul a névutó, az egyeztetés a névutón (1) vagy egy külön elemen jelenik-e meg (0).

A módszertannal kapcsolatban megjegyzendő, hogy 5 darab ellenpélda szükségeltetett ahhoz, hogy egy adott névutó ne kapjon 1-es értéket egy tulajdonságra. Ha például 5 esetben szerepel a névszó előtt, akkor már 0 látható a „poz" oszlop adott sorában.

A „wh" tulajdonság kiértékelése különösen nehéz a korpusz alapján, ezért a legtöbb esetben itt az értéket saját nyelvi intuíciónk alapján határoztuk meg. A ?-et tartalmazó cellák azt jelzik, hogy az adott tulajdonság tesztelésére kiötlött példáink elfogadhatóságában nem vagyunk teljesen biztosak.

Az első, ami szembetűnik a 2. táblázatban, hogy a 4-6. oszlopban gyakori a -. Ez azt jelzi, hogy az adott szerkezetben az adott névutó nem fordul elő. Például a *gyanánt* névutót nem lehet személyes névmáshoz illeszteni, ezért szerepel - a *gyanánt* vektorának utolsó pozícióján.

Ha a kizárólag 1-es értéket tartalmazó vektorú névutókat tekintjük, akkor kirajzolódik a prototipikus névutók köre: ezek szinte teljesen megfelelnek a szakirodalmi kategorizációk alapján is „tiszta névutóknak" nevezett csoport tagjaival. A kivételt egyrészt azoknak a névutóknak egy része jelenti, amelyeknek az egyes szám harmadik személyű névmáshoz illesztett alakja megegyezik az alap alakjukkal (*alá, elé*). Ezeknél megfigyelhető, hogy a névszó előtt is megjelenhetnek *elé a ...-nak* szerkezetben (ld. 8). Kivételt jelent még az *által*, amely főleg a szépirodalmi alkorpuszban, de kisebb számban a személyes alkorpuszban is előfordul esetragos névszóval (9).

(8) *- Mestrében meg sem állunk - szólt elé a kocsisnak.*

(9) *Természetesen szintén szigorúan tévén által.*

Ha eltekintünk a 4-6. tulajdonságok értékétől, kirajzolódik még egy jelentős csoport, amelyek az első három, szintaktikai szempontra 1 értéket kaptak. Mivel a 4-6. tulajdonság mint feltétel nem alkalmazható teljes körűen ezekre a szavakra, kihagyásuk egy algoritmikus feldolgozásnál nem okozhat gondot. Az 1-3. tulajdonságra kapott 1 érték azonban azt jelzi, hogy ezeket a szavakat érdemes lenne névutónak annotálni a korpuszban, mivel mindig főnévi csoportot lezáró pozícióban szerepelnek, közvetlenül a magán testes esetragot nem viselő névszó után.

A táblázatban a kakukktojás a *szemből*: ezt, mivel a fő szintaktikai tulajdonságokra még értéket sem kaphat, lévén hogy nem fordul elő olyan szerkezetekben, nem tekintjük semmilyen szempontból sem névutónak.

2. táblázat. A magyar nyelvészeti szakirodalomban valaha névutóként említett szavak tulajdonságvektorként történő megjelenítése.

	poz	NOM	szom	wh	az_a	pers
alá	0	0	1	1	1	1
alapján	1	0	1	1	0/1	0
alatt	1	1	1	1	1	1
alól	1	1	1	1	1	1
által	1	0	1	1	1	1
alul	1	0	0	1	0	0
át	0	0	0	0	0	0
belül	?	0	1	1	0	0
céljából	1	0	1	1	0	-
együtt	0	0	0	0	0	0
elé	0	0	1	1	1	1
ellen	1	1	1	1	1	1
ellenére	1	0	1	1	0	-
elől	1	1	1	1	1	1
előtt	1	1	1	1	1	1
esetén	1	0	1	1	0	-
felé	1	1	1	1	1	1
felől	1	1	1	1	1	1
felül	1	0	0	0	0	0
fogva	1	0	1	1	0	0
fölé	1	1	1	1	1	1
fölött	1	1	1	1	1	1
fölül	1	0	1	1	0	1
folytán	1	0	1	?	0	-
gyanánt	1	1	1	-	-	-
hasonlóan	0	0	0	1	0	0
helyett	1	1	1	1	1	1
hosszat	1	1	1	-	-	-
innen	0	0	0	?	0	0
iránt	1	1	1	1	1	1
ízben	1	1	1	-	-	-
javára	1	0	0	1	0	1
kedvéért	1	0	0	1	0	1
képest	1	0	0	1	0	0
keresztül	0	0	0	0	0	0
kezdve	0	0	1	1	0	0
kivéve	0	0	0	1	0	0
kívül	0	0	0	1	0	0
kívülre	1	0	0	1	0	0
kívülről	1	0	1	1	0	-
köré	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

	poz	NOM	szom	wh	az_a	pers
⋮	⋮	⋮	⋮	⋮	⋮	⋮
körül	1	1	1	1	1	1
következtében	1	0	1	1	0	-
közben	1	1	1	1	1	-
közé	1	1	1	1	1	1
közel	0	0	0	0	0	0
között	1	1	1	1	1	1
közül	1	1	1	1	1	1
létére	1	1	1	-	-	-
mellé	1	1	1	1	1	1
mellett	1	1	1	1	1	1
mellől	1	1	1	1	1	1
miatt	1	1	1	1	1	1
módjára	1	1	1	-	0?	-
módra	1	1	1	-	-	-
mögé	1	1	1	1	1	1
mögött	1	1	1	1	1	1
mögül	1	1	1	1	1	1
múltán	1	1	1	1	0	-
múlva	1	1	1	-	-	-
nélkül	1	1	1	1	1	1
nézve	1	0	1	1	0	0
nyomán	1	0	1	1	0	-
óta	1	1	1	1	1	-
részére	1	0	1	1	0	1
révén	1	0	1	1	0	-
során	1	0	1	1	0	1
számára	1	0	1	1	0	1
szembe	0	0	0	1	0	0
szemben	0	0	0	1	0	0
szemből	-	-	-	-	0	0
szemközt	0	0	0	1	0	0
szerint	1	1	1	1	1	1
tájban/tájt	1	1	1	-?	-?	-
túl	0	0	0	0	0	0
túlra	1	0	0	1	0	0
túlról	1	0	0	1	0	0
után	1	1	1	1	1	1
útján	1	0	1	1	0	-
végett	1	1	1	1	-	1
végig	0	0	0	0	0	0

Határozottan kirajzolódik a perifférián lévő szavak csoportja is: ezek azok, melyeknek minden értéke 0. Az *együtt, közel, túl, végig* esetében közös, hogy minden szempontból megvizsgálhatóak, tehát minden olyan szerkezetben előfordulnak, amiben a tipikus névutók (zöme), azonban nem úgy viselkednek ezekben, mint a tipikus névutók. Ezeknek az annotációjában érdemes a határozószói jellegükhöz ragaszkodni, viselkedésük a „vonzatos határozószó" terminussal ragadható meg a legjobban: egy esetragos névszót vonzanak, maguk előtt vagy maguk mögött is elfogadják, és távolra is kerülhetnek a vonzatuktól.

A legérdekesebbek azok az elemek, melyekhez az 1 0 1 vektor tartozik (a 4-6. tulajdonságot figyelmen kívül hagyva): ezek a közvetlenül a főnév után szereplő, de esetragos főnevet vonzó névutók. Mint látható, ezek viselkedésükben sokkal közelebb állnak a prototipikus névutókhoz, mint a vonzatos határozószók. A problémát az általuk megkövetelt esetrag okozza: egy névszó és egy prototipikus névutó kapcsolatának elemzése nem sokban különbözik egy esetragos névszóétól; ha az AnaGramma elemzőrendszer balról jobbra történő feldolgozását vesszük alapul, akkor az algoritmus lényege, hogy az esetrag nélküli névszó mellett, ha megtaláljuk a névutót, a kettőt összekapcsoljuk, és együtt lehetnek majd a mondatban egy másik elemnek a vonzatai. A vonzatos határozószók esetében a kívánt esetrag keresletként jelenik meg a mondatban, amit egy, az adott esetragot magán viselő főnévi csoport kínálata fog kielégíteni. Az 1 0 1 vektorú elemek esetében azonban a névutót megelőző főnévi csoport nem esetrag nélküli, az azt lezáró esetrag azt kommunikálja, hogy itt véget ért egy főnévi csoport. A probléma megoldására alább javasolunk egy lehetséges algoritmust.

A 2. táblázat néhány cellája további magyarázatot igényel. Az *alapján* 0/1 értéket kapott az „az_a" tulajdonságra. Ezt azért így jeleztük, mert minden más névutó vagy az egyik, vagy a másik módon viselkedik a mutató névmási determinánssal, az *alapján* azonban mindkét formában előfordul a korpuszban (10a és 10b). Ez csak erre a névutóra jellemző.

(10) a. *az alapján az idézet alapján*

b. *annak az élettani tapasztalatnak alapján*

A 3. táblázatban szerepelnek a korpuszlekérdezéseink során gyakran előkerülő, névutószerű viselkedést produkáló szavak (melyeknek egy része már máshol is említésre került névutó-jelöltként [9]). Mint látható, a közvetlen szomszédosság mindegyikük jellemző tulajdonsága; ezen kívül egy-két kivételtől eltekintve mindig a főnévi csoport után találhatók. Emiatt az 1 0 1 vektorú csoport elemeihez közeli szavak, azzal a különbséggel, hogy morfológiailag összetettebbek, vagy birtokos személyjeles, vagy esszívuszi módhatározó ragos szóalakok. Mondatbeli elemzésük nem sokban különbözik attól, amit a részletes morfológiai elemzésük aktiválna, azonban a jelentésük és a kizárólag ebben a tipikus névutó-pozícióban történő előfordulásuk azt indokolja, hogy felvegyük őket a névutók közé.

Az eddig elmondottak alapján a névutószerű elemek három nagy csoportja körvonalazódik. A gépi mondatelemzés során – az AnaGramma balról jobbra haladó, *kereslet-kínálat* keretrendszerben működő eljárását szem előtt tartva – a következő algoritmusokat gondoljuk megvalósíthatónak:

3. táblázat. A korpuszlekérdezések során megfigyelt névutószerű elemek listája és a tulajdonságokhoz rendelt értékeik.

	poz	NOM	szom	wh	az_a	pers
címén	1	0	1	1	0	-
dacára	0	0	1	1	0	-
eltérően	0	0	1	1	0	0
esetében	1	0	1	1	0	1
érdekében	1	0	1	1	0	1
értelmében	1	0	1	1	0	-
fényében	1	0	1	1	0	-
függően	1	0	1	1	0	0
hiányában	1	0	1	1	0	-
idején	1	0	1	1	0	-
jegyében	1	0	1	1	0	-
keretében	1	0	1	1	0	-
kezdődően	1	0	1	1	0	0
köszönhetően	0	0	1	1	0	0
követően	1	0	1	1	0	0
megegyezően	1	0	1	1	0	0
megelőzően	1	0	1	1	0	0
megfelelően	0	0	1	1	0	0
terén	1	0	1	1	0	-
ürügyén	1	0	1	1	0	-
vonatkozóan	1	0	1	1	0	0

- A prototipikus névutók esetén: ezek a szavak mindig közvetlenül egy testes esetrag nélküli – jelenleg tehát alapértelmezetten `NOM`-nak címkézett – névszó után találhatók. Feldolgozásuk lényege, hogy az esetrag nélküli névszóhoz érkezve, az ablak elemeire kitekintve látja őket az elemző, ezért további analizálás nélkül összekapcsolja a névszót a névutóval, és onnantól kezdve együtt, az esetragos névszókkal parallel módon vesznek részt az elemzésben. A prototipikus névutók POS-tagjének a `NU` kategóriát javasoljuk – ahogy az az ide tartozó elemek nagy részénél jelenleg is van. Ezek a szavak a 2. táblázat azon elemei, melyeknél az első három cellában 1-es érték található.
- A prototipikus névutókhoz képest a perifférián elhelyezkedő elemek esetén: ezek a szavak mindig esetragos névszót vonzanak, a névszó előtt is, más elemek közbeékelődésével is megjelenhetnek. A 2. táblázatban minden cellában 0 értéket kaptak. Ezeknek az elemzése más, régens szerepű elemekével megegyezően történhet. POS-tagjüknek a `HA` kategóriát javasoljuk, de emellett tulajdonságaik között fontos feltüntetni, hogy milyen esetragú névszót vonzanak – ez lesz ugyanis a keresletük a mondat elemzése során.
- Az esetragos névszót vonzó névutók esetén: ezek a szavak, bár mindig közvetlenül a névszó után találhatóak, különböznek a prototipikus névutóktól abban, hogy a névszó előttük testes esetragot visel(het). A 2. táblázatban 1 0 1 a hozzájuk rendelt vektor kezdete. Elemzésük nem egyezhet meg a vonzatos határozószónak tekintett, perifériás névutószerű elemekével, hiszen

nem klasszikus értelemben vett régensek; nem keresik a vonzatukat a mondatban, hanem egyértelműen tudják, hogy közvetlenül maguk előtt található. A problémát az jelenti, hogy ha az esetragos névszó elemzésekor azt nem kapcsoljuk össze a névutójával, az esetragja miatt az elemzés – rövid ideig ugyan, de – tévútra juthat. A 11-es példában a már feldolgozott *láttam* keres egy tárgyesetben lévő névszót – az *előadást* tokenhez érve meg is találhatja, és vonzatául veheti. Ezeknek az eseteknek a kiküszöbölésére azt javasoljuk, hogy ne csupán az esetrag nélküli, de a testes esetragot viselő névszók elemzésekor is vizsgáljuk meg az ablak első elemét, hogy nem névutó-e. Természetesen ez az eljárás is több problémát felvet: gazdaságosságának, pontosságának vizsgálata az itt felvázolt algoritmus implementációja után lehetséges.

(11) *Nem láttam sehol az **előadást** követően a főnököt.*

5. Összegzés

Tanulmányunkban a névutószerű elemek részletes, korpuszalapú elemzését mutattuk be. Megvizsgáltuk, hogy a nyelvészeti szakirodalomban is említett, a névutók egyes csoportjait ezekben a tanulmányokban egymástól élesen elválasztani hivatott tulajdonságok a korpuszon vizsgálva miként jellemzik a névutókat. A 2. táblázat eredményeit vizsgálva látható, hogy jól rendszerezhető a számtalan névutó-jelölt ezeknek a tulajdonságoknak a mentén. Találunk prototipikus és perifériás elemeket is, melyeknek az elemzése egy gépi mondatelemzés során eltérő kell, hogy legyen. Három nagy csoport körvonalazódik, melyeknek feldolgozására három különböző algoritmust javasoltunk. A kutatás következő fázisában ezeknek az implementációját kell megvalósítani, hogy kiértékelhessük a teljesítményüket. Emellett fontos konkrét számokat rendelni a 2. táblázat 0 értékű celláihoz, hogy lássuk azok szignifikanciáját, ezáltal tovább pontosíthassuk a névutók besorolását.

Hivatkozások

1. Prószéky, G., Indig, B., Vadász, N.: Performanciaalapú elemző magyar szövegek számítógépes megértéséhez. In Kas, B., ed.: „Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére. MTA NYTI, Budapest (2016) 223–232
2. Prószéky, G., Indig, B.: Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. Alkalmazott nyelvtudomány **15**(1-2) (2015) 29–44
3. Keszler, B., ed.: Magyar grammatika. Nemzeti Tankönyvkiadó (2000)
4. Kiefer, F., Bánréti, Z.: Strukturális magyar nyelvtan: Mondattan. Strukturális magyar nyelvtan. Akadémiai Kiadó (1992)
5. É. Kiss, K.: The Syntax of Hungarian. Cambridge Syntax Guides. Cambridge University Press (2002)

6. Dékány, É.K.: A profile of the Hungarian DP: the interaction of lexicalization, agreement and linearization with the functional sequence (2012) PhD dissertation. University of Tromsø.
7. Marácz, L.: Dressed or naked: the case of the PP in Hungarian. In Werner, Abraham és de Meij, S., ed.: Topic, Focus and Configurationality. John Benjamins, Amsterdam and Philadelphia (1986) 223–252
8. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014)
9. Ligeti-Nagy, N.: Szövegkorpuszok pontosabb annotációja gépi elemzéshez. In: Többnyelvűség és kommunikáció Kelet-Közép-Európában. XXIV. Magyar Alkalmazott Nyelvészeti Kongresszus, Erdélyi Múzeum-Egyesület (2015) 421–429

II. Szemantika, információkinyerés

Magyar szóbeágyazási modellek kézi kiértékelése

Novák Attila, Novák Borbála

Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar
MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Budapest, Práter u. 50/a.
{novak.attila, siklosi.borbala}@itk.ppke.hu

Kivonat A szóbeágyazási modellek a lexikai szemantikai tudásreprezentáció hatékony eszközének bizonyultak, és természetes módon illeszkednek az utóbbi években elterjedt neurálisháló-alapú modellekbe, ezért alkalmazásuk nagyon népszerűvé vált. Ezeket a modelleket leggyakrabban valamilyen gépi osztályozási vagy annotációs feladatot végző rendszer részeként értékelik ki. Ebben a cikkben ezzel szemben magyar nyelvű szóbeágyazási modellek közvetlen kézi kiértékelésének eredményét mutatjuk be. Az eredmények azt mutatják, hogy a morfológiailag elemzett előfeldolgozott korpuszból készített modellek jobb teljesítményt nyújtanak a kézi kiértékelésünk alapját képező szóhasonlósági teszteken, mint a nyers szövegekből készített modellek.

1. Bevezetés és kapcsolódó kutatások

A szavak megfelelő gépi ábrázolása alapvető feladat a természetesnyelv-feldolgozó rendszerekben. Az utóbbi években elterjedt neurális szóbeágyazási modellek hatékony szóreprezentációs formalizmusnak bizonyultak [1,2,3]. Ezen modellek kiértékelésére számos módszert javasoltak. A modellek kiértékelésével foglalkozó tanulmányok egy része a külső kiértékelés mellett foglal állást: a beágyazási modellek által a szavakhoz rendelt vektorokat valamilyen osztályozási vagy annotációs feladathoz használják jegyekként, és azt mérik, hogy a modellválasztás hogyan befolyásolja a rendszer adott feladatban nyújtott teljesítményét. Ez a módszer nem közvetlenül méri a beágyazási modell minőségét, hanem csak annak hatását jellemzi az azt felhasználó rendszer teljesítményére, ez utóbbi azonban jelentős mértékben függ a feladat jellegétől.

A beágyazási modellek kiértékelésével foglalkozó tanulmányok nagy része [4,3] úgy végez kiértékelést, hogy azt mérik, hogy milyen korrelációt mutat az adott modell által egy adott tesztanyagban szereplő szópárokhoz rendelt szóhasonlósági mérték azzal, amennyire az adott (rendszerint korábban elvégzett) kísérletben résztvevő emberek hasonlónak ítélték azokat. Azonban mint Schnabel és társai kimutatták [5], az emberek, illetve a gépi modellek által felállított különböző tartományokba eső pontszámokon alapuló különböző rangsorok nem feltétlenül jól összehasonlíthatóak, és az így kapott pontszámok aggregálásával kapott eredmények félrevezetőek lehetnek.

Problémát jelent például, hogy csak a szópárok egy részhalmazára vonatkozó rangsorolás áll rendelkezésre a kiértékelés alapjául szolgáló anyagokban. Azonban ahhoz, hogy a rangsor-korreláció tényleg megbízhatóan használható metrika lehessen, az olyan szavak közötti hasonlóság emberek szerint kvantifikált mértékére is szükség lenne, amelyeknek „semmi közük egymáshoz". A kiértékeléshez használt szópárlistákkal kapcsolatban azt is problémaként vetik fel, hogy azok nem kiegyensúlyozott mértékben reprezentálják a szókincset. Ezért azt javasolják, hogy szóbeágyazási modellek kiértékelésére olyan hívószólistát használjunk, amely gyakoriság, szófaj és a jelentés absztraktsága szempontjából is kiegyensúlyozott.

A legtöbb tanulmány elsősorban a szóbeágyazási modellek angolra, illetve más általában viszonylag egyszerű morfológiájú nyelvre való alkalmazására koncentrál. Emberek által a szóhasonlóság mértéke szempontjából kiértékelt szógyűjtemények is csak néhány nyelvre állnak rendelkezésre. A viszonylag korlátos szókincs, illetve a viszonylag kötött szórend ezeknek a nyelveknek az esetében lehetővé teszi, hogy a felhasznált egyszerű neurális hálózatok a szójelentést jól megragadó vektortérmodelleket hozzanak létre.

A kiértékelést általában úgy végzik, hogy egy adott korpuszból különböző algoritmussal, illetve különböző paraméterbeállításokkal készített modelleket hasonlítanak össze, és az algoritmusválasztás, illetve a paraméterbeállítások hatását mérik. A morfológiailag komplex nyelvek esetében azonban fontos kérdés az is, hogy a modell mennyire képes az adathiánnyal kapcsolatos problémákat kezelni.

Ebert és kollégái [6] lemmatizált, illetve tövesített korpuszokból készítettek szóbeágyazási modelleket morfológiai szempontból különböző mértékben komplex nyelvekre (többek között magyarra is), és arra az eredményre jutottak, hogy a szóhasonlósági feladatokban az általuk vizsgáltak közül a lemmatizált korpuszból épített modellek teljesítenek legjobban. Kiértékelésüket a WordNetre alapozták. Azt vizsgálták, hogy a modell által az adott szóhoz generált hasonlósági listán hányadik pozícióban szerepel az első olyan szó, amely az adott szó WordNet-beli reprezentációjából két lépésben elérhető valamely a WordNetben szereplő reláció mentén. Ezen pozíció reciprokának átlagát (MRR: Mean Reciprocal Rank) használták a modelleket kiértékelő metrikaként.

Sajnos a magyar WordNet (HuWN) [7] alacsony és kiegyenlítetlen lexikai fedése miatt nem igazán alkalmas arra, hogy megbízható referenciaanyagként szolgáljon a szóhasonlósági modellek kiértékeléséhez. A HuWN 19400 főnévi synsetjének 83%-a tulajdonnév (elsősorban településnév), ugyanakkor a viszonylag gyakori köznevek fedése nem nagyon jó. Egy másik probléma, hogy az angol (2.0) WordNet struktúráját követi, ezért a csomópontok 6.7%-a olyan szellemcsomópont, amihez egyáltalán nem kapcsolódik magyar lexikai elem. Ebert és munkatársai [6] feltehetőleg többek között ezek miatt a problémák miatt kaptak jóval alacsonyabb MRR pontszámokat magyarra, mint más nyelvekre. A 3 részben leírt kísérleteinkben mindenesetre a magyar WordNetet is felvettük a tesztelt erőforrások közé.

Az itt bemutatott kutatásban a célunk az volt, hogy felmérjük, hogy a korpusz előfeldolgozása milyen hatással van a létrejövő modell minőségére. A kor-

puszt, a vektortérmodell létrehozására alkalmazott algoritmust, és annak paramétereit rögzítettük. Ezután részletes közvetlen kézi kiértékelést végeztünk a különböző módon annotált magyar nyelvű korpuszból készült szóbeágyazási modelleken.

2. Magyar szóbeágyazási modellek

Négy modellt építettünk a `word2vec` eszköz CBOW modelljét használva. A korpusz mindegyik esetben ugyanaz az 1,2 milliárd szavas webről gyűjtött korpusz volt. A modellek paraméterei mindegyik esetben a következők voltak: 10 szavas kontextusablak, 300 dimenzió, minimális korpuszgyakoriság: 5.

Első modellünk a tokenizált, de egyébként nyers korpuszból készült (`SURF`). Ebben a modellben felszíni szóalakok szerepelnek. Ez a modell – legalábbis a viszonylag gyakori szavak esetében – a jelentés mellett jól megragadja a morfológiai hasonlóságokat is (amennyire azok a szavak szintaktikai viselkedését is befolyásolják). A morfológiailag analóg *jó – rossz* és *jobb – rosszabb* szópárok hasonlósága jóval nagyobb ebben a modellben, mint a lemma és valamelyik toldalékolt alak pl. *jó – jobb*, illetve *rossz – rosszabb* hasonlósága. Az 1. táblázat első két oszlopa felszíni szóalakok nyers korpuszból épített modellbeli legközelebbi szomszédait mutatja. A *kenyerek* kérdőszóhoz többnyire többes számú péksütemények hasonlítanak legjobban. Ez a modell azonban érzékeny az adathiányproblémára. A *kenyér* szónak 197 különböző ragozott alakja szerepel a korpuszban, a 'kenyér' jelentésreprezentációja ezek között oszlik meg. A `SURF` modell gyakran nem képes arra, hogy a ritka szóalakok jelentését megbízható módon megragadja. Az 1. táblázat második oszlopában szereplő *Vakkalit* szóalakhoz (a Vakkali név tárgyesete) a modell teljesen oda nem illő szavakat rendel hozzá.

kenyerek(2270) SURF	Vakkalit(5) SURF	kenyér(147000) ANA	Vakkali(23) ANA
kiflik(349)	tevedesnek(5)	hús(136814)	Ánanda(321)
zsemlék(283)	áfa-jának(7)	kalács(10658)	Avalokitésvara(39)
lepények(202)	mot-nak(5)	rizs(31678)	Dordzse(270)
pogácsák(539)	Villanysze(5)	zsemle(6690)	Babaji(82)
pékáruk(771)	oktávtól(5)	pogácsa(11066)	Bodhidharma(210)
péksütemények(997)	Isten-imádat(5)	sajt(46660)	Gautama(574)
sonkák(613)	Nagycsajszi(5)	kifli(9715)	Mahakásjapa(25)
tészták(2466)	-fontosnak(7)	krumpli(37271)	Maitreya(426)
kalácsok(277)	tárgykörbôl(5)	búzakenyér(306)	Bódhidharma(115)

1. táblázat. Egy gyakori és egy ritka szó legközelebbi szomszédai a `SURF` és az `ANA` modellből. A zárójeles számok a korpuszbeli előfordulások számát mutatják.

A korpusz morfológiailag annotált változatából **készítettünk egy lemmatizált modellt is (`LEM`)**, hogy a morfoszintaktikai és szintaktikai hasonlóságok

megragadása helyett inkább a szemantikai analógiák megragadására helyezzük a hangsúlyt, illetve hogy enyhítsük a fentebb említett ritka szavakkal kapcsolatos adathiányproblémát. A lemmákat is tartalmazó annotációt a magyar Humor mofológiai elemzőt [8,9] használó PurePos taggerrel [10] készítettük. Az annotációból csak a lemmákat hagytuk meg, és ezek sorozatán tanítottuk be a szóbeágyazási modellt.

A [11] cikkben leírt módon **létrehoztunk egy másik elemzett korpuszon alapuló modellt is (`ANA`).** Ebben a modellben nemcsak a lemmák szerepelnek, hanem minden eredeti tokent két token reprezentál: a lemmát a szó morfoszintaktikai címkéje külön tokenként követi.

Az alábbi példa a *Szeretlek, kedvesem.* mondat így előfeldolgozott változatát mutatja:

```
szeret [IGE][Ie1] , [PUNCT] kedves [FN][PSe1][NOM]
```

Mivel ebben a reprezentációban a címkéket a hozzájuk tartozó szó mellett tartottuk, a `LEM` modellel ellentétben a címkék által hordozott morfoszintaktikai információ továbbra is szerepet játszott a vektorreprezentáció meghatározásában. Ugyanakkor az adathiányt jól kezeli ez a reprezentáció, mert a szó különböző ragozott alakjait egyetlen lemma képviseli. Az 1. táblázat második két oszlopa az `ANA` modell által generált legközelebbi szomszédokat mutat be. Látható, hogy ez a modell a `SURF` modellnél jobban meg tudja ragadni a ritka lexikai elemek jelentését, mert míg a lemmatizálás enyhíti az adathiányproblémát, a környezetben jelenlévő morfoszintaktikai annotáció megőrzi a grammatikai információ túlnyomó részét (v.ö. a táblázat 2. és 4. oszlopát). A *Vakkali* szóhoz az `ANA` modellben legközelebbi szavak világosan jelzik, hogy ez a modell képes volt annak a ténynek a kihámozására a korpuszból, hogy egy buddhista személyiségről van szó.

Az `ANA` modell egy apró módosításával egy még **kifinomultabb elemzett modellt hoztunk létre (`POS`).** A fő szófajcímkét a lemmán hagytuk, és csak a morfoszintaktikai címke többi részét választottuk le külön tokenként.

Az alábbi példa a *Szeretlek, kedvesem.* mondat így előfeldolgozott változatát mutatja:

```
szeret[IGE] [Ie1] ,[PUNCT] kedves[FN] [PSe1][NOM]
```

Ez a modell szófajonként külön reprezentációt hoz létre a különböző szófajú homonim szavakhoz (pl. *vár, nyúl, csűr, reggeli, ír* stb.), ezzel részben kezeli azt a problémát, hogy a szóbeágyazási modelleket létrehozó algoritmusok többsége nem tud értelmes módon mit kezdeni a homonímia/poliszémia jelenségével. A CBOW/skipgram modellek által létrehozott vektorterekben ugyan valamilyen értelemben sokszor visszanyerhető a több jelentést egybeejtő vektorreprezentációból a több jelentés, de hogy ez mennyire sikerülhet, az nagy mértékben múlik azon, hogy hány jelentése van az adott lexikai elemnek, és hogy azoknak a korpuszbeli jelenléte mennyire kiegyensúlyozott.

3. Kísérletek

Mivel – ellentétben az angollal – olyan magyar nyelvű erőforrások nem léteznek, amelyekben anyanyelvi beszélők jellemeztek volna szópárokat a jelentéshasonlóságuk mértéke szempontjából, és amelyeket a szóbeágyazási modellek kiértékeléséhez közvetlenül lehetne használni, ezért más módszert kellett alkalmaznunk. Egy olyan webes felületet hoztunk létre, melyen keresztül összehasonlító értékeléseket gyűjtöttünk a különböző modellekből kinyert és a felhasználó számára bemutatott példák alapján [5]. A kiértékelésben résztvevőket arra kértük, hogy rangsorolják a modelleket az alapján, hogy egy adott hívószóhoz kinyert kapcsolódó szavak listája mennyire releváns. Az `ANA`, `POS` és `LEM` modellek mellett a `SURF` modell egy módosított változatát (l. alább), egy szabadon hozzáférhető, nyers magyar nyelvű szövegeken tanított skip-gram modellt [12] és egy a magyar WordNeten alapuló modellt is bevettünk a kiértékelésbe.

A nyers korpuszból épített modellek (pl. `SURF`) a szavak toldalékolt alakjait is tartalmazzák, ezek megjelennek a lekérdezésekre kapott válaszokban is. Annak érdekében, hogy ezek a szólisták (amikben tehát egy szóhoz a legközelebbi szavak nagy része ugyanannak a lemmának a különböző alakjai) összehasonlíthatóak legyenek az előfeldolgozott korpuszból létrehozott, csak lemmákat tartalmazó modellből kinyert szólistákkal, a nyers modellek kimenetén utólag alkalmaztunk tövesítést. Az adott hívószóhoz az n legközelebbi szót pedig úgy adtuk meg, hogy az ismétlődő lemmák eliminálása után kapjunk n hosszú listát (`SURFL`). A [6] cikk szerzői hasonló transzformációt alkalmaztak a nyers és lemmatizált modellek összehasonlításához.

Az általunk létrehozott modelleket egy 1,2 milliárd szavas magyar webkorpuszból építettük. A referenciaként használt szabadon letölthető modell egy jóval nagyobb, 4,6 milliárd szavas nyers korpuszból készült [12]. Míg a mi modelljeink tanítása során a CBOW architektúrát alkalmaztuk, addig a nagyobb korpuszból épült modell létrehozása a skip-gram architektúrával történt, és 200 dimenziós vektorokat tartalmaz. Ennek a modellnek szintén az utólag tövesített eredményét vettük be a kiértékelésbe (`SGL`).

A magyar WordNet-et, mint kézzel ellenőrzött szemantikai adatbázist vettük be a kiértékelésbe. [6] alapján minden tesztszóhoz kinyertük azokat a szavakat, amik valamilyen WordNet reláció mentén három lépésen belüli távolságban voltak az adott szótól. Ezt a listát a távolság alapján rendeztük sorba úgy, hogy a vizsgált szó szinonimáit 0 távolságra lévő szavaknak tekintettük.

3.1. A lekérdezések

Schnabel és munkatársai egy olyan 100 szóból álló angol szólistát állítottak össze angol nyelvű szóbeágyazási modellek kiértékeléséhez, ami a lekérdezések kiindulásaként használt szavakat korpuszbeli szógyakoriság, a jelentés absztraktsága, illetve szófaj szempontjából kiegyensúlyozott arányban tartalmazza [5].[1] Ezt a

[1] Az erőforrás elérhető online a `http://www.cs.cornell.edu/~schnabts/eval/` oldalon.

szólistát magyarra fordítottuk úgy, hogy minden szónak a megfelelő szófajú és ugyanabba a jelentéskategóriába tartozó megfelelőjét választottuk, majd ellenőriztük, hogy a magyar nyelvű korpuszban hasonló gyakorisággal szerepel-e. Ha a fordítás nem ugyanabba a tartományba esett, akkor egy másik, hasonló jelentéscsoportba tartozó, ugyanolyan szófajú, de megfelelő gyakoriságú szót választottunk a korpuszból. Ezáltal egy, az eredeti szólistához hasonlóan kiegyensúlyozott magyar nyelvű listát kaptunk. Ezt a listát kiegészítettük még 7 további olyan homonim szóval, amiknek a különböző jelentései különböző szófajúak (pl. *vár*, *reggeli*).

3.2. A rangsorolási feladat

Magyar anyanyelvű annotátorokat kértünk meg arra, hogy a létrehozott webes felületet használva rangsorolják a hat modell által generált válaszokat (`ANA`, `SURFL`, `POS`, `LEM`, `SGL`, `WN`). A kiértékelésben résztvevő felhasználóknak minden körben az előre összeállított listából egy hívószót és az egyes modellekből 1, 5, vagy 9 kapcsolódó szót mutattunk meg. A modellek neve természetesen rejtve volt a felhasználók elől, és a megjelenítés sorrendje is véletlenszerűen változott. Az 1, 5, vagy 9 elemű megjelenített szólisták az egyes modellek szerint a vizsgált szó k legközelebbi szomszédját tartalmazó listák 1, 5, illetve 9 egymás utáni elemét tartalmazó részei voltak, vagy a távolság szerint rendezett lista első elemétől, vagy a 30. elemétől kezdve (egy konkrét kérdés esetén mindig ugyanabból a tartományból). Ez alól egy esetben tettünk kivételt: mivel a WordNet-ből létrehozott modellben a kapcsolódó szavak száma a 107 vizsgált szóból 97 esetben 30-nál kevesebb volt, ezért az olyan kérdéseknél, amikor a többi modellből a 30. pozíciótól számítva jelenítettük meg a kapcsolódó szavakat, a `WN` modellt kihagytuk a rangsorolási feladatból. A magyar WordNet fedése egyébként is elég alacsony volt, a kísérlet során vizsgált 107 tesztszó közül csupán 70 szerepelt benne (65%). Ha egy hívószó esetén valamelyik modell nem adott választ, akkor üres listát jelenítettünk meg a felhasználó számára.

Az annotátoroknak a megjelenített listákat rangsorolni kellett az alapján, hogy mennyire érezték a hívószó és az egyes listák elemei között a kapcsolódást. A rangsoroláskor 1 és 99 közötti pontszámot adhattak (minél magasabb, annál jobb a lista minősége), az üres listákhoz automatikusan 0 pontot rendeltünk és megengedtük az egyenlő pontszámokat is.

Mivel a hasonlóság mértékének meghatározása szubjektív feladat, ezért ha egy annotátor nem tudott egy kérdésben rangsort felállítani (az is előfordulhatott, hogy a ritka szavak esetén a hívószó jelentését sem ismerte), akkor válaszadás nélkül továbbléphetett a következő kérdésre. A válaszok összesítésekor az egyes listákhoz rendelt pontszámok értékét figyelmen kívül hagytuk, csupán a rangsort vettük figyelembe. A kiértékeléshez használt felület és az annotátorok számára adott instrukciók az 1. ábrán láthatók.

A válaszok kiértékeléséhez a nyerési arány metrikát használtuk. Ez azt fejezi ki, hogy egy adott modell az esetek mekkora részében kapott magasabb pontszámot egy másik modellnél. Ez tehát a modellek páronkénti összehasonlítását jelenti minden egyes tesztkérdés alapján a felhasználók által adott rang-

Mennyire hasonlít?

Ebben a kísérletben annak megítélésére kérjük, hogy az alább megjelenő egyes listákban szereplő szavak **jelentése és használati köre mennyire hasonlít a kiemelt hívószó jelentésére illetve használatára.**

A "lista" időnként csak 1, máskor maximum 5, illetve maximum 9 elemet tartalmaz. **Minél jobban hasonlítanak** egy-egy listában szereplő szavak jelentésük és használatuk tekintetében a kiemelt szóhoz, **annál nagyobb pozitív egész számot** írjon be az adott lista felett levő mezőbe az 1-99 tartományban. (Időnként üres lista is megjelenik, ehhez a rendszer automatikusan 0 értéket rendel, amit nem lehet módosítani.)

Ha nagyjából **azonosan jónak** ítél két listát, akkor írjon hozzájuk **azonos számot.** Ha a jelentés ugyan hasonlít, de a szót más szövegkörnyezetekben lehet használni, mint a hívószót (pl. más a szófaja), akkor azt minősítse kevésbé hasonlónak, mint ha a használata is azonos az adott szavaknak. Az ellentétes jelentésű szavakat (pl. *jó-rossz*) minősítse hasonlóbbnak, mint azokat, amiknek a jelentése nem csak abban különbözik, hogy egy adott tulajdonság két ellentétes pólusát nevezik meg. Ha a szó jelentése nagyon hasonlít és a használata (szófaja) azonos (pl. *fut~szalad*), azt minősítse hasonlóbbnak, mint ha a jelentés azonos, de a szavak használata (pl. szófaj) különbözik (pl. *fut~futás*).

Ha túl nehéz megítélni a sorrendet, az összes mező üresen hagyásával átugorhatja a kérdést.

A munka megkezdésekor az e-mail cím mezőbe írja be az e-mail címét, ez alapján azonosítani tudjuk az azonos annotátortól kapott válaszokat. **A legtöbb értékelhető választ beküldőket szeretnénk megjutalmazni,** az erről szóló értesítést a megadott e-mail címre fogjuk küldeni. (Ezért érdemes érvényes e-mail címet megadni. :)) Értékelhető válasznak az számít, ahol a rangsor tényleg a szavak hasonlóságát tükrözi.

Köszönjük a segítségét.

e-mail cím jani@gmail.com
Rangsor

Submit

terem[IGE]

terménytároló	könyvtárszoba	csüng	bárpult	nő	alagsor
udvar	tanácskozóterem	hever	virágágyás	tenyészik	hálószoba
malom	társalgó	elhervad	bódé	megterem	kerthelyiség
veranda	klubhelyiség	csirkeól	fogadócsarnok	hoz	pajta
pince	folyosórész	termel	díszterem	növekszik	lugas

1. ábra. A kiértékeléshez használt felület

sor szerint. Minden „győzelem” egy pontot ér, a végső pontszám pedig a annak az aránya, hogy az összes összehasonlításból hányszor került ki győztesként az adott modell. Ezt a módszert alkalmazzák gépi fordító rendszerek emberi kiértékelése során is, mivel ez bizonyult a legmegbízhatóbb módszernek arra, hogy egymástól egészen különböző rendszerek minősége összemérhető legyen [13]. Az összesítő kiértékelést minden paraméter szerint külön-külön is elvégeztük, így a modellek minőségét a hívószavak gyakorisága, szófaja, szemantikai kategóriája, és a válaszlisták kezdőpozíciója (1 vagy 30) szerint is vizsgálni tudtuk.

Bár a felkért annotátorok magyar anyanyelvű felhasználók voltak, a valódi válaszadás megkezdése előtt egy tesztfázison kellett átesniük. A 107 elemű listán felül kiválasztottunk 5 olyan szót, amire a modellek válaszai egyértelműen rangsorolhatók, így egy gold standard-nek tekinthető választ várhattunk ezekre. Ezt az 5 szót minden felhasználónak az első 10 kérdés között megjelenítettük. Ha egy annotátor ebből az 5 kérdésből legalább 4-re nem tudta a helyes sorrendet felállítani, akkor az ő eredményeit nem vettük figyelembe a kiértékelésnél.

A kísérlet során végül 15 annotátortól kaptunk használható válaszokat, akik mind legalább 14, legfeljebb 102 különböző kérdésre adtak választ. Minden egyes tesztszóra legalább 3-3 rangsorolás érkezett, mind az 1., mind a 30. pozíciótól számított listák alapján.

4. Eredmények

A kiértékelés eredménye az 2. ábrán látható. Az egyik legegyértelműbb tény, ami az ábrán jól látszik, hogy a referenciaként használt 200 dimenziós vektorokból álló skip-gram modell (`SGL`) teljesített legrosszabbul szinte mindegyik kiértékelési szempont esetén ($p < 0.05$). Ez alól kivétel csak a melléknevekre és a nagyon

ritka szavakra (gyakoriság ≤ 500) szűkített eredmény, ahol a `WN` modell csaknem ugyanolyan rossz eredményt produkált (az alacsony fedés miatt), illetve a szófajon átívelő többértelműség esete (lásd a 2d. ábra elemzését később). Az `SGL` modellből a gyakori szavakhoz kapott válaszlisták is sok olyan szót tartalmaztak, amik nem, vagy nagyon gyengén kapcsolódnak a hívószóhoz (pl. a *kenyerek* szó esetén a *búzalisztből, sütéshez, karfiolból* szavak), a ritkább szavakhoz pedig sokszor teljesen értelmetlen eredmények szerepeltek a listában.[2]

A modellek összesített teljesítménye a,z 2a ábra *all frq* oszlopában látható. Mindegyik általunk készített modell jobban teljesített, mint a magyar WordNet. Ez megerősítette azt a feltevésünket, hogy bár az angol szakirodalomban népszerű a WordNet-hez viszonyított kiértékelés, a magyar WordNet azonban nem alkalmas arra, hogy a beágyazási modellek kiértékelésének alapjául szolgáljon: a fedése ehhez nem elég jó.

Ahogy a 2a. ábrán látható, az előfeldolgozott és lemmatizált korpuszból épített szóbeágyazási modellek jobban teljesítettek a ritka szavak esetén, míg a gyakori szavak esetén kevésbé érvényesült az előfeldolgozás hatása.

A listán belüli kezdőpozíciót, mint paramétert megfigyelve (2b. ábra) az látható, hogy az `SGL` modell teljesítménye tovább romlik, ha a kapcsolódó szavak listájában hátrébb szereplő elemeket vizsgáljuk. Megfigyelhető továbbá, hogy az `ANA` és a `POS` modellek minősége a 30. pozíciónál jobb, mint a `LEM` modell minősége. Míg a lemmatizált modell jól kezeli az adathiány problémáját (lásd a ritka szavakra mutatott teljesítménynövekedést a nem lemmatizált modellekhez képest), ebben a modellben a szemantikai hasonlóság elnyomja a grammatikai szempontból való hasonlóság szempontjait. A `LEM` modellből kapott válaszlisták gyakran változatos szófajú elemeket tartalmaznak. Az `ANA` modell ezzel szemben megőrzi a grammatikai (szintaktikai) relációkat is, mivel az előfeldolgozás során megtartottuk a morfoszintaktikai tulajdonságokat kódoló címkéket a szavak környezetében. Ezért a létrejött szóbeágyazási vektorok tanítása során ez az információ is érvényesülhetett. A `POS` modell ezen felül a szófaj alapján megkülönböztethető homonim szóalakok között is különbséget tud tenni. Ezek a modellek tehát egy jó minőségű köztes megoldást biztosítanak a nyers korpuszból tanított szóalapú modellek (ahol a morfoszintaktikai viszonyok ugyan jól megjelennek, de a ritkább szavak esetén az adathiány rontja a modell minőségét) és a szótő alapú modellek között (ahol szinte kizárólag szemantikai szempontok érvényesülnek). Fontos megjegyezni, hogy az általunk létrehozott modellek tanítása azonos paraméterekkel történt, ez mégsem ugyanazt jelenti, hiszen az `ANA` és a `POS` modellek esetén a környezetet figyelembe vevő ablakban valójában fele annyi szó szerepel, hiszen minden szót két tokenként reprezentáltunk a korpuszban. A fele akkora környezet ellenére a modellek minősége nem rosszabb, mint a nagyobb környezetet figyelembe vevő `LEM` és `SURF` modelleké.

[2] Mivel ezt a modellt nem mi készítettük, nem egyértelmű, hogy pontosan milyen tényezők játszhattak szerepet abban, hogy ennyivel gyengébb teljesítményt nyújt a többinél. Nem valószínű, hogy az lenne az oka, hogy a CBOW helyett a skipgram architektúrát használták a modell készítésekor.

(a) gyakoriság (b) pozíció

(c) absztraktság (d) szófaj

2. ábra. A modellek teljesítménye (nyerési arány) különböző szempontok szerint vizsgálva

Ami a jelentés absztrakt vagy konkrét voltát illeti, a modellek teljesítményében ez a szempont nem nagyon játszik szerepet (2c. ábra), hasonlóan teljesítenek mindkét kategóriában. A szófajok szerinti bontásban azonban vannak különbségek (2d. ábra). Egyértelműen látszik, hogy az elemzett korpuszból tanított modellek jobban teljesítenek igék és főnevek esetén. Ez nem meglepő, hiszen ezek a leggyakrabban toldalékolódó szófaji kategóriák. A melléknevek esetén azonban a nyers szövegen tanított modell teljesített jobban.

A homonim alakokra a `POS` modell – mint az várható volt – a mezőnyből messze kiemelkedő eredményt nyújtott (2d. ábra *posamb* oszlopa). Ebben a kategóriában a `WN` modell lett a második legjobb, ami annak köszönhető, hogy a

WordNet-ben (a POS modellhez hasonlóan) szerepel szófaji információ, tehát az azonos alakú, különböző szófajú szavak jól megkülönböztethetők. A WordNet alacsony fedése miatt azonban a POS modell jobban teljesített. Érdekes továbbá, hogy ez volt az egyetlen olyan szempont, amely szerint a LEM és ANA modellek teljesítménye gyengébb volt, mint a nyers korpuszból tanított modelleké (SGL és SURFL).

5. Konklúzió

Cikkünkben egy magyar nyelvű szóbeágyazási modellek kiértékelésére irányuló kísérlet eredményeit foglaltuk össze. A kiértékelésben 6 modellt hasonlítottunk össze, melyek közül négynek a létrehozását is részletesen bemutattuk, a különböző előfeldolgozási lépésekkel együtt. A további két modell közül az egyik egy mások által más anyagon és más architektúrán betanított, általunk csak a kiértékelésben felhasznált szóbeágyazási modell, a másik pedig a magyar WordNeten alapuló modell volt. A modellek kiértékelését egy webes felületen keresztül humán annotátorok által adott rangsorok alapján végeztük.

Az eredmények alapján általánosságban elmondható, hogy a morfológiai annotációt tartalmazó korpuszból tanított modellek jobban teljesítettek, mint a nyers korpuszból tanított modellek, enyhítve az agglutinációból fakadó adatritkaság által okozott problémákat. A bemutatott ANA és POS modellek az egyszerűen lemmatizált modelleknél is jobb teljesítményt nyújtottak, annak ellenére, hogy a tanítás során kisebb kontextust vettek figyelembe a szóvektorok létrehozásakor, mert ezeknek a modelleknek a betanításakor a morfoszintaktikai információ jelen volt az adott lexikai elem környezetében, így a grammatikai hasonlóságok megragadására alkalmasabbak a pusztán lemmatizált modellnél.

Köszönetnyilvánítás

Jelen kutatás az FK 125217 és a PD 125216 számú projekt keretében a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással az FK 17 és a PD 17 pályázati program finanszírozásában valósult meg.

Hivatkozások

1. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
2. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. (2013) 746–751
3. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, Association for Computational Linguistics (2014) 238–247

4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. (2013) 3111–3119
5. Schnabel, T., Labutov, I., Mimno, D.M., Joachims, T.: Evaluation methods for unsupervised word embeddings. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y., eds.: EMNLP, The Association for Computational Linguistics (2015) 298–307
6. Ebert, S., Müller, T., Schütze, H.: LAMB: A good shepherd of morphologically rich languages. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA (2016)
7. Miháltz, M., Hatvani, C., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the Hungarian WordNet project. In: Proceedings of The Fourth Global WordNet Conference. (2008) 311–321
8. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
9. Novák, A.: A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 1068–1073 ACL Anthology Identifier: L14-1207.
10. Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, Incoma Ltd. Shoumen, Bulgaria (2013) 539–545
11. Siklósi, B.: Using embedding models for lexical categorization in morphologically rich languages. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, Springer International Publishing, Cham. (2016)
12. Szántó, Z., Vincze, V., Farkas, R.: Magyar nyelvű szó- és karakterszintű szóbeágyazások. In Tanács, A., Varga, V., Vincze, V., eds.: XIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2017) 323–328
13. Bojar, O., Ercegovčević, M., Popel, M., Zaidan, O.F.: A grain of salt for the wmt manual evaluation. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. WMT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 1–11

The World is Built with our Words to Each Other Basic and Fine-Tuned Intensional Profiles in Hungarian

Anna Szeteli, Gábor Alberti, Judit Kleiber, Mónika Dóla

University of Pécs, Research Team ℜeALIS for Theoretical, Computational and Cognitive Linguistics
anna.szeteli@gmail.com, {alberti.gabor, kleiber.judit, dola.monika}@pte.hu

Abstract: Our paper provides some theoretical background to a program which operates on a continuously changing world-model with possible speakers and listeners speaking about the world and each other. The agents' information states are also continuously changing depending on changes of the model of the outer world, including their messages sent to each other, and each other's information states. As a groundwork, we identify the pragmasemantic components of some basic sentence types and discourse markers compositionally. Then we present how speakers with their psychological egos can be separated from linguistically conventionalized addresser roles and how many pragmasemantic phenomena can be captured through pattern matching between addressers' conventional profiles and the corresponding speakers' information states, including a few elements of politeness. In short, the program is ultimately designed to simulate human intelligence through modeling human communication and language-based cognition in order to improve our theoretical background on the basis of the functioning of the program. And *vice versa*, we intend to improve the "machine" by building in its information treatment mechanisms as much language-based human intelligence as possible.

Keywords: formal pragmasemantics, discourse representation, discourse markers, pattern matching, politeness

1 Introduction

This paper investigates the five basic sentence types and their possible "fine-tuning" by discourse markers in Hungarian, also taking into consideration the effect of the utterances on the world-model and on the interlocutors. In any given situation, an utterance can only be performed by a speaker who is in an appropriate information state to perform it, and, most of the time, a suitable listener is also required. There is a program in preparation—based on ℜeALIS—which aims to present how the success of the conversation is influenced by conventional meaning and the information state of the speaker and of the listener (see Section 5). First, we describe the roots of the Austinian model [1] felicity conditions and their relevance to the formal evaluation of utterances, then we introduce the theoretical framework ℜeALIS 'Reciprocal And Lifelong Interpretation System' [2]. The framework is based on Kamp's Discourse Representation

Theory [3]. It provides an expansion of the Kampian discourse representation to mind representation and is capable of a compositional handling of the linguistically relevant data.

Our primary goal is to define the pragmasemantic content, or 'intensional profile', of different sentence types and discourse markers, and – since intensional profiles can disambiguate between similar meanings or functions of the same utterances – we also aim to establish the relationship between the various linguistic elements under analysis. We claim that this task requires a formal dynamic discourse and mind-representation.

The software itself can be regarded as an improved version of Anton Benz's [4] multi-agent system [5]. Benz argues that the general apparatus for multi-agent system provides us with a natural representation of the information of dialogue participants on events in the outer world, but each participant should be assumed not only to update his own DRS if he gets some new information, but also to update a DRS representing the knowledge of different groups which *commonly* got this information. According to the even more sophisticated approach that ℜeALIS offers, each dialogue participant should potentially have asymmetrical information on others' knowledge, hypotheses, desires, intentions, and not only on the outer world but also on others' continuously changing information states, recursively. This way ℜeALIS captures what cognitive scientists call mentalization [6].

As our software application inherently belongs to a radically new and holistic "pragmalinguistics" theory, it is uneasy to compare its background to software applications based on some different theoretical foundation. An exception is the SDRT-based [7] experimental software dialogue system, RUDI [8], primarily due to its distinguished attention to the relationship between pragmatic phenomena and the external-world model. RUDI ("Resolving Underspecification with Discourse Information") automatically computes some aspects of the content of scheduling dialogues, particularly the intended denotation of the temporal expressions, the speech acts performed and the underlying goals. Following SDRT, it is assumed that a dialogue is coherent just in case every proposition (and question and request) is *rhetorically connected* to another proposition (or question or request) in the dialogue (NB: virtually anomalous conversations can be regarded as conversations with parts linked by very special rhetorical relations partially based on implicit knowledge in the background). The rhetorical relations are viewed as *speech act types* in the RUDI project—that is the point where our ℜeALIS-based project can be regarded as an extension of RUDI, given that in the ℜeALIS theory further relations among pieces of information stored in minds, addressers, addressees, contexts and the external world are (intended to be) taken into account in a completely uniform system.

ℜeALIS essentially follows SDRT, which represents discourse content as a "segmented discourse representation structure", which is a recursive structure of labelled DRSs, with rhetorical relations between the labels. In contrast to traditional dynamic semantics, SDRT attempts to represent the *pragmatically preferred* interpretation of a discourse—just like ℜeALIS. The rule schema used in RUDI contrasts with the plan-recognition approach to computing speech acts [9], which uses *only* the goals of the antecedent utterance, rather than its compositional and lexical semantics directly, to constrain the recognition of the current speech act.

2 Pattern-Matching in Extensional and Intensional Evaluation

Already Austin [1] pointed out that no sequence of utterances can be either true or false, they can only be regarded as "felicitous" or "infelicitious" according to given sets of conditions. Following Oishi [10], we make a distinction between the information that is linguistically encoded in the utterance and which elaborates a complex relation between the addresser's (AR) and the addressee's (ae) conventional beliefs (B), desires (D) and intentions (I) from what the speaker of the utterance actually believes, desires and intends to do in the real world. Every single time when the flesh-and-blood speaker (sp) with his/her information state makes an utterance, he/she takes on the role of the addresser and hands over the role of the addressee to the listener (li).

Thus, our framework makes a clear distinction between the addresser and the speaker of the actual discourse. Similarly, the addressee is distinguished from the listener, and so is the linguistically defined appropriate context from the factual situation in which the interlocutors are talking to each other.

In the theoretical framework ℜeALIS [11] the speaker's information state and the representation of the discourse are defined in the speaker's mind, which is contained in the world-model. Thus, separating the addresser role from the speaker's information sate is not a simple "technical" separation of variables from values but this is the way how the framework can treat numerous complex pragmasemantic phenomena such as lying or withholding information [12].

The implementation of the compositional meanings and the pattern-matching mechanism should be able to present, in a simulated world-model which contains possible speakers and listeners, that the discourse representation of ℜeALIS can capture new properties of a basic discourse.

3 Multiple Worldlets of the Interpreter

Let us present the pragmasemantic point of view above in a more formalized way according to ℜeALIS [11][13]. The *eventuality* (e) or the *propositional content* of an utterance is encapsulated in an intensional profile, which consists of worldlets. Wordlets are labeled DRS-like structures, and all of them can be captured by the prism-effect formula:

$$\mathscr{R}([\mathscr{P}(M\times I\times R\times T\times P)]^*)$$

The five labels indicate the modality (M), the intensity (I), the referent (R) that is the host of the worldlet, the time parameter (T), and the polarity value (P={+,–,0}).

The modality, that is, the attitude to the given content, can be a belief (B), a desire (D), an intention (I), authority (A), experience (E) or a mixture of these as the inner power set symbol '$\mathscr{P}$' shows.

Intensity can be maximal (M) or non-maximal (nM). Non-maximal intensity can be, for example, almost maximal (aM), great (gr) or some (sm). The number of the linguistically encoded intensity degrees can differ in various languages, and we often formulate the values with a number between 0 and 1, as customary in probability theory.

The referent who has the worldlet is mostly the addresser (AR) or the addressee (ae) or both. The set R contains this/these referent(s).

Set T contains the moments at which the worldlet is in the interpreter's mind.

The P={+,–,0} component of the formula (true (+), false (–), not specified (0)) yields eight possible polarity values, as a result of the above mentioned power set symbol.

The Kleene-star indicates the possibility of recursion, because, for example, the interlocutor can desire to acquire a piece of information about the content which is a belief on someone's intention to persuade another person to do something... Finally, the first power set symbol indicates that the interlocutor may have beliefs, desires, intentions etc. about the content at the same time.

4 Amplified Felicity Conditions Encoded in Intensional Profiles

In this part we introduce the ideal intensional profiles and show a few points in the speaker's and in the listener's information state where the failure of pattern-matching points out that the success of the conversation is impossible; which means that it cannot change the world-model in the intended way.

Let us first consider two basic intensional profiles and a fine-tuned one presented in Table 1. First of all, however, it should be noted that it would go far beyond the scope of this paper to attempt to argue for the pragmatic perfection of the profile elements presented. It is in the ℜeALIS-papers referred to in this paper that the establishment of the profiles is a systematically completed task, at least according to the introspective tradition. Here the reader is expected to accept on the basis of the symbols explained in Section 3 that

(i) both the basic sentence types and those fine-tuned by certain discourse markers are conventionally associated with definite information on definite interlocutors' definite beliefs, desires and intentions on definite things that the given sentence decides,

(ii) it is possible to formally capture these linguistically conventionalized pieces on information practically on the ideal circumstances (of the outer world and, primarily, the speaker's information/mental state), and

(iii) as the worldlet labels responsible for capturing this knowledge are nothing else but sequences of quintuples of the symbol types demonstrated in Section 3, it is realistic to write a computer program in which agents' mental/information states are represented as mappings (which are changing from state to state) between a few eventualities in the outer world and sets of worldlet labels.

The exclamative sentence is analyzed as the one which *ab ovo* does not depend on the addressee, as presented in the first column of Table 1. It shows the relation between the addresser's belief and desire states, which are required to appear with different polarity values and there must be a strong desire concerning to the eventuality. The actual belief should be positive (negation is regarded as a discourse marker fine-tuning the hosting basic types). The addresser knows that the eventuality has happened, but in a former moment (s)he did not know about its content, or thought that it had not happened yet. On the other hand, it is "enough" to perform the exclamative sentence if (s)he is very (un)happy because of the fulfillment of the propositional content. In short, coming to know something upsets the addresser.

The ideal purpose of the speaker with taking on the addresser's role of the basic declarative sentence type is stating facts. This follows from the Cooperative Principle of Paul Grice [14] because the Maxim of Quality specifies that the addresser should have enough evidence and should not make a statement if (s)he thinks that it is false. There are two axioms which show the characteristics of the ideal addressee. The addresser should be relevant, so (s)he can state a fact if (s)he thinks that the addressee does not know about it, and, as the third axiom says, one of them should have a relevant desire. The first possibility is that the addressee is assumed to intend to learn the fact, but in the current version of ℜeALIS [13][15], this formula has been generalized in order to capture the case when the addresser is interested in forcing the given piece of knowledge on his partner. (S)he wants her to learn the given fact. In both cases the intention of the addresser is to tell the fact. Moreover, there is an authority axiom encoded in the declarative utterance: the addressee thinks that (s)he can make a statement about fact e, because it is not a secret or there are no moral barriers. As we demonstrated, there are some conditions which depend on the addressee. Hence, the first pattern-matching is between the addresser's factual thoughts on the listener and the role (s)he attributes to her. If we want to know more about the success of the conversation, there should be a second pattern-matching between the applied intensional profile of the addressee and the factual information state of the listener.

The other declarative intensional profile contains a discourse marker *szerintem* which "fine-tunes" the basic sentence-type. The background of the alteration is that the addressee does not have direct evidence about fact e. There are sentences – for example *Szerintem Péter otthon van.* 'In my opinion, Peter is at home.' – in which the predicate is evaluable in the world-model, but in other sentences – for example, *Mari gyönyörű.* 'Mary is beautiful.' – the basis of the inference is very complex and depends on the speaker's mind; it is rather a judgment [15]. The Maxim of Quality dictates that the lack of evidence should be indicated in the profile, so the intensity of the belief is ι and the tripartite components in the second column show that the addresser does not supply the ultimate knowledge but presents that it is her/his ι-strong belief. With the communicative success of the addresser, the addressee will have a belief that *according to the addresser/the addresser states that: the fact is true*, and not a belief that the *fact is true*. We should note at this point that the listener can hold the result-belief of a basic declarative sentence also in these two ways.

Let us continue the review of profiles given in Table 1 by considering the interrogative sentence type and its two related profiles. In contrast to the declarative sentence, a speaker can take the addresser's role of the interrogative one when (s)he does not have the maximal belief about the fact. As the second axiom shows, (s)he should have a belief that the addressee does not believe about her/him that the content is well known for her/him. It is a higher level of mentalization, but the encoding was very motivated, because if the addresser thinks that according to the addressee (s)he should know about the content, (s)he should explain why (s)he does not know about it. Then (s)he knows that (s)he wants to learn the fact, or (s)he thinks that the addressee wants to tell her/him it. (S)he supposes that the addressee can tell the answer and the required axiom for this is that (s)he believes that the addressee knows the answer. The stricter, sufficient axiom is that the addressee has the maximal authority to tell the truth. The intention of the addresser is to arouse the intention of the addressee to tell her/him the right polarity value (i.e. truth value) of the propositional content. If a speaker asks a listener in the

world-model who does not know the fact, it is a failure of the pattern-matching between the listener's mind and the addressee-profile. In this particular case, the speaker cannot improve her/his knowledge about the eventuality.

The discourse marker *ugye* improves the profile without any conflict with the basic intensional profile [13], which is the ideal realization of pragmasemantic compositionality. There is a worldlet, underspecified in its modality, which expresses that the addressee has a strong belief and/or desire that the eventuality is true, but her/his belief is not maximal.

There is another discourse marker, *vajon*, which can have an *ab ovo* addressee-independent profile without the *italicized* axioms. The addresser is speculating, (s)he does not know the truth about the fact, and (s)he knows that (s)he also will not know it in the next moment (τ^+). If (s)he takes an addressee, by eye contact, for instance, (s)he thinks about her that she has not got any knowledge either. It is a failure of the pattern-matching between the speaker's mind and the profile if (s)he asks someone who must know the truth on a high level about the fact with "fine-tuned" by *vajon*.

In Table 2 there is another Janus-faced sentence type which can be interpreted without the *italicized* axioms. The optative sentence expresses that the belief of the addresser has a negative polarity value but (s)he has a positive desire and no authority [16]. If an addressee is taken (by eye contact, for instance), the intention will be to present the addresser's desire, and there appears a presupposition that the addressee also knows that the eventuality has not happened.

The imperative sentences all have these two presuppositions, which stand in the first and the second row of the table. Sentences which are not imperative compositionally but have these presuppositions due to other elements can begin to be functioning as imperatives.

We claim that, in addition to the pattern-matching mechanism sketched above, there is a potential accommodation-process available to the listener, which also influences the perlocution and in this manner the statements in the following ways. (i) The central element of the addresser's profile is the conventionalized intention, and its fulfillment is trivially satisfied if the given process of pattern-matching is successful. (ii) There are pattern-matching problems which can be solved through the accommodation of the missing premises. For example, if the addresser says *Open the door*, the addressee can draw the conclusion that it is closed. Moreover, (iii) any axiom of the profile sent implicitly to the addressee as part of the whole intensional profile of the message can be interpreted as follows: *the addresser intends the addressee to learn that…* In the case of a declarative sentence, for instance, the addressee can freely construct conclusions such as *the addresser intends me to learn that (s)he knows this fact* ((s)he is proud of the fact that (s)he is so clever and well informed) or *(s)he intends me to learn that (s)he knows that I do not know this* ((s)he recalls that (s)he is precisely aware of my unfamiliarity with certain issues). Finally, symmetrical to the former case, (iv) some kind of reduction of the intention is also possible for the addressee. For example, in the sentence *Peti szerint Mari otthon van* 'In Peti's opinion, Mari is at home,' *szerint* 'in someone's opinion' should primarily be interpreted as *the addresser has said that Peti believes something*, but if the listener thinks that (s)he can calmly rely on Peti (and on the addresser), (s)he can directly accept the piece of information that *Mari is at home* as a

fact. This way the addressee has undertaken the risk of accepting a false piece of information but if one wants to exploit information coming from others, (s)he should trust in others.

5 Summary, loose ends, and some remarks on implementation

The paper is devoted to provide some theoretical background to a program which operates on a continuously changing world-model with possible speakers and listeners speaking about the world and each other, and in which the agents' information states are also continuously changing depending on changes of the model of the outer world, including their messages sent to each other, and each other's information states [18].

In Section 1, we presented the representationalist theoretical framework we use, and argued that its implementation promises to provide a lot of information on discourse mechanisms and linguistically relevant phenomena. We claim that, on the basis of truth-evaluation and pragmatically relevant felicity conditions, a program based on ℜeALIS is able to simulate the process of discourse interpretation in a more sophisticated way then earlier systems based on DRT-style theories.

The crucial innovation of ℜeALIS is that the traditional pattern-matching based truth-evaluating mechanisms of formal semantics can and should be generalized in a way that each sentence sent as a message should undergo a multiple mechanism of simultaneously extensional and intensional interpretation (Sections 2 and 3).

Section 4 is devoted to a detailed demonstration of the generalized method of interpretation of the five basic sentence types and a few sentence types constructed from the basic ones by augmented them with certain discourse markers.

Finally, some words on future versions of our program. In the current, first, phase of the project [18], the eventualities considered in the model of the outer world are only such simple states (or at least handled as such) which are just holding true in a given moment of the "game", or do not hold. For instance, Peter is married in certain moments while he is not married in other moments. Bea is working, or not working. In certain moments, Peter adores Bea, while in other moments, he does not adore her. Snowing has also two phases: it is snowing in a certain moment, or not. In the ℜeALIS-framework, there exists a much more sophisticated description of the general internal event structure of eventualities [17]: the temporal axis is cut into five intervals. In the first and the last intervals, the eventuality does not hold true. The second one is the interval of preparation. *He is about to travel home* – by a sentence like this one can refer to this interval, when the agent is packing his suitcases, buying tickets, saying good-byes. The third interval is the cumulative phase. An illustration: *he is just traveling home*. The fourth one is the result state: e.g., *he has traveled home*, that is, *he is at home*. One might think at first glance that it is only a question of quantity whether eventualities are considered only with two phases, or they are regarded as having five states in the above sense, of which in two phases the given eventuality does not hold true while in three other intervals it is being carried out, which manifests itself in different, truth-evaluable, forms from time to time. However, that is not the case. Let us consider, for instance, the type of Hungarian imperative sentence fine-tuned by the discourse marker *csak* 'only'. Thus, *Utazz haza!* 'Travel home' is considered to be augmented as follows:

Utazz csak haza! While this fine-tuned variant, which, depending on intonation, can express encouragement or intimidation, is preferably be performed in the preparatory phase of the eventuality (when the agent seems to be about to travel home), the basic imperative type can readily be performed in the earlier, first, interval, when the agent is likely not to think about traveling home.

All in all, we claim that the implementation of our formal system can be regarded as taking the first steps towards simultaneously representing the outer world in its double role [19]: as (possible) world-models which our words should be aligned to in the course of a post-Montagovian style of dynamic interpretation [3] [7] [9] (words→world) and as world-states (including states of human minds) that the acts of/in/by saying our words result in (world→words).

Acknowledgements

The present scientific contribution is dedicated to the 650th anniversary of the foundation of the University of Pécs, Hungary. We are grateful for the financial support of the University of Pécs (EFOP343). We would also like to thank the two anonymous reviewers for their constructive suggestions and comments.

References

1. Austin, J. L.: How to Do Things with Words. Clarendon Press, Oxford (1975/1962)
2. Alberti G.: ℜeALIS: Interpretálók a világban, világok az interpretálóban. Akadémiai Kiadó, Budapest (2011)
3. Kamp, H., Genabith, J., Reyle, U.: Discourse Representation Theory. In: Gabbay D., Guenthner F. (eds.) Handbook of Philosophical Logic 15. pp. 125–394 Springer-Verlag, Berlin (2011)
4. Benz, A.: Chains and the Common Ground. In: Poesio M., Traum D. (eds.) GötaLog 2000 — Gothenburg Papers in Computational Linguistics 00–5 pp. 181–184 (2000)
5. Fagin, R., Halpern J.Y., Moses Y. and Vardi M.: Reasoning about Knowledge. MIT Press, Cambridge (1995)
6. Premack, D., Woodruff, G.: "Does the chimpanzee have a theory of mind?" Behavioral and Brain Sciences, special issue: Cognition and Consciousness in Nonhuman Species. Cambridge Journals. 1 (4): pp. 515–526 (1978)
7. Asher, N., Lascarides A.: Logics of Conversation. Cambridge Univ. Press, Cambridge (2003)
8. Schlangen, D., Lascarides, A. and Copestake, A.: Resolving Underspecification using Discourse Information. In: Kühnlein, P., Rieser H., Zeevat H. (eds.) Proceedings of the 5th International Workshop on Formal Semantics and Pragmatics of Dialogue. pp. 79–93 Bielefeld (2001)
9. Asher, N., Lascarides A.: Imperatives in dialogue. In: Kühnlein P., Rieser H. and Zeevat H. (eds.) The Semantics and Pragmatics of Dialogue for the New Millenium, New Series 114. pp. 1–24 Benjamins, Amsterdam (2003)

10. Oishi, E.: Discursive functions of evidentials and epistemic modals. In: Cantarini, S. et al. (eds.) Certainty-uncertainty – and the attitudinal space in between. Studies in Language Companion Series 165. pp. 239–262 Benjamins, Amsterdam (2014)
11. Alberti, G., Kleiber, J., Kárpáti, E.: Reális (ℜeALIS) kép a másik elméjéről. In: Márton M. et al. (szerk.) Más elmék. pp. 237–258 L'Harmattan Kiadó, Budapest (2017)
12. Alberti, G., Vadász, N., Kleiber, J.: Ideal and Deviant Interlocutors in a Formal Interpretation System. In: Zuczkowski, A. et al. (eds.) Communicating Certainty and Uncertainty in Medical, Supportive and Scientific Contexts. Dialogue Studies 25. pp. 59–78 Benjamins, Amsterdam (2014)
13. Kleiber, J., Alberti, G.: 2017. Compositional Analysis of Interrogative Imperatives in Hungarian. Manuscript (Univ. of Pécs, Dept. of Linguistics) based on a talk at the Linguistics Beyond and Within Conference. Lublin, 20–21 October (2016)
14. Grice, P.: Logic and conversation. In: Cole P., Morgan, J. L. (eds.) Syntax and Semantics 3. pp. 41–58 Academic Press, New York (1975)
15. Szeteli, A.: *Szerint(em).* A kifejezés funkcióinak vizsgálata formális interpretációs rendszerben. In Böhm G., Czeferner D., Fedeles T. (szerk.) Szemelvények 4. Válogatás a PTE BTK XXXIII. OTDK-n I–III. helyezést elért hallgatóinak pályaműveiből. pp. 88–107. PTE BTK TDKT, Pécs (2017)
16. Kas, B.: Az óhajtó mondatok kategóriája. Nyelvtudományi közlemények 102. pp. 136–174 (2005)
17. Farkas, J., Ohnmacht, M. Aspect and Eventuality Structure in a Representational Dynamic Semantics. In Alberti, G., Farkas, J., Kleiber, J. (eds.) Vonzásban és változásban pp. 353–379 Doctoral School of Linguistics at Univ. of Pécs, Hungary (2012)
18. Nőthig, L., Szeteli A.: Nagyfelbontású pragmaszemantikai igazságértékelés egy játékprogramban. Current volume.
19. Searle, J. R.: Expression and meaning – Studies in the Theory of Speech Acts. Cambridge Univ. Press, Cambridge (1979)

	Exclamative (basic type)	*Szerintem*-Declarative	Declarative (basic type)	Interrogative (basic type)	*Ugye*-Interrogative	*Vajon*-Interrogative
BELIEF	$\langle B,\gamma_{B\tau-},AR,\tau^{-},0-\rangle$	$\langle E,M,AR,\tau,0\rangle$				$\langle B,M,AR,\tau^{+},0\rangle$
	$\langle B,\gamma_{B\tau},AR,\tau,+\rangle$	$\langle B,\iota,AR,\tau,+\rangle$ $M \geq \iota \geq s$ $\iota = \iota' \cdot \iota''$	$\langle B,M,AR,\tau,+\rangle$	$\langle B,M,AR,\tau,0\rangle$	$\langle B,M,AR,\tau,0\rangle$	$\langle B,M,AR,\tau,0\rangle$
	$\gamma_{B\tau}=1$ $(\gamma_{B\tau}-\gamma_{B\tau-})+\lvert\gamma_{D}\rvert$	$\lambda\hat{}\langle B,\iota''',ae,\tau,+-\rangle$			$\langle BD,\gamma_{u},AR,\tau,+\rangle$ $\gamma_{u} \geq gr$	
		$\lambda\hat{}\langle B,M,ae,\tau,0\rangle$ $\hat{}\langle B,\iota,AR,\tau,+\rangle$	λ $\hat{}\langle B,M,ae,\tau,0\rangle$	$\lambda\hat{}\langle B,\gamma'',ae,\tau,+\rangle$ $\hat{}\langle B,M,AR,\tau,0\rangle$	$\lambda\hat{}\langle B,\gamma'',ae,\tau,+\rangle$ $\hat{}\langle B,M,AR,\tau,0\rangle$	*$\lambda\hat{}\langle B,\gamma'',ae,\tau,+\rangle$* *$\hat{}\langle B,M,AR,\tau,0\rangle$*
DESIRE	$\langle D,\gamma_{D},AR,\tau,+-\rangle$	$\lambda\hat{}\lambda'$ $\hat{}\langle B,M,ae,\tau^{+},+-\rangle$ $\hat{}\langle B,\iota,AR,\tau,+\rangle$	$\lambda\hat{}\lambda'$ $\hat{}\langle B,M,ae,\tau^{+},+-\rangle$	$\lambda\hat{}\lambda'$ $\hat{}\langle B,M,AR,\tau^{+},+-\rangle$	$\lambda\hat{}\lambda'\hat{}\langle B,M,AR,\tau^{+},+-\rangle$	$\lambda\hat{}\lambda'$ $\hat{}\langle B,M,AR,\tau^{+},+-\rangle$ $r^{*}=AR$
AUTHORITY		$\langle A,M,AR,\tau,+\rangle$ $\hat{}\langle B,M,ae,\tau^{+},+\rangle$ $\hat{}\langle B,\iota,AR,\tau,+\rangle$	$\langle A,M,AR,\tau,+\rangle$ $\hat{}\langle B,M,ae,\tau^{+},+\rangle$	$\lambda\hat{}\langle A,\gamma',ae,\tau,+\rangle$ $\hat{}\langle B,M,AR,\tau^{+},+-\rangle$ stricter axiom of $\lambda\hat{}\langle B,M,ae,\tau,+-\rangle$	$\lambda\hat{}\langle A,\gamma',ae,\tau,+\rangle$ $\hat{}\langle B,M,AR,\tau^{+},+-\rangle$ stricter axiom of $\lambda\hat{}\langle B,M,ae,\tau,+-\rangle$	*$\lambda\hat{}\langle A,\gamma',ae,\tau,0\rangle$* *$\hat{}\langle B,M,AR,\tau^{+},+-\rangle$* *consequence of* *$\lambda\hat{}\langle B,M,ae,\tau,0\rangle$*
INTENTION		λ'' $\hat{}\langle B,M,ae,\tau^{+},+\rangle$ $\hat{}\langle B,\iota,AR,\tau,+\rangle$	λ'' $\hat{}\langle B,M,ae,\tau^{+},+\rangle$	$\lambda''\hat{}\lambda'''$ $\hat{}\langle B,M,AR,\tau^{+},+-\rangle$	$\lambda''\hat{}\lambda'''$ $\hat{}\langle B,M,AR,\tau^{+},+-\rangle$	*$\lambda''\hat{}\langle B,M,ae,\tau^{+},+\rangle$* *$\hat{}\lambda'\hat{}\langle B,M,AR,\tau^{+},+-\rangle$* *pref.: $r^{*}=\{AR\}$*

Table 1. Intensional profiles in ℜeALIS (BEL: +/0): $\lambda=\langle B,\gamma,AR,\tau,+\rangle$; $\lambda'=\langle D,\gamma_{r^{*}},r^{*},\tau,+\rangle$, where $\Sigma\gamma_{x} \geq 1$ ($x \in r^{*}$) and preferred: $r^{*}=\{AR,ae\}$; $\lambda''=\langle I,M,AR,\tau,+\rangle$; $\lambda'''=\langle I,M,ae,\tau^{+},+\rangle$

	Interrogative (basic type)	Conditional (-nÁl) **re**quest*ion*	Imperative (basic type)	*Nyugodtan*-Imperative	Optative (basic type)
BELIEF	⟨B,M,AR,τ,0⟩	⟨B,M,AR,τ,–⟩; *⟨B,M,AR,τ,0⟩^⟨D,M,ae,τ,+–⟩*	⟨B,M,AR,τ,–⟩	⟨B,M,AR,τ,–⟩	⟨B,M,AR,τ,–⟩
		λ^⟨B,M,ae,τ,–⟩	λ^⟨B,M,ae,τ,–⟩	λ^⟨B,M,ae,τ,–⟩	*λ^⟨B,M,ae,τ,–⟩*
	λ^⟨B,γ'',ae,τ,+⟩ ^⟨B,M,AR,τ,0⟩	*λ^⟨B,γ'',ae,τ,+⟩^ ⟨B,M,AR,^τ,0⟩^⟨D,M,ae,τ,+–⟩*			
DESIRE	λ^λ' ^⟨B,M,AR,τ^+,+–⟩	**⟨D,M,AR,τ,+⟩**; ⟨D,M,AR,τ,+⟩^⟨B,M,AR,τ^+,+⟩ ^⟨D,γ_{ae},ae,τ,+–⟩	λ^λ'	λ^λ'	⟨D,M,AR,τ,+⟩
AUTHORITY	λ^⟨A,γ',ae,τ,+⟩ ^⟨B,M,AR,τ^+,+–⟩ stricter axiom of λ^⟨B,M,ae,τ,+–⟩	trivially satisfied: λ^⟨B,M,ae,τ,+–⟩ ^⟨D,M,ae,τ,+–⟩; λ^⟨A,γ',ae,τ,+⟩	λ^⟨A,γ',ae,τ,+⟩; λ^⟨A,M,AR,τ,+⟩ ^⟨I,M,ae,τ^+,+⟩	λ^⟨A,γ',ae,τ,0⟩	⟨A,M,AR,τ,0⟩
INTENTION	λ''^λ''' ^⟨B,M,AR,τ^+,+–⟩	*λ''^λ'''^⟨B,M,AR,τ^+,+⟩ ^⟨D,M,ae,τ,+–⟩;* **λ''^λ''' λ''^λ'''^⟨B,M,AR,t^+,+⟩ ^⟨D,M,ae,τ^+,+–⟩**	λ''^λ'''	λ'' ^⟨A,M,AR,τ^+,+⟩	*λ'' ^⟨B,M,ae,τ^+,+⟩ ^⟨D,M,AR,τ,+⟩*
NOTE	Ag=ae	Ag=ae	Ag=ae		Ag=ae
	Ag≠AR	Ag≠AR	Ag≠AR		Ag≠AR

Table 2. Intensional profiles in ℜeALIS (BEL: –/0): λ=⟨B,γ,AR,τ,+⟩; λ'=⟨D,γ_{r^*},r^*,τ,+⟩, where $\Sigma\gamma_x \geq 1$ ($x \in r^*$) and preferred: r^*={AR,ae}; λ''=⟨I,M,AR,τ,+⟩; λ'''=⟨I,M,ae,τ^+,+⟩

Egy orosz nyelvű korpusz (NarRu) narratívaelemzése saját fejlesztésű szentiment- és emóciószótárakkal

Nyíri Zsófi[1], Szabó Martina Katalin[1,2,3], Ilyés Virág[1]

[1]Precognox Informatikai Kft.
[2] Szegedi Tudományegyetem, Szláv Intézet, Orosz Filológiai Tanszék
[3] MTA TK „Lendület" RECENS Kutatócsoport
nyirizsofi93@gmail.com
{mszabo, vilyes}@precognox.com

Kivonat: A tanulmány egy orosz nyelvű szentiment- és emóciólexikon, valamint egy szépirodalmi korpusz létrehozásának tapasztalatairól és felhasználásának első eredményeiről számol be. A korpusz és az elemzési eszközök létrehozásának elsődleges célja az volt, hogy a szövegek emóció- és szentimentelemzésével feltárjuk azok narratívastruktúráit. A tanulmányban bemutatjuk a korpusz felépítését és létrehozásának a legfontosabb szempontjait, valamint a szótárak készítésének eszközeit, módszereit, alapelveit és adatait. Ezt követően ismertetjük a korpuszszövegek emóció- és szentimentelemzési eredményeit. Áttekintést adunk arról a megoldásról, amellyel a kapott elemzési eredmények alapján a szövegek narratívaszerkezeteit feltártuk, és a leggyakoribb narratívatípusokat definiáltuk. Végezetül bemutatjuk azt az interaktív dashboardot, amelynek segítségével a korpuszt a létrehozott eszközökkel a szentimentek, az emóciók, valamint a narratívaszerkezetek szempontjából vizsgálhatóvá tettük.

1 Bevezetés

A dolgozatban egy olyan kutatásról számolunk be, amelynek célja különböző orosz nyelvű szépirodalmi szövegek narratívaszerkezetének a feltárása szentiment- és emóciószótárak segítségével. A kutatáshoz a vizsgálati anyagot és az elemzési eszközöket magunk hoztunk létre [1].

A munka egyik fontos célja az volt, hogy olyan szótárakat és korpuszt fejlesszünk, amelyek segítségül szolgálhatnak az orosz nyelvre irányuló emóció-, szentiment-, valamint narratívaelemzési feladatainkhoz. A kutatás alapjául vett tanulmányban Reagan és szerzőtársai [2] szentiment- és emócióelemzéssel vizsgálták angol nyelvű szövegek narratíváit, ami alapján azt feltételeztük, hogy a korpusz szövegeinek szentiment- és emócióelemzési eredményeivel rá tudunk mutatni az egyes szövegek narratívastruktúráira, illetve sikerül erre a szövegsajátságra vonatkozóan típusokat meghatároznunk.

Annak céljából, hogy különböző korszakokban keletkezett irodalmi művek struktúráit összevethessük egymással, a vizsgálati korpuszban három alkorpuszt különítettünk el. A kutatás első fázisában elvégzett szótáras elemzéssel megkaptuk a

korpuszban reprezentált korszakokra jellemző szisztematikus mintázatokat az emóció- és a szentimenttartalmak vonatkozásában. Ezt követően, az eredmények alapján feltártuk a szövegek narratívastruktúráit, a különböző szerkezettípusok korpuszbeli kvantitatív adatait, és definiáltuk a fő narratívatípusokat.

A dolgozatban ismertetjük a vizsgálati korpusz létrehozásának a menetét, módszereit, alapelveit, a vizsgálati eszközök adatait, valamint szólunk a fejlesztés során felmerülő elméleti és gyakorlati dilemmákról is. Ezt követően ismertetjük az elvégzett elemzési megoldásokat, majd a feldolgozással kapott eredményeket mind a szentiment- és emóciótartalmakra, mind a narratívastruktúrákra vonatkozóan, és bemutatjuk az azokat prezentáló interaktív dashboardot. Végül szólunk a munka további tervezett lépéseiről is.

2 A korpusz és a szótárak létrehozása

A munka első szakaszában létrehoztuk a vizsgálathoz szükséges szépirodalmi korpuszt, valamint az orosz nyelvű emóció- és szentimentszótárakat.

Az emóciószótárunk nyolc emóciókategóriát foglal magában, míg a szentimentlexikon egy, a negatív és egy, a pozitív értékelő tartalommal rendelkező szentimentkifejezéseket tartalmazó listát. Az emóciókifejezéseket egy orosz nyelvű monográfia mellékletére támaszkodva gyűjtöttük össze [3]. A szentimentszótárak esetében a SentiRus [4] orosz nyelvű, valószínűségi mutatókat tartalmazó lexikon listáját dolgoztuk fel, rendszereztük, valamint egészítettük ki az emóciószótárunk anyagával [1].

A korpusz három alkorpuszra oszlik, a szövegek keletkezési ideje alapján. A korpusz szövegeit automatikus módszerekkel gyűjtöttük le egy szabadon hozzáférhető műveket tartalmazó honlapról[1].

2.1 A vizsgálati korpusz létrehozása

A szövegkorpusz létrehozásának első lépése a források feltérképezése, majd a legmegfelelőbb forrás[2] kiválasztása volt. Választásunk egyik fontos szempontja az volt, hogy az oldal lehetővé tegye a szövegek txt formátumú automatikus gyűjtését.

A szövegek kiválasztásánál a korpusz előre meghatározott paramétereit követtük. Fő kiindulási pontnak az 1917-es orosz forradalmat választottuk. Hipotézisünk szerint ugyanis az eseményre a szépirodalomban is reflektáltak, és ezek a hatások a szövegek szintjén is kimutathatóak lesznek. Figyelembe véve a forradalom, valamint az annak hatására született művek közötti időrést, a szövegek határának az 1895-1955 közé eső időszakot választottuk. Ezt a későbbiekben további három periódusra osztottuk: az 1. korszakot 1895-1910, a 2. korszakot 1910-1920, a 3. korszakot pedig 1920-1955 közé helyeztük. Az 1. periódus a forradalom előtti, a 3. periódus a forradalom utáni

[1] http://www.royallib.com
[2] http://www.royallib.com

időszakot reprezentálja, míg a 2. periódus a forradalom hatására vagy az azt közvetlen megelőző években létrehozott szövegeket foglalja magába.

A létrehozott adatbázis minden szöveg esetében tartalmazza annak címét, szerzőjének nevét, valamint a szöveg keletkezésének a dátumát.

A munka gyűjtési fázisában fő célunk egy minél nagyobb adatbázis létrehozása volt. A szövegek kiválasztásánál tehát a munka e szakaszában nem követtünk semmilyen kritériumot a keletkezés időpontján kívül.

A szövegeket automatikus módszerrel gyűjtöttük, txt formátumban és UTF-8 karakterkódolással, hogy azok előfeldolgozó lépések nélkül feleljenek meg az elemző eszközök bemeneti kritériumainak. A mondatra és a szavakra bontáshoz az NLTK csomag alapeszközeit [5], a POS taggeléshez és a morfológiai elemzéshez pedig egy kimondottan orosz nyelvű szövegekre fejlesztett programot, a Pymorphy2-t használtuk [6]. A feldolgozás eredményeként megkaptuk a szövegek szavainak kisbetűs tokenjét, lemmáját és a POS-taget. Az output fájl tartalmazott néhány szisztematikus hibát (pl. az idézőjelben szereplő vagy a latin szavak felismerésének a hiánya következtében), de azokat manuális korrigáltuk.

A létrehozott korpusz adatait az 1. táblázat mutatja be.

	1. Periódus (1895-1910)	2. Periódus (1910-1920)	3. Periódus (1920-1955)	Összesen
Szöveg	46	48	82	176
Szerző	10	15	22	51
Átlagos szövegenkénti szószám	35 516	18 666	34 956	30 660
Összes szószám	1 633 741	895 978	2 866 470	5 396 189

táblázat.1. A korpusz adatai.

Ahogyan a táblázatban látjuk, az adatbázisban összesen 51 szerző 176 szövege található, melyek közül legtöbb, összesen 22 szerző 82 szövege a 3. periódusban keletkezett, hiszen az öleli fel a leghosszabb időintervallumot. Ugyanakkor érdemes felhívni a figyelmet arra, hogy a 2. periódus átlagos szószáma jóval alacsonyabb, mint az egész korpuszra vonatkozó átlagos érték, míg az 1. és a 3. periódusban ugyanez valamivel magasabb. Ebből arra lehet következtetni, hogy az 1917-es forradalmat közvetlen megelőző és követő időszakban összességében rövidebb művek születtek.

2.2 Az emóció- és a szentimentszótár létrehozása

Az emóciószótárak létrehozásában egy orosz nyelvű monográfia mellékletére támaszkodtunk [3]. A melléklet összesen 8500 lexémát tartalmaz, amelyek a szerző alapján az orosz nyelvben az emóciók kifejezésére képes elemek.

A monográfia [3] 37 emotív jelentést, azon belül számos további alkategóriát, valamint sajátos szemantikai tartalmi típusokat különít el, és ebben a komplex rendszerben dolgozza fel a lexikai anyagot. A kiinduló emotív jelentéseken belül

felsorolt lexikai elemeket 7 további kategóriába sorolja (érzelmi állapot, érzelmi állapot kialakulása, érzelmi hatás, érzelmi viszonyulás, érzelem külső kifejeződése, érzelmi jellemzés és érzelmi kvalitás). A jelentések típusait a következő alkategóriák szerint különíti el: fő,- alapjelentés, nominatív és metaforikus, valamint képzett jelentés.

Saját szótáraink kialakításánál az első és talán legfontosabb dilemmának az emóciókategóriák meghatározása bizonyult, hiszen erről a kérdésről nincs egységes vélekedés sem a hazai, sem a külföldi irodalomban [7,8]. Tekintettel arra, hogy a kutatási tervek között szerepelt az angol és magyar nyelvű szövegekkel történő kontrasztív vizsgálat, az ezeken a nyelveken már meglévő és számunkra elérhető szótárak [8,9,10,11,12] felépítésére támaszkodtunk, és egy nyolc érzelemkategóriából álló séma mellett tettük le a voksunkat. A nyolc emóciókategória az öröm, a düh, a bánat, a félelem, a feszültség, a vonzalom, az undor és a meglepődés volt [9].

A kiinduló anyagként szolgáló lista [2] feldolgozását automatikus és manuális módon végeztük el [1]. A mellékletet digitalizáltuk, majd az anyagot manuálisan javítottuk és az elemeket listákba rendeztük. A szótár kialakításakor a szófaj kritériumát tágan értelmeztük, így minden, emóciót kifejező, vagy annak meglétére utaló elemet felvettünk a szótárba.

A kialakított szótár adatait és a kategóriák közötti megoszlási arányait az 1. ábra mutatja be.

ábra. 1. Az emóciószótárak adatai és megoszlási arányai.

A szentimentszótár létrehozásához első lépésként kiválasztottunk egy 5000 elemet tartalmazó, orosz nyelvű lexikont [4]. A lexikont a szerzők a következőképpen készítették el: A munka első fázisában gyűjtöttek egy kisméretű, pozitív és negatív szentimentértékű elemekből álló listát, amelyeket egy korpuszra illesztettek, majd lekérdezték a szótár elemeinek környezeteit. Az eredmények alapján a korpusz minden szavához, amely szerepelt valamelyik környezetben, egy nullától egyig terjedő valószínűségi értéket rendeltek. A szám azt mutatja, hogy mekkora valószínűséggel

hordoz pozitív vagy negatív értéket az adott elem. A létrehozott szótár tehát azon a feltevésen alapul, hogy minél gyakrabban szerepel egy adott elem pozitív vagy negatív környezetben, annál nagyobb valószínűséggel lesz maga is azonos értékű (a módszer problémáiról részletesebben l. [4]).

Részben e valószínűségi mutatókra támaszkodva határoztuk meg, hogy a listából mely elemek kerüljenek be a szótárainkba. A 0.250-nél alacsonyabb valószínűségi mutatóval rendelkező elemeket eleve nem vettük figyelembe, a többi elemet pedig egyenként megvizsgáltuk, és, amennyiben elfogadtuk azt, besoroltuk a pozitív vagy a negatív lista valamelyikébe. Ezután a szentimentszótárakat az emóciószótár anyagával tovább bővítettük, ugyancsak egyenként vizsgálva a lehetséges szentimentkifejezéseket. E bővítési megoldás indokolható, hiszen, bár az emócióelemzés az érzelmek, a szentimentelemzés pedig a nyelvi értékelések azonosítását célozza, a két jelenség szorosan összefügg egymással, számos emóciókifejezés egyben értékelést is hordoz [9]. Az elkészített szentimentlexikon adatait a következő táblázat mutatja be.

Szentimentkifejezés	Elemszám
Pozitív	1982
Negatív	2675
Összes	4657

táblázat.2. A szentimentlexikon adatai.

3 Az emóció- és a szentimentelemzés eredményei

Ahhoz, hogy az emóció- és a szentimentelemzéssel olyan görbéket kaphassunk, amelyek időrendben mutatják a szövegek cselekményét, minden szöveget 100 részre osztottunk, majd a szótáras elemzéssel minden részre kiszámoltuk az emóció- és a szentimentértékeket. Az eredményeket egy interaktív dashboardon vizualizáltuk, ami lehetővé teszi az adathalmazon belüli keresést korszakonként, művenként, részenként, és megmutatja a különböző emóció- és szentimentértékeket a kiválasztott fentebbi feltételek szerint. Az alábbi ábra a negatív szentimentek előfordulásának vizualizációját mutatja be az interaktív dashboardon.

ábra. 2. Az interaktív dashboard egy részlete.

A bal oldalon lévő szentimentgörbék a korpuszon belüli szentimentértékeket mutatják. A 3 egymás alatt vízszintes görbe a 3 alkorpuszt reprezentálja, felülről alulra a következő sorrendben: 1. Periódus, 2. Periódus, 3. Periódus. Balról jobbra haladva látjuk a korábban 100 részre osztott szövegek szentimentértékeit, követve azok cselekményét. A dashboard ezen részlete így az összes szöveg szentimentértékét tartalmazza, jelen ábrán nem szűkítettük a keresést konkrét művekre. A 2. ábra jobb oldalán a harminc leggyakoribb szentimentkifejezés látható, valamint azok korszakonkénti előfordulása. Emellett feltüntettük azt is, hogy korszakonként összesen hány negatív értékítéletet kifejező szentimentkifejezést sikerült beazonosítani (1. Periódus: 1622, 2. Periódus: 1490, 3. Periódus: 1818).

Az emóció- és szentimentértékeket, az egész korpuszban és korszakonkénti megoszlásban is, a 3. táblázatban tüntettük fel Mivel a korszakokat reprezentáló alkorpuszok szószáma eltér, az emóció- és szentimentkifejezések előfordulási számát a teljes korpusz és az alkorpuszok szószámához arányítva is bemutatjuk, valamint feltüntetjük a korpuszhoz viszonyított százalékos megoszlást is. A szentimentkifejezéseknél a százalékértékeket a pozitív és a negatív szentimentkifejezésekre együtt értjük.

	1. Periódus (1895-1910)	2. Periódus (1910-1920)	3. Periódus (1920-1955)	Összesen
Összes szószám	1 633 741	895 978	2 866 470	5 396 189
Szentimentkifejezések száma	2894: 0,18%	2612: 0,29%	3228: 0,11%	8834: 0,16%
Ebből pozitív	1272	1222	1410	3904
Ebből negatív	1622	1490	1818	4930
Emóciókifejezések száma	2091: 0,13%	1843: 0,21%	2350: 0,09%	6284: 0,12%

táblázat.3. Emóció- és szentimentértékek a NarRu korpuszban.

Ahogyan azt az előfordulási arányok mutatják, az emóció- és a szentimentkifejezések legnagyobb százalékban a 2. periódusban fordulnak elő, annak ellenére, hogy ez öleli fel a legrövidebb időintervallumot. Míg az 1. periódus hasonló adatai alapvetően egyeznek az egész korpuszra számított értékekkel, a legnagyobb eltérés a 2. és a 3. periódus között mutatkozik. Utóbbiban ugyanis csupán kis százalékban voltak beazonosíthatók az emóció- és a szentimentkifejezések.

Az emócióelemzés során azt az eredményt kaptuk, hogy az emóciókifejezések emóciókategóriánkénti megoszlása összhangban áll az emóciószótár elemeinek emóciókategóriánkénti százalékos megoszlásával, ezért azt külön táblázatban nem reprezentáljuk.

4 A szövegek narratívaelemzése

A narratívaelemzés alapját a 100 részre feldarabolt szövegek egyes részenkénti szentimentértéke adta.

A kapott értékekből főkomponens elemzéssel nyertük ki a korpuszban domináló narratívagörbéket. Az elemzést elvégeztük a szentimentértékek összességére és átlagára is. Az irodalomban talán jellemzőbb a szentimentértékek összegének alkalmazása [2], miután az átlagok számítása révén veszítünk adataink változatosságából.

A szentimentértékek összességét nézve 6, valamint 12 főkomponenst nyertünk ki az adatokból. Előbbi esetében az adataink kezdeti varianciájának kicsivel több mint 50, utóbbi esetében majdnem 60%-át át őriztük meg. A szentimentértékek átlagából 6 főkomponenst különítettünk el, és adataink kezdeti varianciájának 80%-át őriztük meg.

A kapott narratívagörbéket mozgóátlagot számolva kisimítottuk. Ehhez az adott időpont előtt és utána lévő 5 szövegrész szentimentértékének átlagát vettük figyelembe, kisebb súllyal számolva azokat az értékeket, melyek távolabb estek az adott szövegrésztől.

Pearson-féle korrelációt számoltunk, hogy kiderüljön, mely művek simított narratívagörbéi állnak legközelebb az egyes narratívatípusokhoz.

A kapott narratívatípusokat összevetettük Reagan és szerzőtársai [2] angol nyelvű szövegek elemzésével kapott narratívagörbéivel, amelyek a szentimentértékek összességére számolt narratívákkal mutatták a nagyobb egyezést. A szentimentértékek összességére végzett elemzésből az alábbiakban az áttekinthetőség kedvéért a 6 főkomponenst megkülönböztető eredményeket prezentáljuk:

ábra. 3. A szentimentek összesített értékeiből definiált 6 főkomponens.

A 6 főkomponens közül az 1. egy pozitívan stagnáló görbe, a legegyszerűbb cselekményszerkezetnek tekinthető. Az 5. görbe egy 'u' alakú verzió, ahol a történet közepén lévő mélypont után visszatérünk a kezdeti állapothoz, míg a 6. főkomponens egy emelkedő narratívagörbe. A 2. és a 3. görbe mutatja a legbonyolultabb cselekményt. A 4. főkomponens az 5.-hez hasonlóan visszatér a kiinduló ponthoz, annyi különbséggel, hogy a történetben két, egy szolidabb és egy drasztikusabb mélypont mutatható ki, amelyek között egy enyhe pozitív felemelkedés látható.

Narratívagörbéiket Reagan és szerzőtársai [2] minden esetben két verzióra osztották az alapján, hogy az pozitív vagy negatív véggel záródik. A kutatás során ezt a felosztást nem alkalmaztuk, de amennyiben ebből a szempontból tekintjük saját eredményeinket, azt látjuk, hogy azok, az 5. főkomponenstől eltekintve, mindegyik főkomponens inkább pozitív értékkel záródik, bár a 2. és a 4. görbe a történet végén enyhe esést mutat. A következő ábra Valerij Brjuszov Tüzes angyal c. művének cselekményét mutatja be. A történet a 4. főkomponenshez kapcsolható 0.783123 korrelációs értékkel. A narratívagörbén a történet fordulópontjait és főbb eseményeit jelöltük be.

ábra. 3. V. Brjuszov Tüzes angyal c. műve.

A regény cselekménye a 4. főkomponenshez tartozik. A történet vége visszatér a kiinduló ponthoz, pozitív értékkel kezdődik és azzal is záródik. A pozitív kezdés a két kulcsfigura találkozása és kettejük Kölnben töltött ideje köré szerveződik, míg a pozitív vég a főszereplő hazatérésére és missziójára utal. A 4. főkomponens két mélypontja is magyarázható. A szolidabb mélypont egy párbajjal és annak negatív kimenetével, a drasztikusabb pedig a két szereplő elválásával. A mélypontok közötti enyhe pozitív felemelkedés a sérült párbaj utáni ápolásával köthető össze.

5 Összegzés

A tanulmányban egy olyan kutatásról számoltunk be, amelynek célja az volt, hogy különböző orosz nyelvű szépirodalmi szövegek narratívaszerkezetét szentiment- és emóciószótárak segítségével feltárjuk, valamint a legfőbb narratívaszerkezeteket definiáljuk. A kutatáshoz minden vizsgálati és elemzési eszközt magunk hoztunk létre [1], amelyek fontos adalékul szolgálhatnak a jövőbeli, orosz nyelvre irányuló tartalomelemzési feladatainkhoz is.

A kutatás első lépéseként létrehoztunk egy összesen 51 író 176 szövegét tartalmazó szépirodalmi korpuszt. Annak céljából, hogy a különböző korszakokban keletkezett művek struktúráját összevethessük egymással, az adatbázisban három alkorpuszt különítettünk el.

Az első feldolgozási szakaszban a korpuszt mondatokra és szavakra bontottuk, amihez az NLTK csomag alapeszközeit használtuk [5]. A szövegek POS taggeléséhez és a morfológiai elemzéséhez egy orosz nyelvű szövegekre fejlesztett programot, a Pymorphy2-t alkalmaztuk [6]. Az ily módon feldolgozott korpuszt ezután az emóció- és szentimentszótárunkkal, szótárillesztéses megoldással elemeztük. A szótáras elemzésekkel megkaptuk a korpuszban bemutatott korszakokra jellemző szisztematikus mintázatokat az emóciók és a szentimentek vonatkozásában.

A szótárakat ugyancsak e kutatás keretében, magunk hoztuk létre. Az emóciószótárunk, amelyben nyolc érzelemtípust különböztetünk meg, összesen 3291 elemet tartalmaz. A szentimentszótár egy negatív és egy pozitív értékelő nyelvi elemeket tartalmazó listából áll ls összesen 4657 szentimentkifejezést tartalmaz.

A kapott eredmények alapján feltártuk a szövegek narratívastruktúráit, valamint megállapítottuk, mely szövegek tartoznak a különböző szerkezettípusokhoz. Végül az eredményeket egy interaktív dashboard segítségével vizualizáltuk. A dashboard nyújtotta lehetőségeket szeretnénk még jobban kiaknázni, és azt további kutatásokra és elemzésekre is felhasználni.

Az elkészített korpusz, a vizsgálati eszközök, valamint az interaktív dashboard a Precognox Informatikai kft. TAS elnevezésű nyelvfeldolgozó rendszerének részét képezi (http://precognox.hu/tas-szoveganalitikai-rendszer). A dashboardot hamarosan kutatási célra szabadon hozzáférhetővé tesszük a honlapon.

Köszönetnyilvánítás

A kutatást a Precognox Informatikai Kft., az Emberi Erőforrások Minisztériuma Új Nemzeti Kiválóság Programja, valamint az Európai Kutatási Tanács (European Research Council) az Európai Unió Horizont 2020 kutatási és innovációs programjának keretében (ERC_CoG_2014_648693 sz. szerződésben) támogatta.

Bibliográfia

1. Nyíri, Zs.: Egy orosz nyelvű emóció- és szentimentszótár, valamint egy orosz nyelvű szépirodalmi korpusz létrehozásának tapasztalatai. In: II. Szláv Filológiai Konferencia, Budapest (2017) Megjelenés előtt
2. Reagan, A. J., Mitchell, L., Kiley, D., McLaughlin Danforth, C., Dodds, P. S.: The emotional arcs of stories are dominated by six basic shapes. In: EPJ Data Science (2016) Elérhető: https://arxiv.org/pdf/1606.07772v2.pdf
3. Бабенко, Л. Г.: Лексические средства обозначения эмоций в русском языке. Издательство Уральского университета, Екатерибург (1989)
4. Chetviorkin, I. I., Loukachevitch, N. V.: Extraction of Russian Sentiment Lexicon for Product Meta-Domain. Proc. of COLING 2012: Technical Papers (2012) 593– 610
5. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit, O'Reilly Media (2009) Elérhető: http://www.nltk.org/book_1ed/
6. Морфологический анализатор Pymorphy2 (2015) Elérhető: https://pymorphy2.readthedocs.io/en/latest/
7. Liu, B.: Sentiment Analysis and Opinion Mining. Draft (2012) Elérhető: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
8. Szabó, M. K., Morvay, G.: Emócióelemzés magyar nyelvű szövegeken. In: Gecső, T., Sárdi, Cs. (eds.): Nyelv, Kultúra, Társadalom, Tinta Könyvkiadó, Budapest (2015) 286-293
9. Szabó, M. K., Vincze, V., Morvay, G.: Magyar nyelvű szövegek emócióelemzésének elméleti és nyelvtechnológiai problémái. In: Reményi, A. Á., Sárdi, Cs., Tóth, Zs. (eds.):

Távlatok a mai magyar alkalmazott nyelvészetben, Tinta Könyvkiadó, Budapest (2016) 282-293
10. Szabó, M. K.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. In: Gecső T., Sárdi, Cs. (eds): Nyelv, Kultúra, Társadalom, Tinta Könyvkiadó, Budapest (2015) 278-286
11. Wilson, T., Wiebe J., Hoffmann P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005 (2005) Elérhető:
12. Strappava, C., Mihalcea. R.: Learning to identify emotions in text. Elhangzott: UMM CSci Senior Seminar Conference, University of Minnesota, Morris. Morris, Amerikai Egyesült Államot (2008) Elérhető:
http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.217.62

Forrás

http://www.royallib.com
http://precognox.hu/tas-szoveganalitikai-rendszer

Kacsa vagy nem kacsa? Magyar nyelvű álhírek automatikus azonosítása

Vincze Veronika[1,2]

[1]Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék
Szeged, Árpád tér 2.
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Kivonat Az interneten olvasott hírek nem mindig származnak megbízható forrásokból, illetve nem mindig valósághűek. Cikkünkben gépi tanulási technikák segítségével különítjük el a magyar nyelvű valós híreket az álhírektől. A szakirodalom áttekintése után bemutatjuk a vizsgálati anyagot, majd ismertetjük módszereinket. Az eredmények alapján elmondhatjuk, hogy gazdag nyelvi jellemzőtéren alapuló megoldásunk 80%-os eredményeséggel képes a megfelelő csoportba sorolni az ál- és valódi híreket: elsősorban a morfológiai jellemzők bizonyultak leghatásosabbnak a gépi tanulási folyamatban.
Kulcsszavak: hírek, álhírek, osztályozás, információkinyerés

1. Bevezetés

Manapság a világhálón számos dokumentum születik, melyek száma napról napra gyarapszik. E dokumentumok bárki számára hozzáférhetőek, amennyiben szerzőjük azt szabadon elérhetővé teszi. Különösen a közösségi médiában gyakori jelenség a tartalmak megosztása a többi felhasználóval. Ezzel azonban együtt jár az is, hogy sokszor nem lehetünk biztosak abban, hogy kitől származik az adott bejegyzés, hiszen az interneten sokszor csak felhasználónévvel vagy álnévvel szerepelnek az emberek, emiatt pedig az olvasott információ hitelességét sem vehetjük készpénznek, ha nem megbízható forrásból származik.

Az internetkorszak beköszöntével együtt járó változások egyike, hogy sokan például a világhálóról tájékozódnak nyomtatott sajtó olvasása vagy egy tévéhíradó megtekintése helyett (vagy mellett). Ennek megfelelően a nagyobb napilapok sok esetben saját online portált üzemeltetnek, melyre elektronikus változatban feltöltik a nyomtatott sajtóban megjelent cikkeiket, emellett pedig léteznek online hírportálok is, melyek hírei nem jelennek meg nyomtatásban, kizárólag online olvashatók. A megbízható hírportálok mellett azonban léteznek valótlan tartalmakat közlő oldalak is. Ezek egyik fajtájához tartoznak az úgynevezett hoax oldalak, melyek álhíreket, kétes valóságtartalmú állításokat, reklámokat tartalmaznak. Az oldal szerzőinek célja az olvasók megtévesztése, különféle hamis

információk terjesztése, illetve pénzszerzési tevékenység. A hoax oldalak mellett azonban léteznek olyan álhíreket közlő oldalak is, melyek elsődleges célja az olvasók szórakoztatása, és e tevékenységük nyilvánvaló az olvasó számára is. Cikkünkben ez utóbbi típusba tartozó álhírek automatikus azonosítását tűzzük ki célul: gépi tanulási technikák segítségével különítjük el a valós híreket az álhírektől. A szakirodalom áttekintése után bemutatjuk a vizsgálati anyagot, majd ismertetjük módszereinket. Az eredmények alapján elmondhatjuk, hogy gazdag nyelvi jellemzőtéren alapuló megoldásunk eredményesen képes a megfelelő csoportba sorolni az ál- és valódi híreket.

2. Kapcsolódó irodalom

A világban számos munka foglalkozik az álhírek automatikus felismerésével. Például megtévesztő termékismertetők automatikus azonosítását célozza [1] a szentimentelemzés szemszögéből, illetve pletykák felismerésére is született több tanulmány [2,3]. Angol nyelvű kísérletek szerint elsősorban szintaktikai [4] és nyelvi jellemzők [5] segítik az álhírek felismerését, [6] pedig részben felügyelt módszerekről számol be.

Az angol nyelvre több adatbázis is született az utóbbi időben, melyekben álhírek és valódi hírek egyaránt előfordulnak. [7] összesen 221 állítást tartalmaz annak valódisági értékével együtt, míg a Liar korpusz csaknem 13 000 rövid, politikai tartalmú állításból áll [8]. Az Emergent korpuszt 300 állítás és 2500, az állításokhoz kapcsolódó hír alkotja, valódisági értékükkel [9]. Tudomásunk szerint magyar nyelvre azonban ez az első munka, mely álhírek automatikus felismerését célozza.

Szövegek	**Szövegek száma**	**Mondatszám**	**Tokenszám**
Valódi hírek	40	1001	15941
Álhírek	40	719	10376

1. táblázat. A korpusz adatai.

3. A korpusz

Vizsgálatunkat egy 80 hírből álló korpuszon végeztük. A korpusz szövegeit az internetről töltöttük le. A valódi hírek több országos és megyei hírportálról származnak (pl. www.index.hu, www.origo.hu, www.delmagyar.hu). Az álhíreket két forrásból gyűjtöttük össze. Egyrészt valós híreket közlő napilapok április elsejei, az olvasók megtréfálására szolgáló álhíreit válogattuk ki, másrészt a www.hircsarda.hu oldalról töltöttünk le cikkeket. A Hircsárdán megjelenő hírek leginkább az aktuálpolitikai és közéleti események parodizálására épülnek, céljuk vállaltan az olvasóközönség szórakoztatása. A legtöbb hír 2017 szeptemberéből

és októberéből származik, ez alól természetesen kivételt képeznek az április elsejei hírek. A híreket véletlenszerűen válogattuk, arra azonban ügyeltünk, hogy egy adott eseményről csak egy hír szerepeljen a korpuszban.

A korpusz alapvető adatait az 1. táblázat mutatja be.

Az alábbiakban közlünk egy-egy példát valódi és álhírekre is, a szövegek eredeti helyesírását változatlanul hagyva.

Moszkvában készült robbantani az Iszlám Állam[1]

Az Iszlám Állam terrorszervezet moszkvai pokolgépes merényletekre készülő alvósejtjét számolta fel az orosz Szövetségi Biztonsági Szolgálat, az FSZB – közölte hétfőn a testület társadalmi kapcsolatokkal foglalkozó osztálya.

A hattagú csoport forgalmas helyeken és tömegközlekedési járműveken tervezett robbantásokat. Az FSZB vasárnapi elfogási akciója során két nagy hatóerejű, házi készítésű pokolgépet, robbanószerkezetek előállításához szükséges anyagokat, két pisztolyt, két kézigránátot és kommunikációs eszközöket foglalt le a terroristáktól. Az FSZB szerint a külföldről irányított, észak-kaukázusi származású tagokból álló csoport egy világméretű kalifátus létrehozásának ideológiáját követte.

Zuckerberg megvette Magyarországot[2]

„Fasza kis startupnak tűnik" – kommentálta országunkat Budapest romkocsmáit látogatva a Facebook vezére. Magyarország ezentúl Facebook appként működik tovább.

Lapunknak sikerült kiderítenie a váratlanul Budapesten feltűnő Mark Zuckerberg látogatásának célját. Ez pedig nem más, mint Magyarország akvizíciója, amivel a Facebook portfolióját szeretné szélesíteni a világ egyik legbefolyásosabb embere. „Ha fényképezni akarok ott az Instagram, a térképi infók végett pedig meglesz a Waze. Most pedig veszek egy országot. LOL" – indokolta a vásárlást a Facebook vezetője.

Elemzők szerint a vételre a kutyája beszélhette rá, aki szintén magyar származású és puli. Úgy tudjuk, hogy Magyarország Magyarország ezentúl Facebook appként működik tovább. A jelenlegi vezetést valószínűleg lecserélik és hipszter dizájnerekkel, programozókkal, egy kávéautomatával és Széles Gáborral töltik fel az igaztatótanácsot. „Yeah, Gábor is lots of fun. Különösen tetszik a perpetuum mobile alapú golfpálya koncepciója. Zseniális. Nekem jó ideje ő a háttérképem" – mondta el Zuckerberg munkatársunknak, aki kutyaeledelnek álcázva jutott a közelébe.

Hogy pontosan mi lehet a célja az országunkkal az egyelőre nem ismert, de valószínűleg hamarosan mobilra is optimalizálják, megjelennek rajta reklámok és bárki örökbe fogadhat rajta kidobott kismacskákat.

[1] http://www.origo.hu/nagyvilag/20171002-moszkvaban-akart-robbantani-az-iszlam-allam.html

[2] http://hircsarda.hu/2013/05/29/zuckerberg_megvette_magyarorszagot

4. Gépi tanulási kísérletek

Kísérleteinkben szerettük volna megvizsgálni, mennyire hatékonyan különíthetők el automatikusan az álhírek a valósi hírektől. Ehhez első lépésként a magyarlanc [10] elemzővel automatikusan előelemeztük a szövegeket, hozzájutva így azok szótövesített változatához, illetve morfológiai és szintaktikai elemzéséhez. A nyelvi elemzésre építve számos nyelvészeti alapú jellemzőt használtunk fel a kísérletben, figyelembe véve, hogy az álhírek és valódi hírek megkülönböztetése elsősorban szemantikai-pragmatikai kérdés, hiszen egy jól megírt álhír nem feltétlenül különbözik morfológiai és szintaktikai felépítésében a valódi hírektől. Ennek alátámasztására statisztikai szignifikanciavizsgálatokat is végzünk az egyes jellemzőkre nézve (l. 5. rész).

A felhasznált jellemzőket az alábbiakban részletezzük.

- **Statisztikai jellemzők**:
  - mondatok száma,
  - szavak száma,
  - mondatok átlagos hossza.
- **Morfológiai jellemzők**:
  - főnevek, igék, melléknevek, határozószavak, tulajdonnevek és ismeretlen szavak száma és aránya a szószámhoz képest,
  - múlt és jelen idejű igék száma és aránya az igék számához képest,
  - feltételes és felszólító módú igék száma és aránya az igék számához képest,
  - E/1. és T/1. igék száma és aránya az igék számához képest,
  - felsőfokú és középfokú melléknevek száma és aránya a melléknevek számához képest.
- **Szintaktikai jellemzők**:
  - alanyok, tárgyak, jelzők, határozók, alárendelő és mellérendelő mondatok száma és aránya.
- **Szemantikai jellemzők**:
  - pozitív és negatív jelentésű szavak száma és aránya a szószámhoz képest (a listákat [11] alapján készítettük el),
  - bizonytalanságra utaló szavak száma és aránya a szószámhoz képest ([12] alapján felállított osztályokba sorolva), érzelmekre utaló szavak száma és aránya a szószámhoz képest ([13] alapján felállított osztályokba sorolva).

  A szemantikai jellemzők vizsgálatakor egyszerű listaillesztéses módszert használtunk: amennyiben az egyes szavak lemmája megegyezett bármelyik listaelemmel, akkor találatként számoltuk.
- **Pragmatikai jellemzők**:
  - beszédaktust jelentő igék száma és aránya,
  - szó szerinti idézetek és nyilatkozatok száma és aránya.

  Utóbbiakat a szövegben előforduló idézőjelek és mondatkezdő gondolatjelek számával mértük.

A feladatot bináris osztályozásként kezeltük, modellként pedig döntési fákat [14] alkalmaztunk tízszeres keresztvalidációval. Alapmódszerként a többségi osztályozást választottuk, mely 50%-os pontosságot ért el.

5. Eredmények

Gépi tanuló rendszerünk 80%-os pontosságot ért el, azaz 80 dokumentumból 64-et sorolt be a megfelelő osztályba. A részletesebb eredmények a 2. táblázatban láthatók. Mint látható, az alapmódszernél jóval hatékonyabb eredményeket sikerült elérni. A számokból az is kiderül, hogy a rendszer jobban teljesít a valódi hírek felismerésében, míg az álhírek közül többet minősített valódi hírnek, 40 esetből 9-szer tévedve.

Szövegek	Pontosság	Fedés	F-mérték
Valódi hírek	0,786	0,825	0,805
Álhírek	0,816	0,775	0,795
Összesen	0,801	0,800	0,800

2. táblázat. Eredmények.

Megvizsgáltuk az egyes jellemzőcsoportok hatékonyságát is porlasztásos módszerrel. A 3. táblázatban látottak szerint a leghatékonyabbnak a morfológiai jellemzők bizonyultak, azonban a szemantikai és szintaktikai jellemzők is pozitívan járulnak hozzá az eredményekhez.

						Különbségek			
Jellemzők	**Szöveg**	**P**	**R**	**F**	**Acc**	**P**	**R**	**F**	**Acc**
összes	álhírek	0,816	0,75	0,795					
	valódi hírek	0,786	0,825	0,805					
	összes	0,801	0,8	0,8	80				
morfológia nélkül	álhírek	0,488	0,5	0,494		-0,328	-0,25	-0,301	
	valódi hírek	0,487	0,475	0,481		-0,299	-0,35	-0,324	
	összes	0,487	0,488	0,487	48,75	-0,314	-0,312	-0,313	-31,25
szintaxis nélkül	álhírek	0,78	0,8	0,79		-0,036	0,05	-0,005	
	valódi hírek	0,795	0,775	0,785		0,009	-0,05	-0,02	
	összes	0,788	0,788	0,788	78,75	-0,013	-0,012	-0,012	-1,25
szemantika nélkül	álhírek	0,769	0,75	0,759		-0,047	0	-0,036	
	valódi hírek	0,756	0,775	0,765		-0,03	-0,05	-0,04	
	összes	0,763	0,763	0,763	76,25	-0,038	-0,037	-0,037	-3,75

3. táblázat. Porlasztásos eredmények. P: pontosság (precision), R: fedés (recall), F: F-mérték (F-measure), Acc: pontosság (accuracy).

Egyes jellemzők statisztikailag szignifikáns különbséget mutatnak a két csoport között, melyeket a 4. táblázatban mutatunk be részletesebben. Eszerint tehát a morfológiai jellemzők közül a határozószavak száma, illetve az ismeretlen szavak, múlt és jelen idejű igék és T/1. igék aránya, a szintaktikai jellemzők közül az alanyok és mellérendelések száma, a tárgyak aránya és a jelzők és alárendelések száma és aránya, míg a szemantikai jellemzők közül az episztemikus

és weasel típusú bizonytalanságot jelölő szavak száma (és aránya), illetve a bánatot kifejező szavak száma térnek el szignifikánsan a valódi hírek és az álhírek csoportjai között.

Jellemző	**p-érték**
Morfológiai jellemzők	
Határozószavak száma	0,0146
Ismeretlen szavak aránya	0,0001
Múlt idejű igék aránya	0,0215
Jelen idejű igék aránya	0,0184
T/1. igék aránya	0,0078
Szintaktikai jellemzők	
Alanyok száma	0,036
Tárgyak aránya	0,0362
Jelzők és alárendelések száma	0,0344
Jelzők és alárendelések aránya	0,0071
Mellérendelések száma	0,0459
Szemantikai jellemzők	
Episztemikus bizonytalanságot jelző kifejezések száma	0,0051
Episztemikus bizonytalanságot jelző kifejezések aránya	0,0183
Weasel típusú bizonytalanságot jelző kifejezések száma	0,0398
Bánatot kifejező szavak száma	0,027

4. táblázat. Szignifikáns különbségeket mutató jellemzők.

6. Az eredmények megvitatása

Az eredmények vizsgálata alapján elmondhatjuk, hogy gépi tanulási kísérleteink 80%-os pontossággal különítik el egymástól a valódi és álhíreket. Az is egyértelműen látszik, hogy elsődlegesen a morfológiai jellemzők bizonyultak hasznosnak, ám a szintaktikai és szemantikai jellemzők hozzáadott értéke is kiderül a 3. táblázatból. A statisztikai és pragmatikai jellemzők azonban nem jelentettek számszerűsíthető többletet a mérések során.

A fentiek alapján azt is kijelenthetjük, hogy elvárásunk – miszerint elsősorban szemantikai-pragmatikai különbségeket találunk az álhírek és valódi hírek között – nem teljesen igazolódott be. Egyrészt a morfológiai jellemzők bizonyultak leghatásosabbnak a gépi tanulási folyamatban, másrészt a statisztikailag szignifikáns jellemzők sem mutatják a szemantikai-pragmatikai jellemzők kiemelkedő szerepét. Eredményeinkből úgy tűnik, hogy a korpuszunkban található álhírek szókészlete és mondatszerkesztése is erősen eltér a valódi hírekétől: azoknál több határozószót, több jelen idejű igét és több T/1. igét használ. A mellé- és alárendelések száma azonban alacsonyabb, ami arra utal, hogy kevesebb összetett mondat található az álhírekben. Ugyanakkor a jelenre való utalások, a többes szám

első személyű igék gyakori használata is a közösségi tudatot erősíti az álhírekben, ami akár manipulatív stratégia is lehet [15], az olvasó meggyőzését célozva. Továbbá, több bizonytalanságra utaló kifejezést találhatunk a valódi hírekben, azaz az álhírekben inkább tényszerű állítások jelennek meg, ezzel is „tényként" beállítva a közölt tartalmakat. Érzelmek terén pedig különösen a bánatra és a haragra jellemző szavak használata gyakoribb a valódi hírekben, ellenben a szeretet kifejeződése az álhírekben gyakoribb, azaz a pozitív szemlélet talán jobban az álhírek sajátja.

A rendszer által tévesen osztályozott hírekre mutatunk néhány példát az alábbiakban.

Tévesnek ítélt valódi hírek:

- *Időkapszulát ástak el a Spitzbergákon*
- *A trópusi erdők több szenet bocsátanak ki, mint az amerikai gépjárműpark*
- *Észak-Korea nehézbombázókat röptet az óceán felett*

Valódinak címkézett álhírek:

- *Itt a kézműves iPhone*
- *Közmunkásokat képeznek át macskává*
- *Tarr Béla rendezi a Trónok harca 7. évadját*

A hibásan kategorizált cikkeket áttekintve megállapíthatjuk, hogy elsősorban a dokumentumok hosszúsága játszott szerepet a téves osztályozásban. A rövidebb valódi híreket gyakran álhírnek, míg a hosszabb álhíreket valódinak minősítette a rendszer. Mindemellett minél több bizonytalanságra utaló nyelvi elem volt megtalálható az álhírben, annál valószínűbb, hogy a rendszer valódi hírnek tekinti.

7. Összegzés

Ebben a munkában magyar nyelvű hírek és álhírek automatikus elkülönítéséről számoltunk be. Eredményeink alátámasztják, hogy gazdag – morfológiai, szintaktikai, szemantikai és pragmatikai – jellemzőtérre építve sikeresen lehetséges felismerni az álhíreket, különösen a morfológiai jellemzők játszanak fontos szerepet a folyamatban.

Jelen cikkben olyan álhírekkel dolgoztunk, melyek elsődleges célja az olvasó szórakoztatása volt, nem pedig megtévesztése. A jövőben azonban hoax típusú hírekre is szeretnénk kiterjeszteni gépi tanuló megközelítésünket, azaz az olvasót szándékosan megtévesztő, illetve manipuláló információk automatikus felismerése is célunk.

Köszönetnyilvánítás

A tanulmány az Emberi Erőforrások Minisztériuma UNKP-17-4 kódszámú Új Nemzeti Kiválóság Programjának támogatásával készült.

Hivatkozások

1. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, Association for Computational Linguistics (2011) 309–319
2. Ma, J., Gao, W., Wong, K.F.: Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, Association for Computational Linguistics (2017) 708–717
3. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M.: Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. IJCAI'16, AAAI Press (2016) 3818–3824
4. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Jeju Island, Korea, Association for Computational Linguistics (2012) 171–175
5. Pérez-Rosas, V., Mihalcea, R.: Experiments in open domain deception detection. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Association for Computational Linguistics (2015) 1120–1125
6. Hai, Z., Zhao, P., Cheng, P., Yang, P., Li, X.L., Li, G.: Deceptive review spam detection via exploiting task relatedness and unlabeled data. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Association for Computational Linguistics (2016) 1817–1826
7. Vlachos, A., Riedel, S.: Fact checking: Task definition and dataset construction. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, MD, USA, Association for Computational Linguistics (2014) 18–22
8. Wang, W.Y.: "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada, Association for Computational Linguistics (2017) 422–426
9. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, Association for Computational Linguistics (2016) 1163–1168
10. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763–771
11. Szabó, M.K.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. In: Segédkönyvek a nyelvészet tanulmányozásához 177. Tinta, Budapest (2015) 278–285
12. Vincze, V.: Uncertainty detection in Hungarian texts. In: Proceedings of Coling 2014. (2014)
13. Szabó, M.K., Vincze, V., Morvay, G.: Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In: Távlatok a mai magyar alkalmazott nyelvészetben. Tinta, Budapest (2016)
14. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)

15. Bártházi, E.: Manipuláció, valamint manipulációra alkalmas nyelvhasználati eszközök a sajtóreklámokban. Magyar Nyelv **104**(4) (2008) 443–463

A Rákosi-éra pártjegyzőkönyveinek feldolgozása, elemzése és vizualizációja szövegalapú kapcsolatháló-elemzési módszerekkel

Gulyás Attila[1], Szabó Martina Katalin[1,2,3], Ifj. Boros István[1], Havadi Gergő[1]

[1]MTA TK „Lendület" RECENS Kutatócsoport
[2]Szegedi Tudományegyetem, Szláv Intézet, Orosz Filológiai Tanszék
[3]Precognox Informatikai Kft.
{gulyas.attila, szabo.martina, boros.istvan, havadi.gergo}@tk.mta.hu

Kivonat: A jelen dolgozatban a hatalmi hálózatok szöveg alapú feltárását célzó projektünk egy részletét, a Rákosi-jegyzőkönyvek feldolgozását és az ezek alapján elkészített mintavizualizációt mutatunk be. A Rákosi-éra hatalmi hálózatának fejlődése és működése jól rekonstruálható a történelmi dokumentumoknak − levéltári anyagoknak, jegyzőkönyveknek − és interjúknak köszönhetően. Kutatásunkat a hatalmi hálózat mögött rejlő informális kapcsolatok feltárását célozza meg. Vizsgálati anyagaink között változatos forrásból származó, nagy mennyiségű irat szerepel, amelyek döntő többségükben különböző pártbizottságok ülését megörökítő jegyzőkönyvek. Ezen dokumentumok feldolgozása komoly kihívást jelent, mivel a számuk igen nagy, a digitalizált anyagok minősége pedig a közel használhatatlantól a jól olvashatóig terjed. Dolgozatunkban bemutatjuk a dokumentumok feldolgozásának az alapelveit, eszközeit és módszertanát, illetve ismertetjük, hogyan állíthatóak elő kapcsolathálók ezekből a dokumentumokból, majd tárgyaljuk a szövegalapú kapcsolatháló-elemzés módszerét, és betekintést nyújtunk az ehhez tartozó vizualizációs technikákba. Végezetül egy mintaelemzéssel szemléltetjük az alkalmazott módszert, megmutatjuk a szövegalapú kapcsolatháló-elemzésben rejlő lehetőségeket. Kutatásunk egyik fő célja az, hogy a szövegeken szentiment- és topikelemzést hajtsunk végre a jövőben, megerősítve, vagy éppen megcáfolva korábbi eredményeinket.

1 Bevezetés

1.1 A kutatás történelmi háttere

A második világháborút követően hazánkban kialakult politikai helyzetből 1949-re az MDP (Magyar Dolgozók Pártja) került ki győztesen.[1] Az ezt követő években egy szűk

[1] A Magyar Dolgozók Pártja 1948 júniusában jött létre, miután a Magyar Kommunista Párt „egyesült" a szociáldemokrata párttal (SzDP). Ezt hivatalosan fúziónak, vagyis a két munkáspárt egyesülésének nevezték, valójában azonban a meggyengített és megtizedelt szociáldemokrata párt maradványát olvasztotta magába az MKP (Magyar Kommunista Párt).

hatalmi elit a párthierarchia mellett a kapcsolatrendszerén keresztül biztosította uralmát. Kapcsolatrendszer alatt ugyanakkor nem csupán a politikai életben kiépített kapcsolatokról beszélünk, hanem az azon kívül, az informális életben zajló kapcsolatokról is. A későbbiekben is számos példát láthatunk arra, hogy pártfunkcionáriusok éppen informális kapcsolataikkal erősítették meg a politikai kapcsolataikat, vagy éppen az informális kapcsolataikból kovácsoltak politikai tőkét [1]–[3].

Kutatásunkban a politikai kooperáció során létrejött kapcsolatokat vetjük össze a párthierarchia által diktált struktúrával a kapcsolatháló-elemzés eszközeit segítségül hívva. A kutatás felfogható egyfajta történelmi elitkutatásnak, amelyet társadalomtudományos eszközökkel (a hálózatkutatás módszerével) végzünk, a történeti források prozopográfiai feldolgozásán és vizsgálatán keresztül (néhány ígéretes hazai kísérlet erre, például [4], [5].

Az elit Andorka alapján a társadalmi hierarchia csúcsán elhelyezkedő kis létszámú – az uralkodó osztálynál kisebb – csoport [6]. A politikai elit ennek egyik szegmense, típusa; a sztálinista típusú kommunista diktatúrákban az erőforrások (gazdaság, kultúra, társadalom és kapcsolódó tőkék) felett kizárólag regnáló csoport.[2]

A Rákosi éra hatalmi / politikai elitje szervezetileg viszonyalag könnyem körülhatárolható és definiálható: a Párt (MDP) politikai vezető testületeinek a Titkárság, a Politikai Bizottság, valamint az 1953-ig létező Szervező Bizottság) tagjaiból és póttagjaiból állt össze. Ezen belül is elsősorban (informálisan is) azok a csúcspolitikusok tartoztak bele megkérdőjelezhetetlenül és maradandóan a hatalmi elitbe, akik kiemelt pozíciókkal, személyes hatással és kapcsolatrendszerrel rendelkeztek (és ennél fogva információkkal bírtak) az államigazgatásban, a tömegszervezetek (pl. a szakszervezetek), kultúrális élet irányításában (a *Szabad Nép* főszerkesztője) avagy az erőszak szervezetek (ÁVH, Honvédség, Rendőrség) ellenőrzésében. Őket nevezhetjük a „a pártvezetés legfelső körének".

Jellemző a Rákosi korszak hatalmi elitjére, hogy épp oly könnyen eshetett ki közülük valaki, ahogy bekerült (lásd. erről a folyamatos éberség hisztéria és ellenségkép fenntartását igazoló, Sztálini (szovjet) mintájú koncepciós perek sorát: Rajk, Marosán, Kádár, Kállai etc.).

Ebből a szempontból a későbbi Kádár alatti elit réteg lényegében a Rákosi alatti elitre épült, annak is gyorsan mozgosított, új másod- és harmad vonalára (Apró, Gáspár, Hegedüs, Komócsin, Münnich, Piros, Szalai, Vég vagy Czinege). Különösen érdekes lehet tehát, hogy ebből az elitből vezető pártelit mellett az említettek milyen, a párthierarchiában értelmezhető, de nem abból fakadó kapcsolatrendszert építhettek ki.

A MDP elnöke ugyan Szakasits Árpád lett, ám tényleges vezetője Rákosi Mátyás főtitkár volt. A párt létszáma meghaladta az egymillió főt. 1956 júliusában Rákosit leváltották az MDP éléről, utódjává a tőle politikájában nem sokban különböző Gerő Ernőt választották, akit október 25-én Kádár János váltott fel.

[2] Abban a közelmúlt történelmének kutatásával foglalkozó szakemberek szinte egybehangzóan egyetértenek, hogy nem érték- vagy presztízsalapon, hanem pozíciók vizsgálata mentén van leginkább értelme a pártállami elitbesorolásoknak illetve a hatalmi elit vizsgálatának (Rácz 2014).

1.2 Röviden a szövegalapú kapcsolatháló-elemzésről

A hálózatelemzés jelentőségére mutatnak rá azok az újabb, és egyre szaporodó kutatási eredmények, amelyek a legkülönfélébb hálózatos szerveződésekben szabályszerű mintázatok létrejöttére mutatnak rá. Barabási a következőképpen fogalmaz: „Hálózatok mindenhol vannak. Az agy axonok által összekötött idegsejtek hálózata, maguk a sejtek pedig biokémiai reakciók által összekötött molekulák hálózatai. A társadalmak szintén hálózatok […]. A hálózatok átjárják a technológiát is: az internet, az elektromos hálózatok, valamint a szállítási rendszerek csupán néhány példa erre.” [7]

A 1990-es évek végén jelennek meg az első olyan tudományos megállapítások, amelyek az eddig egymástól függetlennek tűnő hálózatos rendszereknek (pl. úthálózat, világháló, emberi kapcsolatrendszerek stb.) közös tulajdonságaira irányítják a figyelmet, továbbá amellett érvelnek, hogy ezek a tulajdonságok matematikailag leírhatók és elemezhetők [8]–[11].

A szövegalapú kapcsolatháló-elemzés alatt a szövegeket kapcsolathálóként értelmező és a társadalmi kapcsolathálók elemzési módszereit (SNA) használó paradigmát értjük [12]. Az SNA módszere a matematikából jól ismert gráfelméleti gyökerek mellett a fizika gyakorlatiasabb megközelítése alapján fejlődött ki [13].

A kapcsolatháló-elmélet és ezen belül a társadalmi kapcsolathálók elemzése hazánkban a köztudatba Barabási nyomán robbant be [14]. A szociológián belül a módszer leginkább a kapcsolati tőke vizsgálatában uralkodó megközelítés [15]–[17]. A kapcsolatháló-elemzés népszerűsége egyrészt abból fakad, hogy a módszer „társadalmi beágyazottsággal” rendelkezik [18].

Azonban nem csak társadalmi, (személy alapú) kapcsolatok, hanem tulajdonképpen „minden ábrázolható hálózatként” [12]. „Még a nyelv is, amit gondolataink közvetítésre használunk, önmagában véve nem más, mint szintaktikai kapcsolatokkal összekötött szavak hálózata” [7]. Ennek megfelelően a szöveget reprezentáló kapcsolatháló csomópontjai nem egyes személyek, vagy személyek csoportja, hanem az egyes szövegrészek, legtöbbször szavak. Gyakori ugyanakkor az úgy nevezett biparias kapcsolatháló is, melyekben általában a detektált kulcsszavak vagy témák mellé személyeket társítanak [19]. Így lehetséges például tudományos publikációk elemzése alapján arra következtetni, hogy mely kutatóhoz mely területek társíthatók. Ennek segítségével következtethetünk arra, hogy ki, hogyan és milyen terület felé tolja azt a közösséget, mely az eszköz vagy adott téma fejlesztésével, művelésével foglalkozik. A szavak közötti reláció, uniplex [20] feltétele a bizonyos szöveg tartományon belüli együttes előfordulás.

A szövegalapú kapcsolatháló-elemzés tehát nemcsak egy újfajta reprezentációját jelentheti az szövegeknek, hanem a mögöttes tartalmak megismerésében is segítséget nyújthat.

2 A kutatás kérdései és forrásai

Kutatásunk célkitűzése a Rákosi-korszak (1948-1956) alatti politikai hatalmi elit látens hálózatának a felrajzolása, többfajta történeti forrás feldolgozása és elemzése segítségével. Vizsgáljuk a látens és manifeszt hierarchia, valamint a kapcsolatok dinamikájának történeti hálózatát.

Számos korábbi dolgozat feldolgozta már a pártelit látens kapcsolatainak alakulását: kiváló példákat olvashatunk az informális szférában kötött kapcsolatokról a közös vadászatokon keresztül [1], vagy említhetjük Aczél György erősen kapcsolathálókon alapuló, meglehetősen sikeres politikai működését [3]. Ennek nyomán kutatásunk egyik fő hipotézise az, hogy a politikai, vagy akár a politikai tevékenységeket kiegészítő elfoglaltságok terén együtt működő párttagok között olyan látens kapcsolat is létrejöhetett, amely a párthierarchiával párhuzamosan formálta a politikai szférában való tevékenységüket.

A forrásadataink az állampárt (MDP) politikai vezető testületeinek 1948-1956 között keletkezett üléseinek szerkesztett jegyzőkönyvei (Politikai Bizottság, Titkárság, illetve Szervező Bizottság). A hálózatban szereplő személyek pontos azonosításához egyéb történelmi dokumentumokat (biográfiákat, káderlapokat, és életrajzi adatbázisokat) használtunk fel, melyekből kiolvasható az adott személy politikai funkciója, és a pártéletben betöltött szerepe mellett számos további adat (iskolák, lakhelyek, különböző politikailag fontos eseményeken való részvétel stb.), amelyből ugyancsak informális kapcsolataikra következtethetünk. A rendelkezésünkre álló források tehát változó állapotban lévő gépelt, illetve gyakran kézi jegyzeteket is tartalmazó dokumentumok, amelyek feldolgozása és elemzése komoly kihívást jelent (részletesebben l. lentebb).

3 A feldolgozás és a kutatás módszertana

A korpusz létrehozását és feldolgozását az alábbi ábrának megfelelő legfontosabb lépésekben végeztük el:

1. ábra. A kapcsolathálózat létrehozásának a lépései

A jelen fejezetben e folyamatot részleteiben ismertetjük.

3.1 A korpuszszövegek feldolgozásának első lépései

Ahhoz, hogy a későbbi feldolgozási lépések anyagát létrehozhassuk, mindenekelőtt a szövegek digitalizálására volt szükség.

A digitalizálást OCR-eszközzel végeztük el. Ezt követően egy jelentős szövegkorrekciós fázist kellett beiktatnunk, elsősorban a további munkálatok szempontjából kardinális nyelvi elemek, a tulajdonnevek formai problémái miatt. A problémáknak több oka volt. Egyrészt, a szövegekben előforduló tulajdonnevek többsége az átalakítás során sérült, azaz különböző karakterhibák kerültek a szövegbe. Másrészt, a feldolgozott szövegekben találkozhatunk olyan írájelekkel is, amelyeket e történelmi dokumentumokban a nem történelmi szövegekhez képest sajátosan alkalmaztak, így például a per-jelet zárójel funkciójában, amelyet az OCR-eszköz automatikusan nem ismer fel, és ezért nem alakít is át zárójellé. Hasonló, a szöveg típusából fakadó egyedi jellemző az is, hogy bizonyos esetekben a tisztségviselők nevét úgy emelik ki, hogy a név minden egyes karaktere közé szóközt tesztnek („P e t r ó c z i”). Nyilvánvaló, hogy a nem történelmi szövegeken trénelt algoritmusok gyakorta nem az elvárásainknak megfelelően kezelik ezeket az egyedi megoldásokat. Végezetül, A nevekbe számos további okból kifolyólag is kerülhettek hibák a digitalizálás során. Így például adódhatott papír öregedéséből, az alkalmazott tinta gyenge minőségéből, vagy akár a digitálizálás során alkalmazott szkennelő eszköz nem kielégítő minőségéből is.

Az alábbi ábrán e típushibákra mutatunk néhány példát. A bal oldalon a forrás részletei, a jobb oldalon azok digitalizált verziói láthatóak.

2. ábra. (fentről lefelé) Rosszul digitalizált nevek; speciális név formátum; különleges konvenciók (a ‘/’ karakter használata zárójelként)

Az OCR eszközzel végzett feldolgozást követően tehát a hibásan beolvasott nevek javítása volt az első lépés. Ehhez annak céljából, hogy az emberi erőforrás igényét csökkentsük, egyedi szoftverrel támogatott módszert használtunk, amely a rendelkezésre álló névtér-adatbázis alapján megkísérelte beazonosítani a szövegben fellelhető lehetséges neveket. Egy adott karaktersort akkor tekintett a program egy valós név lehetséges elírt alternatívájának, ha a karaktersor Levenshtein-távolsága [21] kevesebb volt a teljes név hosszának 30%-ánál. Így leginkább az OCR algoritmus által elkövetett típushibákat tudtuk javítani, például: Ger6 – Gerő, Kádár − K6d6r. A további hibákat manuálisan, az eredeti jegyzőkönyvekkel összevetve végeztük el. Ez utóbbiak azok az esetek voltak, amelyeket a szoftver azonosított ugyan, de akkora volt a torzulás mértéke (esetleg töredezett a név stb.), hogy azokat nem tekinthettük típushibának.

3. ábra. Kép a szoftverből

Az ezt követő lépés a nevek beazonosítása, amely tulajdonképpen a szövegekben előforduló névelemeknek az általuk jelölt entitásokhoz történő kapcsolását jelenti.

A tulajdonnevek azonosítása nem triviális feladat, ugyanis az egyes tulajdonnevek különböző alakokban is megjelenhetnek. A munka egy rendelkezésre álló névlista alapján történik, szoftverrel támogatott, ugyanakkor túlnyomórészt humán munkával. Ez a félautomatikus megoldás a következőképpen zajlik: Az azonosítást végző feldolgozók számára a szoftver minden egyes, a szövegben megtalálható név esetén felajánlja a lehetséges alternatívákat, kiegészítő információkat szolgáltatva az adott személyekről életrajzi adatok formájában.

A nevek azonosítását követően a szoftver az összes nevet a névtér-adabázis azonosítóját felhasználva taggé alakítja – például "Rákosi Mátyás", vagy "Rákosi" helyére a "rakosi_matyas_8538" szót teszi a szövegben.

Azt a jelenséget, amikor ugyanarra az entitásra (személy, hely, szervezet stb.) különböző nyelvi jelölőkkel hivatkozhatunk, *koreferenciá*nak nevezzük [22]. A fentebb ismertetett munkánk eredményeképpen tulajdonképpen megtörténik a szöveg koreferenciaviszonyainak részbeni az azonosítása, azaz összekapcsoljuk azokat a nyelvi jelölőket, amelyek egyazon személyre referálnak [23], [24]. Azért hangsúlyozzuk, hogy az anonosítás csupán részbeni, mert az egyéb, nem tulajdonnévi alakú, de személyre referáló nyelvi elemeket (pl. névmások, E/3 ragozású igék stb.) nem vettük figyelembe, ugyanis arra nem volt szükségünk: ez a típus a jegyzőkönyvekben nem fordul elő, az egyes személyekre mindig tulajdonnévvel utalnak.

A tulajdonnevek azonosítása a további munkavégzés szempontájóbl rendkívül fontos lépés volt. Ez tette lehetővé ugyanis, hogy az együtt említettségeket megfelelően meg lehessen határozni, tehát azokat az eseteket is helyesen tudjuk kezelni, amikor teljes nevet, vagy pedig csak vezetéknevet említenek a jegyzőkönyvben.

A kapcsolatok elemzéséhez az első feladatunk a kapcsolathálók létrehozása volt, amelyekben az adott kapcsolatháló csomópontjait a személynevek, a köztük lévő

éleket pedig a közöttük detektálható valamilyen kapcsolat (kooperáció) megléte vagy hiánya adja.

Az együtt említést úgy definiáltuk, hogy a két egyedített jelölő egy adott bekezdésben 5 szó távolságon belül fordul elő. Fontos itt megjegyeznünk, hogy a digitalizálás miatt nem mindig lehetséges mondatok azonosítása, ezért muszáj a bekezdésre hagyatkoznunk, amelyek jól elkülönülnek. A munka jelenleg ennél fázisnál, azaz a nevek beazonosításánál és a kapcsolathálók létrehozásánál tart.

Mivel a jegyzőkönyvek egy jelentős részénél igen jó minőségű (80% feletti) az OCR minősége, lehetőségünk nyílhat egy más típusú és egyben részletesebb, szövegalapú kapcsolatháló-elemzés elvégzésére is.

3.2 Módszertan - a szövegstruktúra elemzése

A szövegalapú kapcsolatháló-elemzés egy nagyon fontos eredménye a szövegstruktúra olyan módon történő felrajzolása, amely rámutat a szövegben rejlő témákra (ez nem összetévesztendő a topikelemzéssel), és az ehhez tartozó szövegrészekre. Itt ezt a módszert mutatjuk be egy egyszerű példán keresztül.

A megoldás bemutatásához egy kis méretű tesztkorpuszt hoztunk létre hat darab véletlenszerűen kiválasztott Titkársági jegyzőkönyvből, melyeket igyekeztünk a lehető legteljesebb mértékben helyreállítani. Az így kapott anyag tehát a szövegfelismerésből származó hibákat nem tartalmazó, értelmes magyar nyelvű szöveg. Mivel a jegyzőkönyvek az elemzés szempontjából fontos egyedi struktúrával rendelkeztek, azt az eredeti dokumentumoknak megfelelő módon megőriztük. A kapcsolatháló-elemzést megelőzően a korpuszt a feldolgozáshoz szükséges tidy-text jellegű formátumra hoztuk, hogy a feldolgozó szoftverek számára elemezhető legyen.

Az itt bemutatott modell strukturális modell, mivel a szöveg teljes (a stopszavak nélküli) szókészlete (wordcount) megtalálható benne, és a fő célkitűzése a nagymennyiségű szöveg tartalmak feldolgozása a szöveg strukturájának vizualizálásával. A kapcsolathálóként vizualizált és kezelt szövegrészletek legyakoribb formája a szemantikus hálózatok (semantic networks).

A szemantikus hálózatok felépítése során rendszerint szótövezést (stemming, lemmatization) és N-grammokat alkalmaznak. Mivel a szövegalapú kapcsolatháló-elemzés a korpusz összes tartalmas szavának eredeti alakjára kiváncsi, nem alkalmazza a fenti eljárásokat, melynek pozitívumai és hátrányai egyaránt vannak. A nem információ hordozó szavak, másképpen a funkciószók (pl. a kötőszavak), valamint a jelen kutatásban vizsgált dokumentumokban, azok műfajából adódóan gyakran előforduló szavak (pl. *jegyzőkönyv*) szűrésére stoplistát alkalmaztunk, amelyet magunk állítottunk össze manuális módon a szövegek kézi elemzése alapján. A szavak közti kapcsolatot saját munkánk esetében is az együttes előfordulás adta. Bár az együttes előfordulás többfajta szövegegységen belül is értelmezhető, így lehet dokumentum, paragrafus, mondat, vagy egy bizonyos, előre meghatározott Δx szónyi távolság. A Δx szónyi távolság az adott szó szövegbeli pozíciójától mindkét irányban számított távolságot jelenti. Magunk ez utóbbi megoldást alkalmaztuk.

Jelen elemzésünkben csakis egy lexémából álló szavakat vizsgáltunk, szóösszetételeket és többszavas kifejezéseket nem vettünk figyelembe. Az általános

gyakorlattól eltérően nem végeztünk szótövezést, ugyanis az – amint azt a későbbiekben megmutatjuk (l. lentebb) – fontos információk elvesztését eredményezhette volna a számunkra.

Az így meghatározott, a kapcsolatháló alapjául szolgáló együtteselőfordulás-mátrixot a WORDij[3] [25] felhasználásával hoztuk létre. Végül, a kirajzolódó kapcsolatháló áttekinthetősége és a releváns kapcsolatok kiemelése érdekében csak azon szavak közé kerülnek élek, amelyek egy mondaton belül három szónál kisebb távolságon belül és legalább két alkalommal közösen előfordulnak.

A kapcsolathálózatok vizualizációinak az interpretálhatóságát jelentősen növeli, ha a pontok és az élek egyaránt színesek [26]. A kapcsolatháló megfelelő (bizonyos könnyen is objektívan interpretálható attribútumok szerinti) színezése nem csak az eredmények interpretációja mellett az elemzést is segíti. Az egyes közösségdetektáló algrotimusok megjelenítése, nagy méretű kapcsolathálók esetén nehezen kivitelezhető, és a szín atribútum hozzárendelése a közösségek detektálásnak fundamentális alapját képezi. A fentebb felsorolt fundamentális alapok kivetelezésének eszközéül a kapcsolatháló-elemzés és vizualizáció terén széles körűen elterjedt szoftvert a Gephi 0.9.1[4] [27] választottuk. E szoftver a vizualizációhoz kapcsolódó fejlett algoritmusokkal és grafikai, animációs képességekkel rendelkezik.

A kapcsolathálóban a pontok nagyságát a közöttiség központiság mutatója [12], [28], [29] adja, a színét pedig az, hogy mely kontextuális klaszterbe tartozik. Az élek akkor kapnak egyedi színezést, ha az általa összekötött pontok azonos klaszterbe tartoznak. Azon pontok kerülnek azonos klaszterekbe, melyek inkább összekötöttek, egy azonos nagyságú és sűrűségű véletlenszerű [13] gráf esetén várható élek számához képest. Az így előállt klaszterek tehát tulajdonképpen a szövegben fellelhető "témák" feleltethetők meg.

3.3 Eredmények

A továbbiakban a tesztkorpuszunkon végzett elemzés eredményét mutatjuk be. A célunk itt inkább a szövegalapú kapcsolatháló-elemzés bemutatása, mintsem az adott témával kapcsolatos hipotézisek tesztelése - ez már csak abból is fakad, hogy az előzőekben leírtaknak megfelelően egy a teljes forrásanyag állományhoz képest egy nagyon nagy mértékben leszűkített teszt korpuszt elemeztünk.

A korpuszt leképező kapcsolatháló egy 806 pontból (szavak száma) és 783 élből áll. Ebből az látszik, hogy a háló ritka, azonban a hálóban található élek inkább egy szűkebb csoportban találhatóak. Ezt a csoportot mutatjuk a 4. ábrán.

[3] http://wordij.net/
[4] https://gephi.org/

- Személynevek, rangok, személyhez kapcsolódó utasítások, ügyek (13,4%)
- Formaságok, a bürokratikus mechanizmus elemi (5%)
- Utasítások, kinevezések, hozzárendelés (8,2%)
- Az államapparátus és karhatalom elemei (2,9%)
- Sajtóorgánumok, internacionális kapcsolatok, kulturális események (2,6%)

4. ábra. A tesztkorpuszt leképező kapcsolatháló

Ebben a pontok jelentős része 5-nél kevesebb kapcsolattal rendelkezik, s nagyon alacsony azon pontok száma, melyek negyvennél több kapcsolattal rendelkeznek. A pontok fokszámának átlaga 1.943, a legnagyobb előforduló fokszám 104. Ez alapján látható, hogy korpusz szókészletének varianciája alacsony és gyakori az azonos szópárok előfordulása.

A háló modularitása 0,65, ami azt mutatja, hogy a kapcsolathálóban azonosított klaszterek pontjai egy véletlenszerű gráf pontjainál jobban összekötöttek, tehát valós csoportokról beszélhetünk. A hálóban 420 közösséget találhatunk meg, amelyek közül öt darab nagyobb, mint a kapcsolatháló 2,5 %-a. Azok a pontok, melyek nem rendelkeznek kapcsolatokkal, magányos közösségeket alkotnak. Ezek nem láthatóak a fenti ábrán. Az azonosított témaklaszterek a teljes szöveg körülbelül egyharmadát tették ki.

Az elemzés során előálló klaszterek által lefedett témákat ugyancsak a 4. ábrán mutatjuk be. Látható, hogy a szöveg kapcsolathálóként való elemzése alapján jól elkülöníthetőek a fontos témák. Ez az eredmény már egy ilyen kis korpusznak az elemzésénél is fontos segítséget nyújthat, és egyúttal felhívja a figyelmünket egy a feldolgozással kapcsolatos érdekes jelenségre is.

Mint azt korábban említettük (l. fentebb), a feldolgozás során elhagytuk a szótövezést. Ennek jelentőségét az “Utasítások….” klaszter kiválóan mutatja, mivel ennek egyik központi szava az “elvtársat”, amely szót elveszítettünk volna, hogyha egy szótövezett korpuszon végezzük az elemzést.

4 Összefoglalás, további tervezett lépések

Dolgozatunkban ismertettük jelenleg futó kutatásunk nyelvtechnológiai szempontból is releváns munkálatait, valamint egy kisebb méretű korpuszon megmutattuk a szövegalapú kapcsolatháló-elemzés módszerének alkalmazását.

Kutatásunkban a Rákosi-korszakból származó pártgyűlési jegyzőkönyvek elemzésével foglalkozunk, célunk a párthierarchia mögött létező látens hierarchia felfedése a személyek közt zajló együttműködés, illetve bizonyos ügyekben való együttes érintettség vizsgálatával.

A feldolgozott és elemzett jegyzőkönyvekben a kapcsolatrendszert a nevek együtt-említettségével modelleztük.

A munkát több lépcsőben végeztük el. Mindenekelőtt, a digitalizált szöveg minősége szükségessé tette a szöveg korrekcióját, illetve a szövegekben lévő nevek korrekcióját. Kutatásunk jelenlegi fázisában az ezen korrekciót követő kapcsolatháló alkotást végezzük, amelyhez a jelentős élőmunka mellett komoly történelmi háttérismeretek szükségesek. Az ezt követő elemzéshez a szövegalapú kapcsolatháló-elemzést használjuk fel, melyet egy kis méretű tesztkorpuszon mutattunk be. Megmutattuk, hogy a szövegeket kapcsolathálóként értelmezve úgy csoportosíthatóak a szövegben található szavak, hogy a csoportok elemzésével beazonosíthassuk a szöveg által lefedett témákat.

Az elmondottakon túl, ugyanennek az elemzésnek az eredményeként arra is rámutattunk, hogy az általános elemzési megközelítéssel ellentétben nem célravezető

az ilyen elemzést megelőzően szótövezést végezni, mivel az elfedheti adott szavak különleges funkcióit. Jelen eredményeink szerint ugyanis a szövegben azonosított témák egyike középpontjában egy toldalékkal ellátott szó áll.

A munka további lépéseként azt tervezzük, hogy az előfeldolgozott és névelem-azonosított szövegen szentiment- és emócióelemzést hajtunk végre. E feldolgozási lépésekhez a szótárillesztés módszert kívánjuk alkalmazni, amely például a gépi tanulás mellett egyszerűbb és költséghatékonyabb információkinyerési módszer [30]. A szentimentelemzéshez olyan lexikonra van szükségünk, amely a lexikai szinten pozitív vagy negatív értékelő tartalommal rendelkező nyelvi elemeket tartalmazza [31], [32]. Az emóciók felcímkézéséhez pedig egy olyan szótárra, amely a különböző érzelmek nyelvi realizációit tartalmazza [33], [34]. Nyilvánvaló, hogy a korpuszban feldolgozott szövegek sajátságai okán a kiinduló szótárak szöveganyagát majd jelentősen módosítani kell, valamint azok kiegészítésére lesz szükség.

E két tartalomelemzési megoldástól azt reméljük, hogy a segítségükkel további, a kapcsolathálózat szempontjából fontos szemantikai tartalmakat tárhatunk fel a jövőben.

5 Bibliográfia

[1] K. Bozsonyi, Z. Horváth, and Z. Kmetty, ‘A hatalom hálója - A Kádár-kori hatalmi elit hálózati struktúrája az együttvadászási szokások alapján’, *Korall*, no. 47, pp. 157–184, 2012.

[2] G. Majtényi, *K-vonal - Uralmi elit és luxus a szocializmusban - Uralmi elit és luxus a szocializmusban*. Nyitott Könyvműhely, 2010.

[3] E. Sík, ‘Aczélhálóban’, *Szociol. Szle.*, vol. 3, pp. 64–77, 2001.

[4] I. G. Kovács, *Elitek és iskolák, felekezetek és etnikumok - Társadalom- és kultúratörténeti tanulmányok*. Budapest: L’Harmattan, 2011.

[5] A. Rácz, ‘A budapesti hatalmi elit prozopográfiai vizsgálata 1956-1989’, Budapest, 19-Dec-2014.

[6] R. Andorka, *Bevezetés a szociológiába*. Budapest: Osiris, 2006.

[7] A.-L. Barabási, ‘A hálózatok tudománya: a társadalomtól a webig’, *Magy. Tud.*, no. 11, pp. 1298–1308, 2006.

[8] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, Reprint edition. Cambridge: Cambridge University Press, 2008.

[9] L. Kovács, *Fogalmi rendszerek és lexikai hálózatok a mentális lexikonban*. Tinta Könyvkiadó, 2013.

[10] D. J. Watts, *Small Worlds*. Princeton University Press, 1999.

[11] D. J. Watts, ‘The “New” Science of Networks’, *Annu. Rev. Sociol.*, vol. 30, no. 1, pp. 243–270, 2004.

[12] D. Paranyushkin, ‘Identifying the Pathways for Meaning Circulation using Text Network Analysis’, Oct-2011. [Online]. Available: http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/. [Accessed: 23-Nov-2017].

[13] P. Erdős and A. Rényi, 'On Random Graphs I.', *Publ. Math. Debr.*, vol. 6, pp. 290–297, 1959.

[14] Barabási Albert-László, *Behálózva - A hálózatok új tudománya*. Helikon, 2002.

[15] R. Angelusz and R. Tardos, 'A gyenge kötések ereje és gyengesége', in *Hálózatok, Stílusok, Kultúrák*, Budapest: ELTE Angelusz Róbert Társadalomtudományi Szakkollégium, 2012, pp. 101–127.

[16] M. Granovetter, 'The Strength of Weak Ties', *Am. J. Sociol.*, vol. 78, no. 6, pp. 1360–1380, May 1973.

[17] M. Granovetter, 'A gyenge kötések ereje. A hálózatelmélet felülvizsgálata', in *Társadalmak rejtett hálózata*, R. Tardos and R. Angelusz, Eds. Magyar Közvéleménykutató Intézet, 1991, pp. 371–400.

[18] R. Németh, 'Módszerek a kvantitatív társadalomkutatási paradigmákban', *SOCIO.HU*, vol. 3, no. 10.18030/SOCIO.HU.2014.3.27, pp. 1–42, 2014.

[19] M. Sedighi, 'Using of co-word analysis method in mapping of the structure of scientific fields(case study: The field of Informetrics)', *J. Inf. Process. Manag.*, vol. 30, no. 2, pp. 373–396, Feb. 2015.

[20] K. Takács, *Kapcsolatháló elemzés; Társadalmi kapcsolathálózatok elemzése|Digitális Tankönyvtár*. Budapest: Budapesti Corvinus Egyetem, 2011.

[21] V. I. Levenshtein, 'Binary Codes Capable of Correcting Deletions, Insertions and Reversals', *Sov. Phys. Dokl.*, vol. 10, p. 707, 1966.

[22] J. Zheng, W. W. Chapman, R. S. Crowley, and G. K. Savova, 'Coreference resolution: A review of general methodologies and applications in the clinical domain', *J. Biomed. Inform.*, vol. 44, no. 6, pp. 1113–1122, 2011.

[23] E. Simon, 'A magyar nyelvű tulajdonnév-felismerés módszerei', Budapest, 2013.

[24] V. Vincze and R. Farkas, 'Tulajdonnevek a számítógépes nyelvészetben', in *Általános nyelvészeti tanulmányok XXIV.*, Akadémiai Kiadó, 2012, pp. 97–119.

[25] Danowski, J. A., 'WORDij version 3.0: Semantic network analysis software'. University of Illinois at Chicago, 2013.

[26] L. C. Freeman and V. Duquenne, 'A note on regular colorings of two mode data', *Soc. Netw.*, vol. 15, no. 4, pp. 437–441, 1993.

[27] M. Bastian, S. Heymann, and M. Jacomy, 'Gephi: an open source software for exploring and manipulating networks', presented at the International AAAI Conference on Web and Social Media, 2009.

[28] U. Brandes, 'A faster algorithm for betweenness centrality', *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, Jun. 2001.

[29] L. C. Freeman, 'A Set of Measures of Centrality Based on Betweenness', *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[30] F. Drávucz and M. K. Szabó, 'A beszélői szubjektivitás vizsgálata szentiment- és emóciókorpuszokon', in *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből*, 2017, pp. 39–49.

[31] M. K. Szabó, 'Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái', in *Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához*, vol. 177, T. Gecső and C. Sárdi, Eds. 2015, pp. 278–285.

[32] M. K. Szabó, ‘A nyelvi értékelés mibenlétének kérdése a számítógépes értékeléselemzés (szentimentelemzés) szempontjából’, in *LingDok 15. Nyelvészdoktoranduszok dolgozatai*, Z. Gécseg, Ed. Szeged: Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola, 2016, pp. 153–172.

[33] M. K. Szabó and G. Morvay, ‘Emócióelemzés magyar nyelvű szövegeken’, in *Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához*, vol. 177, T. Gecső and C. Sárdi, Eds. 2015, pp. 278–285.

[34] M. K. Szabó, V. Vincze, and G. Morvay, ‘Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái’, in *Távlatok a mai magyar alkalmazott nyelvészetben*, Budapest: Tinta Könyvkiadó, 2016, p. 282−292.

III. Korpusz, alapmodulok

Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl .hu domainjából

Indig Balázs

Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
{indig.balazs}@itk.ppke.hu

Kivonat Napjainkban az interneten fellelhető szövegek mennyisége és könnyű elérhetősége miatt a nagy korpuszok előállítása egyre könnyebb és szabványosabb. A kihívást manapság az internetes oldalak változékonysága jelenti. Bizonyos tartalmak eltűnnek és újak keletkeznek, valamint az oldalak közötti kapcsolatok is folyamatosan változnak. Egyre népszerűbbek az olyan közhasznú kezdeményezések, melyek „az internetet akarják archiválni". Az ilyen szolgáltatások a leszedett tartalmakat ingyen elérhetővé teszik bárki számára. A cikkben az így elérhetővé tett tartalmak minőségi jellemzőit vizsgáljuk korpusz-előállítás szempontjából, különös tekintettel az Amazonon tárolt Common Crawl archívumból elérhető magyar nyelvű tartalmakra. Célunk egy reprodukálható, könnyen elérhető technikákkal előállított magyar nyelvű webkorpusz létrehozása, mely publikusan elérhető. A korpusz így a nemrégiben elérhetővé vált `e-magyar` digitális nyelvfeldolgozó rendszer kiváló nyersanyagát fogja képezni.
Kulcsszavak: web crawler, korpuszépítés

1. Bevezetés

Az interneten megjelenő tartalmak számszerű növekedésével és változásuk gyorsulásával együtt megjelent az az igény, hogy a jövő számára lenyomatot készítsünk és egyfajta „digitális levéltárban" megőrizzük az egyes weboldalak különböző verzióit. Az internet rohamos fejlődése miatt a technológiák gyorsan váltják egymást, így bizonyos, egy időben valamilyen szempontból fontosnak tartott tartalmak gyakran megváltoznak, majd eltűnnek.

Célszerűnek tűnik tehát egyfajta archiválási technika bevezetése. A Wikipedia például a kezdetektől fogva „vandálbiztos", azaz minden változtatás egy új verziót hoz létre az adott oldalból és a változások visszakövethetők, ahogy az napjainkra a programok forráskódjánál alapvető elvárás. Ez az ismérv jelenleg még csak nagyon kevés oldalra jellemző. Ezért több elképzelés született az internetes oldalak külső fél általi archiválására, melyek napjainkra szabványosítva lettek. Ebben a formátumban több szervezet is közzé teszi az általa gyűjtött tartalmakat.

A nyelvtechnológia célja természetesen elsődlegesen nem az archiválás, hanem az archivált anyagok elemzése és feldolgozása. A cikkben az *Amazon* által

tárolt és szolgáltatott Common Crawl archívum példáján mutatjuk be, hogy hogyan használható egy ilyen archívum arra, hogy reprodukálható módon korpuszt lehessen belőle építeni, mely később az eszközök fejlődésével újra és újra, jobb minőségben születhet meg. A létrejött szabadon elérhető korpusz kiváló kiegészítője a nemrégiben publikált szabadon elérhető teljeskürű eszközláncnak, az `e-magyar` digitális nyelvfeldolgozó rendszernek [1].

A cikkben először összehasonlításképpen bemutatjuk az elérhető magyar nyelvű korpuszokat, majd a web archiválásának fontosabb szereplőit és szabványait. Ezek után a létrehozandó korpuszhoz szükséges módszert ismertetjük, a cikk végén pedig a létrehozott korpusz és a Common Crawl archívumban található magyar nyelvű tartalmak jellemzőit tárgyaljuk.

2. Előzmények

2.1. Fontosabb magyar nyelvű korpuszok

Magyar nyelvre több nagyobb korpusz is létezik, melyek részben vagy egészben a webről származnak (lásd az 1. táblázat). A bárki számára feldolgozott állapotban azonnal hozzáférhető *Webkorpusz* [2] szűretlen változata 1,48 milliárd, míg a szűrt változata 589 millió szót tartalmaz. A maga idejében a legnagyobb magyar nyelvű korpusz volt, mely közvetlenül elérhető egy rendkívül megengedő licenc alatt. A korpusz forrásául a .hu domain 2003-as állapota, azaz 18 millió letöltött oldal szolgált. A gyűjtött anyag jól reprezentálja az írott nyelv igen széles skáláját. A készítés során a többször is megtalálható szövegek kiszűrésre kerültek, ahogy azok a fájlok is, amelyek nem tartalmaztak használható szöveget. A szűrés több lépcsőben történt, aszerint, hogy egy oldalon belül a szavak hány százalékát ismerte fel a helyesírás-ellenőrző program. A szerzők által is jó minőségűnek tartott, és közzétett korpuszrészlet a teljes korpusz 4%-át tartalmazza.

A Webkorpusz készítése során a letöltött oldalak teljes HTML változata nem lett megőrizve a nagy tárhelyigény miatt. Viszont az elkészítéséhez használt eszközök minősége sokat javult azóta, melyekkel az eredeti HTML oldalak hiányában nem reprodukálható a letöltött nyers oldalak feldolgozásának folyamata. Ez a tény a korpusz minőségén történő utólagos javítást rendkívüli módon megnehezíti.

A *Magyar Nemzeti Szövegtár (MNSZ)* [3,4] két fő verzióját különböztetjük meg. Az első változata 187 millió szövegszót tartalmaz, míg a második kiadás (MNSZ 2) az elsőt magába foglalva 1,348 milliárd szóból áll[1]. A korpusz különlegessége, hogy nem csak webről származó szövegeket tartalmaz, hanem van egy határon túli magyar nyelvű szövegeket valamint egy beszélt nyelvi átiratokat tartalmazó alkorpusza is.

A nyers szövegek, illetve a feldolgozáshoz szükséges eszközök nem szabadon letölthetőek egy eszközláncként, ahogy maga a korpusz is csak kérésre elérhető

[1] A korpuszon nem történt duplikációszűrés, így tényleges méret ennél valamivel kisebb lehet az ismétlődések miatt.

szöveges formában, valamint regisztráció után a keresőfelületen használható. A reprodukció és a minőség javítása ezért egyedül a korpusz készítőitől függ.

A *Pázmány korpusz* [5] teljes egészében a 2015-ös magyar webről származó szövegeket tartalmaz. Az 1,24 milliárd token célzottan kiválasztott magyar hírportálokról származó oldalakon található, duplikációszűrt szövegekből áll. A korpuszban meg vannak különböztetve továbbá a hírportálok szerkesztett, a hírhez tartozó szövegrészei, valamint a szerkesztetlen, a hírhez érkezett kommentek szövegeit tartalmazó alkorpuszok.

A készítésekor a szerzők az általuk kitalált speciális *Goldminer-algoritmust* [6] használták, mely *crawl*-olás közben egy mintavétel alapján egy oldal kinézetét jobban meghatározva hatékonyabb sablonszűrést (*boilerplate removal*) tesz lehetővé, mint az addigi state-of-the-art megoldás, ami külön-külön vizsgálta az oldalakat. Sajnos a korpusz jelenleg nem érhető el szabadon. A készítéshez használt eszközök egy része elérhető[2], de nem közvetlenül felhasználható formában. A nyers HTML oldalak ebben az esetben sem lettek megtartva a magas tárhelyigény miatt.

Név	Méret (token)	forrás	készítés éve
Webkorpusz (teljes)	1,48 milliárd	web vegyes	2003
Webkorpusz (4%)	589 millió	web vegyes	2003
Magyar Nemzeti Szövegtár I	187 millió	újság, beszédátirat, web vegyes	2001
Magyar Nemzeti Szövegtár II	1,348 milliárd	újság, beszédátirat, web vegyes	2017
Pázmány korpusz	1,24 milliárd	web hírportálok, kommentek	2015

1. táblázat. A főbb magyar nyelvű korpuszok adatai

Látható, hogy a korpuszépítés során problémát jelent a nyers HTML oldalak tárolása, illetve az esetek jó részében a korpuszt készítő programok sem elérhetőek teljes mértékben. A következő fejezetben az előbbi probléma egy lehetséges megoldását ismertetjük.

2.2. Web archiválása napjainkig

Az internet gyorsan változó világában nagyon fontos a pontos metainformációkkal együtt történő precíz és gyors archiválás, ugyanis az egyes tartalmak elérhetősége néha igen rövid időt ölel csak fel. A verziókezelés, a duplikátumok szűrése és a keresés is fontos szempont a későbbi feldolgozás céljából. A feladatot több nagyobb szervezet iparági összefogással próbálja megoldani.

Az egyikük, az *Internet Archive* egy San Franciscó-i nonprofit digitális könyvtár, amely szabad hozzáférést biztosít számos digitalizált tartalomhoz, melyek közé tartoznak többek között az archivált weboldalak, valamint szabad jogállású könyvek digitalizált változatai is. A teljes archívum mérete 2016 októberében elérte a 15 petabájtot.

[2] `https://github.com/endredy/GoldMiner`

Bár szabadon is feltölthetőek tartalmak, az anyagok nagy része mégis internetes crawlerek segítségével kerül a rendszerbe. A webarchívumuk, a *Wayback Machine*[3] már több mint 308 milliárd oldalt rögzített a metaadataival együtt.

A szervezet által archiválásra kifejlesztett fájlformátum a *WARC (Web ARChive)* – mely az Internet Archive ARC fájlformátumának (ARC) általánosított, szabványosított változata – meghatározza a pontos módját annak, hogy hogyan kell egységesen elraktározni a sokféle digitális erőforrást a hozzájuk tartozó összes metainformációval együtt. Ezen felül támogatja a keresést és az egyes oldalak közötti navigációt, valamint a duplikátumok hatékony tárolására vonatkozó verzióinformációkat. Manapság ez a formátum a web archiválásának szabványos módjává vált.

A *Common Crawl* egy nonprofit szervezet, amely a saját maga által webről leszedett és archivált tartalmakat nyers formában szabadon hozzáférhetővé teszi a lehető legszélesebb közönség számára. Céljuk, hogy az internetet „tehermentesítsék" azáltal, hogy nem kell mindenkinek magának letöltenie a weboldalakat. 2017 novemberére a webarchívumuk mérete 260 TB-ra rúg, amely 3,2 milliárd weboldalról származik[4]. Az archívumot 2012 óta az Amazon szolgálja ki a publikus adathalmazok programjának részeként[5].

Az eszközök tekintetében 2013-tól kezdve a saját crawler helyett az *Apache*-féle szabadon elérhető *Nutch webcrawler*-t használják a saját implementáció helyett, valamint átálltak a szabványos WARC formátumra. 2015-től kezdve az összes archivált tartalom indexelve is elérhető[6], így könnyebben kereshető az URL-ek szerint is.

A CommonCrawl három formátumban teszi elérhetővé a leszedett tartalmat:

- WARC: a leszedett nyers adatok
- WAT: a WARC formátumban található adathoz tartozó származtatott metadatok
- WET: a WARC fájlokból kinyert szöveges tartalom (a kinyerés módjáról nincs elérhető információ)

A későbbiekben a WARC formátummal foglalkozunk, mivel az számunka elég információt tartalmaz: (1) a *crawl*-olási folyamat pontos leképezését adja, (2) tárolja a HTTP kérést és választ a fejléccel együtt, (3) valamint a tényleges tartalom metaadatai is rögzítve vannak.

3. A magyar web a Common Crawl tükrében

Az látható, hogy a webről történő korpuszépítés módját nagyban meghatározzák az eddig döntően angol nyelvű törekvések. Szerencsére ezek egy az egyben átültethetők magyar nyelvre a szabadon elérhető gazdag eszköztárat felhasználva.

[3] `http://archive.org/web/`
[4] `http://commoncrawl.org/2017/11/november-2017-crawl-archive-now-available/`
[5] `https://aws.amazon.com/public-datasets/common-crawl/`
[6] http://index.commoncrawl.org/

A magyar szövegek kinyerésének szempontjából két problémát fedezhetünk fel. Egyfelől, a szövegkinyerő eljárások terén még számíthatunk fejlődésre, így a leszedett tartalmat meg kell őrizni. Másfelől pedig, a futtatott crawlerek céltalanul bolyonganak a weben, és így az egyes oldalak fontosságuktól függetlenül vannak reprezentálva az archívumban. Ezért célszerű lehet a jövőben azonos technológiával saját crawlereket futtatni, speciálisan a magyar web archiválásának céljából, valamint mindig a state-of-the-art sablonszűrő algoritmus használatával kinyerni a szöveget a leszedett tartalomból.

Jelen cikkben a célunk, hogy megvizsgáljuk, hogy a szabadon hozzáférhető anyagból milyen mennyiségű és minőségű magyar nyelvű korpusz gyártható. Az így készült korpusz előnye az eddigiekhez képest – a méretétől és tartalmától függetlenül – az, hogy az eredmény reprodukálható és szabadon hozzáférhető lesz, valamint folyamatosan frissül és bővül, lehetővé téve diakron elemzéseket is.

3.1. Az architektúra

Az archívum az Amazon által biztosított API-n keresztül hozzáférhető. A fellelhető technikai leírásokból látható, hogy az Amazon legfőbb prioritása, hogy a saját felhős rendszerein használható legyen az adathalmaz, így az elérhető példaprogramokból a felhőn kívüli használat hiányzik. Az elérhető példakódok többnyire Python2 és JAVA nyelven íródtak, de más nyelvekre is elérhetőek, mivel az API a szabványos HTTPS protokollra épül. A cikkben implementált eljárásokhoz a Python nyelvet használjuk, ezért a későbbiekben csak a Python nyelvű API-val foglalkozunk. Bár a példaprogramok csak Python2 nyelven érhetőek el, a szükséges könyvtárak Python3 nyelven is elérhetőek [7], így lehetővé téve a példaprogramok egyszerű átírását, amivel a Python nyelv 3-as dialektusának legújabb verziói adta előnyöket is tudjuk használni.

Az archívum indexe külön szerveren található, az Amazontól függetlenül. Egyszerű REST API kérésekkel hívható és JSON formátumban adja vissza a kívánt kimenetet. A keresett URL-eket tartalmazó WARC fájlok letöltéshez szükséges mezők nevei viszont nehezen találhatók meg a dokumentációban, és a példa programok nem tartalmazzák a lehetséges mezőket. Az API dokumentációja szerencsére kellően részletes a szűrési lehetőségek és az elérhető mezők tekintetében. Maga az archívum havonta frissül, így az index is havi bontásban érhető el. Az indexben URL szerint és a *crawl*-olás hónapja szerint lehet keresni. Maga az index alatt lévő szoftver nyílt forráskódú és szabadon felhasználható saját, azonos formátumban lévő archívumok indexelésére is.

A kívánt oldal letöltéséhez szükséges egy az indexből kiszedett speciális mezőkből összerakott URL, melyet az Amazon sajátos konvencióinak megfelelően kell átadni. Ebben segít az SDK-ban található magas szintű API, melyben a következő lépéseket kell végrehajtani a megfelelő oldal gzip formátumban történő letöltéséhez:

1. Létrehozunk egy munkamenetet az *s3* szolgáltatásra.

[7] `http://boto3.readthedocs.io/en/latest/`

2. Explicit módon az anonim bejelentkezést választjuk.
3. Kiválasztjuk a *commoncrawl* nevű *bucketet*.
4. Megadjuk a *key*-ként az indexből kapott fájlnevet.
5. Megadjuk a *range* paraméterként (szöveges formátumban kötőjellel elválasztva) az indexből kapott *offset* és *length* értékeket.
6. Elküldjük a HTTP kérést, melynek eredményeként megkapjuk a WARC fájlt.

Az így kapott fájl tartalmazza a HTTP kérésre és válaszra vonatkozó adatokat is a tartalom mellett. A metaadattól ezek után egyszerűen különválasztható a tényleges tartalom a WARC fájlformátum tervezéséből adódóan. Ezek után a megkapott oldalból ki kell nyerni a szöveges adatot, és fel kell dolgozni az így kapott nyers szöveget a szokásos szerelőszalagon.

Elmondhatjuk tehát, hogy az architektúra három részből áll, az indexből ki kell nyerni a kívánt oldalak letöltéshez szükséges adatait, le kell tölteni az egyes oldalakat, és a tárhelykapacitás korlátossága miatt röptében feldolgozni őket az ismert eszközökkel. A szabadon elérhető, Python nyelven jól használható eszközök hiánya miatt, csak a *JusTextet* [7] alkalmaztuk demonstrációs céllal, mivel így a letöltött és feldolgozott anyag tárolása nem ütközött korlátokba. A *JusText* program sajátosságából fakadóan elvégezte a szöveg bekezdésekre tagolását is, amelyet megtartottunk.

Szükséges lenne még egy duplikációszűrés[8], valamint a feldolgozásban következő lépésként a mondatrabontás és tokenizálás, mely utóbbi a már említett e-magyar rendszer első lépését adja. Ezek a műveletek túlmutatnak a jelenlegi cikk célján, így nem tárgyaljuk őket. A következő fejezetben az index és a letöltött szövegek jellemzőit, valamint a letöltés tapasztalatait foglaljuk össze.

3.2. Tapasztalatok

A letöltés során fel kell készülni hálózati hibákra, valamint az adat különböző hiányosságaira is. Egyszeri tesztekből könnyen kideríthető, hogy nagyon lényeges szempont az ütemezés kérdése, mivel egy szálon maximum 50kb/s-re van limitálva a kapcsolat sebessége, valamint a kapcsolatok élettartalma is korlátozott. Ilyenkor a program úgy érzékeli, hogy a másik fél bontotta a kapcsolatot és a túlterhelés elkerülése végett automatikusan egy kis időt várakozik az újracsatlakozás előtt.

A többszálú futtatáshoz a legyegyszerűbb módszer a *GNU parallel* [8] használata. Így akár több gépen, párhuzamosan több szálon is tudjuk folytatni a letöltést. A feladat jellege miatt nem szabad ragaszkodni a magonként egy szál futtatásához, mivel így nem használjuk ki az erőforrásainkat. Az optimális szám tapasztalataink szerint kb. a processzormagok 3,25-3,5 szöröse.

Mivel az indexben minden információ benne van, amely a *crawl*-olás folyamatával kapcsolatos, ezért ajánlott a letöltendő oldalakat *a priori* megszűrni a rendelkezésre álló mezők alapján:

[8] A duplikációszűrés műveletének elvégzésére ígéretes program az Onion [7], mely a `http://corpus.tools/` oldalról elérhető.

- robots.txt: a crawler működéséből adódóan minden domainen először a szabványos robots.txt-t keresi, mely nem tartalmaz számunkra használható információt.
- HTTP 3xx, 4xx, 5xx kódok: nagyon sok weboldal alkalmaz átirányítást, melyek külön elemként jelennek meg az indexben, de szöveges tartalommal nem rendelkeznek.
- mime-típus: a leszedett tartalmak sokfélék lehetnek, számunkra jelenleg csak a HTML formátumú adat értékes, így érdemes erre is szűrni.

Az így megszűrt indexbeli találatok sem jelentenek garanciát arra, hogy az adott kulcson elérhető tartalom, és valóban használható lesz számunkra. Ezért további, *a posteriori* szűrőket vezettünk be a letöltött tartalom vizsgálatára, melyek már nem az indexben található információkon alapulnak. Így külön választottuk az alábbi hibatípusokat, melyek alapján a letöltött oldalak eldobásra kerültek:

- hossz túl rövid: ha a tartalom nem volt legalább 13 karakter hosszú („<html> </html>”)
- dekódolási hiba: nem helyes UTF-8 tartalom
- jól formáltsági hiba: nem jól formált XML tartalom
- nem található kulcs: az adott oldalra mutató kulcson nem található tartalom
- nincs szöveg: a sablonszűrés után nem maradt szöveges tartalom
- egyéb hibák: ide tartoznak a hálózati hibák és egyéb nem várt jelenségek (melyek többszöri próbálkozás után is fennállnak)

Bár a fent felsorolt hibák egy része javítható – például a helyes karakterkódolás meghatározásával –, jelenleg mégis az eldobásuk mellett döntöttünk, hogy felmérjük az ilyen hibák arányát, valamint el akartuk kerülni, hogy külön kelljen választani azokat az oldalakat, ahol a hibás kódolás miatt egyes szövegrészek egymástól eltérő karakterkódolásban vannak. Azokban az esetekben, ahol az indexben található elem nem volt megtalálható az API által a felhőben, külön választottuk azt, későbbi vizsgálat céljából.

4. Eredmények

A fenti szűrések elvégzése után is sok olyan eset fordult elő, ahol az *a priori* szűrés nem volt elég hatékony. Például a hiányzó mezők, valamint a szerveren rosszul feltüntetett adatok miatt sok elem az *a posteriori* szűrésen került kiszűrésre. A 2. táblázatban az *a posteriori* szűrés eredménye látható.

Mivel az Amazon 50kb/s-ben korlátozza az egy kapcsolatra jutó sávszélességet, ezért több szálon kellett futtatni a letöltést. Az összesen 64 processzormagon futtatott letöltés esetében a teljes letöltés kb. 5 napig tartott. A letöltési időt növelte, hogy az egyes HTTP kérések nem voltak újrahasznosíthatóak, mivel az Amazon bizonyos idő után mindenképpen felbontotta őket, amit a program úgy érzékelt, mintha időtúllépés miatt megszakadt volna a kapcsolat, és ezért

típus	darab	%
letöltött oldal	73 187 167	54.723988
nincs szöveges tartalom	60 294 819	45.084037
hossz túl rövid	189 006	0.141325
kulcs nem található	44 240	0.033079
dekódolási hiba	18 446	0.013793
jól formáltsági hiba	4 718	0.003528
egyéb hiba	208	0.000156
kitömörítési hiba	126	0.000094
összesen	133 738 730	100

2. táblázat. A különféle elemek típusai az *a posteriori* szűrés után

kis várakozás után mindig újra létrehozta a kapcsolatot. Így nem a processzorok teljesítménye vagy a háttértár sebessége jelentette a nehézséget.

Az egyes domaineken belüli oldalak eloszlása hatványeloszlást mutatott. A 10 leggyakoribb domain a 3. táblázatban látható. Ezek nagy része webáruház, melyben többnyire rövid, pársszavas termékleírások találhatók egymás alatt. Ezzel magyarázható többek között a szöveges tartalmak alacsony aránya a letöltött oldalakhoz képest.

domain	darab	%
arukereso.hu	2 274 806	1,53
kirakat.hu	2 136 393	1,44
munkatarsaim.hu	2 094 399	1,41
docplayer.hu	2 058 703	1,39
index.hu	2 006 132	1,35
conrad.hu	1 947 660	1,32
lira.hu	1 772 434	1,20
zoover.hu	1 647 372	1,11
olcsobbat.hu	1 607 617	1,08
globalplaza.hu	1 602 395	1,08

3. táblázat. A tíz leggyakoribb domain gyakorisága és százalékos aránya

Összességében az elkészült korpusz 73 187 167 dokumentumból származó, 52 114 233 (nem tokenizált) szövegszóból és 1 396 844 paragrafusból áll, amely a 2.1. fejezetben bemutatott korpuszok tartalmánál jóval kisebb. Ez a szám a duplikációmentesítés után várhatóan még csökkenni fog. A létrejött korpusz előnye a többi felsorolt korpuszhoz képest, hogy reprodukálható, valamint rövid idő alatt előállítható mégis kellően nagy mennyiségű szövegből áll. A jövőben jó alanya lehet a sablonszűrő algoritmusok tesztelésének, valamint folyamatos bővítésével mérete is egyszerűen tovább növelhető.

5. Összegzés

A cikkben bemutattuk a kurrens webarchiválásra szolgáló technológiákat, melyekkel könnyen létrehozhatóak saját, szabványos eszközökkel feldolgozható webarchívumok és azokból kinyerhető korpuszok is. Megvizsgáltuk, hogy a szabadon elérhető, az Amazonon tárolt Common Crawl archívum .hu domainből származó része mennyire alkalmas korpusz építésére. Az előállításhoz szükséges eszközöket szabadon elérhetővé tettük[9], mellyel egy új, szabadon elérhető és reprodukálható magyar nyelvű korpuszt hoztunk létre. Habár a korpusz mérete nem haladja meg az elérhető korpuszokét, fontos tulajdonsága, hogy bármikor reprodukálható. A reprodukálhatóság előnye pedig, hogy a szöveget feldolgozó eszközök fejlődésével egy szinttel korábbról kezdődően elvégezhető a korpusz újbóli létrehozása – a fejlett eszközöknek köszönhetően – jobb minőségben.

A jövőbeli terveink között szerepel a hibák javítása, illetve a későbbi crawlok hozzáadás a korpuszhoz. További terv, hogy megvizsgáljuk, hogy mennyi duplikáció van a kinyert szövegben, mert az ilyen duplikációk nagyban torzítják a statisztikai méréseket. Végső célunk, hogy a CommonCrawl archívumtól a feldolgozott korpuszig vezető teljes út lefedése – mindenki számára elérhetően, ahol lehet, meglevő eszközökkel –, mely a későbbiekben az felhasznált eszközök fejlődésével újrajátszható marad.

Hivatkozások

1. Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholcz, I., Novák, A., Indig, B., Prószéky, G., Farkas, R., Vincze, V.: Az `e-magyar` digitális nyelvfeldolgozó rendszer. In Vincze, V., ed.: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017), Szeged, Szegedi Tudományegyetem Informatikai Intézet, Szegedi Tudományegyetem Informatikai Tanszékcsoport (2017) 49–60
2. Halácsy, P., Kornai, A., László, N., András, R., Szakadát, I., Viktor, T.: Creating open language resources for hungarian. In N, C., ed.: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). (2004) 203–210
3. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002) European Language Resources Association, Paris, ELRA (2002) 385–389
4. Oravecz, C., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Calzolari, N., et al., eds.: Proceedings of the 9th International Conference on Language Resources and Evaluation, May 26-31, 2014, Reykjavik, Iceland, ELRA (2014) 1719–1723
5. Endrédy, I.: Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz. PhD thesis, PPKE-ITK, Budapest (2016) PhD disszertáció.
6. Endrédy, I., Novák, A.: More Effective Boilerplate Removal - the GoldMiner Algorithm. Polibits - Research journal on Computer science and computer engineering with applications (48) (2013) 79–83
7. Pomikálek, J.: Removing boilerplate and duplicate content from web corpora. PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic (2011)

[9] `http://github.com/ppke-nlpg/commoncrawl-downloader`

8. Tange, O.: Gnu parallel - the command-line power tool. ;login: The USENIX Magazine **36**(1) (2011) 42–47

Normo: Egy automatikus normalizáló eszköz középmagyar szövegekhez

Vadász Noémi[1,2], Simon Eszter[1]

[1]MTA Nyelvtudományi Intézet
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
E-mail: {vadasz.noemi, simon.eszter}@nytud.mta.hu

Kivonat A cikk egy automatikus normalizáló eszközt ismertet középmagyar szövegek normalizálásához. A Normo két modulból áll: egy memóriaalapú modulból és egy szabályalapúból, amely karakter- és tokenszintű környezetfüggő újraíró szabályokat tartalmaz. Az eszköz meggyorsítja és megkönnyíti a középmagyar szövegek kézi normalizálását, amelynek eredménye a további nyelvfeldolgozó eszközök bemenete. Az eszköz ismertetése után a modulok teljesítményét külön-külön és egyben is kiértékeljük.
Kulcsszavak: normalizálás, szabályalapú normalizálás, memóriaalapú normalizálás, történeti szövegek, középmagyar

1. Bevezetés

Az annotált nyelvi erőforrások elérhetősége egyre fontosabb szerepet kap a nyelvészet több területén: a nyelvtechnológiai fejlesztéseken kívül az elméleti és történeti nyelvészeti kutatásoknak is kiváló alapanyagot szolgáltatnak a korpuszok. A történeti korpuszok az adatok és a nyelvi jelenségek gazdag tárházát adják – de csak akkor, ha a releváns információ elektronikusan interpretálható és előhívható módon van tárolva bennük. A nyelvtörténészek és nyelvtechnológusok egyik legfontosabb együttműködési terepe a történeti korpuszok építése. Az elmúlt évtizedekben sorra indultak olyan projektek, melyek egy adott nyelv valamely régebbi változatának digitalizálását és feldolgozását célozták – elsősorban indoeurópai nyelvekre, például [1,2]. Ebbe a sorba illeszkedik az Ómagyar Korpusz [3] is, amely tartalmazza az összes fennmaradt ómagyar kori (896–1526) szövegemléket, valamint bizonyos középmagyar kori (1526–1772) írott és nyomtatott szövegeket, továbbá néhány szövegemlék normalizált és morfológiailag elemzett és egyértelműsített változatát[1].

Napjainkban a korpuszépítési munkálatok során elsősorban már digitalizált szövegekből indulnak ki; de nem ez a helyzet a történeti dokumentumokkal. Az elektronikus formátumok (sőt az elektromosság) előtti korból származó szövegekkel való foglalkozás sokkal idő- és munkaigényesebb folyamat, és bizonyos

[1] http://omagyarkorpusz.nytud.hu/

esetekben más módszereket is igényel, mint a mai szövegek esetében. A helyesírás és a központozás a régebbi nyelvváltozatok korában nem volt konzisztens, ezért a modern szövegekben alkalmazott sztenderd előfeldolgozó lépések (tokenizálás, mondatra bontás, morfológiai elemzés és egyértelműsítés) nem végezhetők teljesen automatikusan, és nagyon sok kézi ellenőrzést igényelnek.

A magyar írásosságot a latin nyelvű és vallásos tárgyú irodalom fordításának igénye hívta életre, de a latin ábécé magyarra alkalmazása számos problémát vetett fel. A legfőbb gond abból fakadt, hogy nyelvünk hangrendszerének több eleme a latinban ismeretlen, így ezek jelölésére új jeleket kellett bevezetni. A helyesírás ezekben a századokban távolról sem volt egységes. A különböző helyesírási rendszerekben is ritka az egy hang–egy betű megfelelés (vagyis amikor egy hang jelölésére mindig ugyanaz a betű használatos, és az adott betűnek mindig egy hangértéke van), de egy alakulóban levő helyesírási rendszerben ilyenfajta következetesség még annyira sem várható el. Ezért szükség van egy ún. normalizálási lépésre, amelynek során az eredeti betűhű szóalakokat mai magyar helyesírású szavakra alakítjuk át. A többféle, különböző nyelvtörténeti szakmai érvekkel alátámasztható lehetséges feldolgozási forgatókönyvek egyik gyakori közös átalakító lépése ez a fajta normalizálás [4,5]. A szövegfeldolgozásnak ez a lépése kritikus fontosságú, enélkül ugyanis a (félig) automatikus annotáció hatékonysága a következő lépésekben drámaian visszaesik [6].

Mivel a normalizálás nyelvtörténeti szakértelmet kívánó, rendkívül időigényes manuális munka, megpróbáltuk kiváltani gépi eljárással. A normalizálás nyelvtechnológiai szempontú kutatásának igen gazdag eszköztára van, mivel a normalizálási feladatot jellemzően más nyelvtechnológiai feladatokkal szokták párhuzamba állítani és azok eszközkészletét használni a feladat megoldására. Az egyik megközelítés szerint a normalizálás során egy forrásnyelvről (jelen esetben a régi nyelvváltozatról) fordítunk egy célnyelvre (a modern nyelvváltozatra), így a gépi fordítás módszerei használhatók [7,8]. Egy másik megközelítés a normalizálást egy ábécéről egy másikra való átírásként, vagyis transzliterációként fogja fel, amire a Shannon-féle zajos csatorna modellt kiválóan lehet alkalmazni [9,10]. A normalizálási munkák nagy része viszont kézzel írott vagy korpuszból kigyűjtött megfeleltetési szabályok [11] és/vagy távolságmetrikák [12] alkalmazásán alapul.

A cikkben egy olyan normalizáló eszközt ismertetünk, amelyet középmagyar szövegek kézi normalizálásának a támogatására fejlesztettünk. Jelen állapotában az eszköz a Károli Gáspár-féle középmagyar bibliafordításra lett optimalizálva, ami a teljes Biblia első magyar nyelvű fordítása 1590-ből. Ennek ellenére a Normo nem csak egy memória- és szabályalapú eszköz ennek az egy szövegnek a normalizálásra, hanem egy olyan keret, amely további szótárak és szabályok hozzáadásával alkalmassá tehető más történeti szövegek normalizálására is. Nem kétséges, hogy a feladat a fent felsorolt technikák bármelyikével is megoldható volna, de mi különféle, elsősorban a projektből fakadó, praktikus kényszerek és követelmények hatására a memória- és szabályalapú technikák ötvözését választottuk.

A cikk az alábbiak szerint épül fel. A 2. fejezetben azt ismertetjük, hogy melyek azok az alfeladatok, amelyeket a kézi normalizálást végző annotátorok elvé-

geznek, vagyis meghatározzuk, hogy mi az ideálisan elérendő cél. A 3. fejezet a Normo által alkalmazott módszerekről szól: a 3.1. fejezet írja le a memóriaalapú modult, míg a 3.2. fejezet a szabályalapú normalizálást ismerteti. A 4. fejezetben a Normo alapos kiértékelését adjuk, összehasonlítva egy távolságmetrikát alkalmazó másik, történeti szövegek normalizálására kifejlesztett eszközzel. A cikket az 5. fejezetben összefoglalás zárja.

2. A kézi normalizálás

A kézi normalizálás során az annotátorok olyan feladatokat is elvégeznek, amelyek kívül esnek a szűkebb értelemben vett normalizálás feladatkörén, de a szöveg további feldolgozását készítik elő. Így a tulajdonképpeni normalizálással párhuzamosan bekerülnek a megfelelő lókuszjelölők a szövegbe, megtörténik a (tag)mondatokra bontás és a kis/nagybetűsítés is. A tokenizálás javítása is megtörténik, ahol a mai magyar helyesírásnak megfelelően bizonyos szavakat fel kell bontani, míg másokat össze kell vonni – természetesen jelölve a változtatásokat. Ahol szükséges, ott az annotátorok megjegyzést vagy értelmezést is adhatnak az egyes szavakhoz, valamint ebben a fázisban történik az igekötők kódolása is a megfelelő igéhez vagy igenévhez. Ennek megfelelően a szöveg kézi normalizálásán egy tágabb értelemben vett előkészítő munkát értünk, amely magába foglalja a szűkebb értelemben vett normalizálás mellett a fent ismertetett egyéb előkészítő feladatokat is.

A gépi normalizálás során ezekből az alfeladatokból néhányat ki tudunk váltani automatikus eszközökkel. A tokenizálás annyiban tud automatikusan történni, amennyiben azt az egyes tokenek felszíni jegyei lehetővé teszik. Egyrészt a tapadó írásjeleket leválasztjuk a szavakról, másrészt a sorvégi kötőjellel jelölt elválasztott szóelemeket összevonjuk, és speciális karakterekkel jelöljük az elválasztás és összevonás tényét és helyét. Bár a ma használatos logikai–grammatikai írásjelezés kibontakozása csak a 17. században kezdődik, a mondatra bontásban mégis csak a korabeli központozásra támaszkodhatunk, így minden írásjel után beszúrunk egy tagmondathatár-jelölő üres sort. A szűkebb értelemben vett normalizálás során ugyanazokat az elveket követjük, mint amiket az annotátorok követnek a kézi normalizálás során. Egyrészt a ma nem létező összes szót, toldalékot, morfológiai konstrukciót meg kell tartani, vagyis morfémát nem toldhatunk be és nem hagyhatunk el. Másrészt viszont el kell hagynunk minden fonológiai és helyesírási esetlegességet, vagyis egységes, amennyire lehet, a mainak megfelelő helyesírásra kell törekedni.

Mivel a normalizált szöveg képezi a további szövegfeldolgozó lépések bemenetét, a cél a lehető legpontosabban normalizált szöveg, ezért a normalizálást annotátorok végzik kézzel. Ugyanakkor a kézi normalizálás időigényes és drága feladat. Az automatikus normalizáló eszköz segítségével le lehet rövidíteni a normalizáláshoz szükséges időt és költségeket. Az annotátorok visszajelzései alapján az automatikus normalizáló kimenetének ellenőrzése és javítása nem csak gyorsabb, hanem egyszerűbb is a betűhű szöveg kézzel történő normalizálásánál, amikor minden egyes szóalakot egyesével kell begépelni. A Normo kimenetének kézi

javítása és a teljesen kézi normalizálás közötti munkaráfordításbeli különbségről lásd a 4. fejezetet.

3. Módszerek

A NORMO két fő modulból áll: egy memória- és egy szabályalapú modulból. A szabályalapú modul karakter- és tokenszintű környezetfüggő szabályokból áll. A karakterszintű szabályok a tokeneken belül, a tokenszintű szabályok a szóhatárokon keresztül operálnak, sőt ez utóbbiak között vannak olyanok is, amelyek tagmondathatárt is változtatnak.

A tokeneken belüli változtatások kétféleképpen történhetnek: a memóriaszótár vagy a karakterszintű újraíró szabályok segítségével – vagyis ez utóbbi két modul egymástól független. Mindkettő kimenete viszont bemenetét képezheti a tokenszintű szabályoknak, amely kimenete adja a NORMO végső kimenetét.

3.1. Memóriaalapú normalizálás

A NORMO memóriaalapú moduljának az alapja egy szótár, amely a leggyakoribb szóalakokat és azok kézzel normalizált változatát tartalmazza egy `tsv` fájlban. A memóriaalapú modul legnagyobb előnye a pontosság, hiszen a szótárban szereplő összes szóalak kézzel lett normalizálva. Az automatikus normalizálás során azok a szóalakok, amelyek a szótárban is szerepelnek, egyszerűen lecserélődnek a szótárban szereplő megfelelőjükre. A szótár alapján normalizált szóalakoknál további javításra nincs szükség, így a további normalizáló eljárások (a karakter- és tokenszintű szabályok) számára stopszóként viselkednek.

Mivel a memóriaszótár kézzel készült, a lehető legmagasabb pontosságot várjuk el. A fedés viszont a szótár méretétől függ. A jelen cikkben bemutatott eredményekhez a Károli-biblia négy evangéliumának és az Újtestamentum bevezetésének a leggyakoribb szavaiból készítettük el a szótárt. Az automatikus normalizálás kézzel ellenőrzött kimenete a későbbiekben tovább növelheti a szótár méretét – és így a fedést is. A kiértékeléskor a szótár 502 szóalakot tartalmazott.

A memóriaalapú normalizálás hátránya, hogy a normalizálandó szóalak és a szótárban szereplő szóalak között teljes karakterszintű egyezés szükséges. Vagyis ha egy betűhű szó egyszer nagybetűvel, egyszer kisbetűvel szerepel a szövegben, de a szótárban csak nagybetűvel van, akkor a kisbetűseket elveszítjük. Ezt megelőzendő, a memóriaalapú normalizálás előtt kisbetűsítjük a szöveget. A memóriaalapú normalizálás alkalmazása után csak azok a szavak lesznek nagybetűsek, amelyeknek a normalizált megfelelője nagybetűs a szótárban, tehát a tulajdonnevek.

3.2. Szabályalapú normalizálás

A NORMO szabályalapú modulja kézzel írt újraíró szabályokat tartalmaz. A szabályok két forrásból származnak: egyrészt nyelvtörténeti kutatások eredményeképpen, másrészt korpuszalapú megfigyelések alapján fogalmaztuk meg őket. Az

újraíró szabályok a lecserélendő karaktersorozat szomszédos karakterei vagy a szóhatárok alapján működnek, tehát környezetfüggők. Illeszkedés esetén a környezetfüggő szabály alapján az eredeti karaktersorozatot lecseréljük a normalizált karaktersorozatra. Ennek megfelelően a szabályok alkalmazása által eredményezett szóalak nem feltétlenül egy valid magyar szó.

Karakterszintű szabályok. A karakterszintű újraíró szabályokat egy tokenen belül alkalmazzuk. A szabályok között vannak olyanok, amelyek a középmagyar és a mai magyar karakterkészlet különbségeit oldják fel úgy, hogy a mai magyar karakterkészlet részét nem képező karaktereket (pl. *Æ* és *ę*) a mai magyarban megtalálható megfelelőjükre fordítja át (*e* és *é*). Ennek megfelelően a normalizált alakok már nem tartalmaznak olyan karaktereket, amelyek nem képezik részét a mai magyar karakterkészletnek. Továbbá olyan szabályok is vannak, amelyek a mai magyar helyesírásban tükröződő fonotaktikai szabályszerűségeket fedik le. Például szóvégen csak hosszú *ó* vagy *ő* szerepelhet, a rövid párjuk nem, így a karakterszintű szabályok segítségével minden szóvégi *o* és *ö* karaktert rendre *ó* és *ő* karakterre cserélünk.

A karakterszintű szabályok esetén nincs szükség teljes karakteregyezésre egy szóalakban, hiszen ezek a szabályok a szóalakon belül operálnak. Ennek megfelelően egy tő különböző végződésű változataiban ugyanúgy végbemennek a karakterszintű szabályok (ellentétben a 3.1. fejezetben említett memóriaalapú normalizálással). Ugyanakkor a karakterek együttes előfordulásai különböző viselkedést mutatnak annak megfelelően, hogy egy morfémán belül vagy morfémahatáron vizsgáljuk őket. A magyar nyelv agglutináló jellege miatt ezzel a problémával gyakran kell szembesülnünk. Például a morfémán belüli intervokális környezetben szereplő i-t j-re kell fordítanunk, morfémahatáron viszont i-nek kellene maradnia. A Normo nem kezeli különbözőképpen a morfémán belüli és a morfémahatáron álló karaktercsoportokat, ezért ezekben az esetekben számba kell vennünk a szabály alkalmazásával létrejövő helyes és helytelen szóalakok számát. Ha a szabály több helyes szóalakot hoz létre, mint helytelent, akkor a szabályt alkalmazzuk, majd a helytelen alakok javítását a kézi annotátorokra bízzuk.

A 88 darab karakterszintű újraíró szabályt reguláris kifejezések formájában implementáltuk. Mivel egy szóalakon belül több újraíró szabályt is alkalmazhatunk, valamint egy újraíró szabály kimenete bemenete lehet egy másiknak, ezért a reguláris kifejezések sorrendje nagyon fontos. Például a [ttz → tsz] és a [tz → c] szabályok kimenete az alkalmazásuk sorrendjétől függ. Ezért minden egyes szabály megfogalmazásakor számba kellett vennünk a korábban megfogalmazott szabályok be- és kimeneteit.

Tokenszintű szabályok. A tokenszintű szabályokat a fent ismertetett memóriaalapú és a karakterszintű normalizáló modulok kimenetén alkamazzuk. Ezek a szabályok a szóhatárokon keresztül operálnak, így módosítják a normalizált szöveg tokenszámát. A 7 darab tokenszintű szabály közül egy szóalakra csak egyet

alkalmazunk, szemben a 3.2. fejezetben ismertetett karakterszintű szabályokkal, amelyek közül akárhány alkalmazható egy szóalakon belül.

A tokenizálást befolyásoló szabályok között megfogalmaztunk egészen általánosakat is (pl. a felsőfokot jelölő *-leg* prefix és az azt követő szó összevonása), de vannak közöttük egyedi, de gyakori eseteket kezelő szabályok is (a *szent* szóalak és a *lélek* tövű szó egybeírása). A NORMO tokenizálást befolyásoló szabályait az 1. táblázat foglalja össze.

összevonás		betűhű	normalizált
	igekötő + ige	***le*** *megy* →	***le****megy*
	mutató nm. + kérdő nm.	***az ki*** →	***aki***
	személyes nm. + névutó	***én utánam*** →	***énutánam***
	felsőfok + melléknév	***leg*** *jobb* →	***leg****jobb*
	szent + lélek	***szent lélek*** →	***szentlélek***
szétszedés			
	szó + *is*	*porát****is*** →	*porát* ***is***
	szó + *nélkül*	*orvos****nélkül*** →	*orvos* ***nélkül***

1. táblázat. A szavak tokenizálását befolyásoló szabályok.

A karakterszintű szabályokhoz hasonlóan a tokenszintű szabályok alkalmazása sem mindig helyes mai magyar szót eredményez. A tokenszintű szabályok számára problémát okozhat a homográfia, például a *meg* szó esetében, amely lehet igekötő vagy kötőszó. Míg igekötőként az őt követő igével egybeírandó, kötőszóként külön tokennek kellene maradnia. A homográfiából fakadó hibák javítását a kézi annotátorokra bízzuk.

A fentieken kívül vannak olyan tokenszintű szabályok is, amelyek a mondatra bontást befolyásolják. Ilyen szabály például az, amelyik a *hogy* kötőszó és a vonatkozó névmások elé vesszőt és tagmondatot jelző üressort illeszt. Jelen cikkben csak a szószintű tokenizálást végző szabályok teljesítményét értékeljük ki.

4. Kiértékelés

A NORMO teljesítményének kiértékeléséhez a Károli-biblia négy evangéliumát és az Apostolok cselekedeteit használtuk. A szövegek kézzel normalizált változatát vettük gold standardnak, azzal vetettük össze az automatikus normalizáló kimenetét. A szövegek tokenszáma a betűhű változat alapján 114 580 token, a normalizált változat alapján 109 289 token.

A kiértékeléshez két mérőszámot használtunk. A *normalizálási pontosság* mutatja a gold standardnak megfelelően normalizált tokenek arányát, így a tulajdonképpeni szószintű normalizálás teljesítményét. Ez a mérőszám a NORMO

moduljai közül a memóriaalapú modul és a szabályalapú modul karakterszintű szabályainak a teljesítményét méri.

A normalizálási pontosság kiszámolásakor figyelembe kellett vennünk azokat a szavakat is, amelyeket a normalizálás során akár a kézi annotátor, akár a Normo tokenszintű szabályai szétválasztottak vagy összevontak. Az eredeti betűhű szöveggel összevetettük a kiértékeléskor gold standardként használt kézzel normalizált szöveget és a Normo kimenetét is, így a szétválasztott vagy összevont szavak normalizálási pontosságát is ki tudtuk számolni.

A karakterszintű szabályok egyik rendszerben sem tudják jól kezelni a tulajdonneveket, mivel azok nem feltétlenül követik az adott nyelv fonológiai és helyesírási szabályait. Mivel a szöveg, amivel dolgozunk, bibliafordítás, nagyon sok bibliai nevet találunk bennük. A kézi normalizálás során a különböző bibliafordításokban és bibliai históriákban említett tulajdonnevek is normalizálásra kerültek, vagyis az adott nevek különbözőképpen használt alakjai a Szent István Társulati bibliafordításnak[2] megfelelően egységes formára lettek hozva. A hibák nagy részét a nevek adják – kezelésük még nem megoldott a Normo-ban.

A szabályalapú modul tokenszintű szabályainak, tehát a tokenizálást befolyásoló szabályoknak a teljesítményét egy másik mérőszám fejezi ki. Ennek kiszámolásához három kategóriát állapítottunk meg: a) álpozitív: a Normo összevont két tokent, amelyek a gold standardban nincsenek összevonva, vagy szétválasztott egy tokent, amely a gold standardban nincs szétválasztva; b) álnegatív: a Normo nem vont össze két tokent, amelyek a gold standardban össze vannak vonva, vagy nem választott szét egy tokent, amely a gold standardban szét van választva; és c) valós pozitív: a Normo összevont két tokent, amelyek a gold standardban is össze vannak vonva, vagy szétválasztott két tokent, amely a gold standardban is szét van választva. Ezeknek a hibafajtáknak a számosságát ismerve kiszámoltuk a pontosságot, a fedést és az F-mértéket, amelynek eredményét a normalizálási pontossággal együtt a 2. táblázat mutatja.

	tokenszintű pontosság (%)	tokenszint fölötti pontosság (%)	fedés (%)	F-mérték (%)
Máté	81,04	87,90	69,40	77,56
Márk	80,62	86,18	65,80	74,62
Lukács	81,58	84,56	67,50	75,07
János	83,10	94,20	70,60	77,03
ApCsel	79,95	90,72	67,13	77,16
átlag	81,23	88,63	68,16	77,06

2. táblázat. A Normo kiértékelése a Károli-biblia egyes részein.

[2] http://szentiras.hu/SZIT

A kiértékeléskor a szótárunk 502 szóalakot tartalmazott. Feltételezzük, hogy minél nagyobb a memóriaalapú normalizáláshoz használt szótár, annál magasabb normalizálási pontosságot érhetünk el – ugyanakkor minimalizálni szerettük volna a szótár készítéséhez használt manuális munkát. Leteszteltük, hogy különböző méretű – ötvenesével növelt – szótárak használata mellett a NORMO memóriaalapú modulja önmagában milyen teljesítményt ér el, vagyis ennél a kiértékelésnél a szabályalapú modult nem működtettük. Az eredményeket az 1. ábra mutatja.

1. ábra: Az ötvenesével növelt szótárméret mellett kapott normalizálási pontosság a NORMO memóriaalapú moduljának kiértékelésekor. A vízszintes tengelyen a szótár mérete, a függőleges tengelyen a normalizálási pontosság látható.

Az eredmények alapján látható, hogy a NORMO 502 szóalakot tartalmazó szótára önmagában is elég magas normalizálási pontosságot ér el – ami még kiegészül a szabályalapú modulok teljesítményével is. Mivel a szótárba az egyes szóalakok gyakoriságuk alapján kerültek bele (tehát az 502 leggyakoribb szó), ezért a szótár további növelése nem okozna nagy javulást a normalizálási pontosságban – a szótárméret növelésével javuló normalizálási pontosság görbéje kilapulna.

A kimenetet emellett a NORMA automatikus normalizáló eszköz [12] eredményével is összevetettük. A NORMA is szabályalapú, de abban különbözik a NORMO-tól, hogy a karakterszintű újraíró szabályokat a tanítóanyagként megkapott, kézzel normalizált szövegből egy módosított Levenshtein-algoritmus segítségével tanulja ki. A tanítóanyag megegyezett a NORMO memóriaalapú moduljában használt szótárral. A NORMA a tanítóanyag mellett egy célnyelvi szólistát is használ, amely a Szent István Társulati bibliafordítás szavait tartalmazó szólista. A karakterszintű szabályok hatékonyságának összevetéséhez a NORMO esetében csak a szabályalapú modult működtettük, a memóriát nem használtuk. Az eredményeket a 3. táblázat tartalmazza.

	pontosság	
	Normo (%)	Norma (%)
Máté	68,29	61,58
Márk	66,43	61,70
Lukács	67,03	61,31
János	68,59	59,14
ApCsel	66,98	60,18
átlag	67,36	60,67

3. táblázat. A Normo és a Norma rendszerek normalizálási pontosságának összehasonlítása.

Az eredmények azt mutatják, hogy ekkora tanulóanyag rendelkezésre állása esetén a nyelvtörténeti tudás hozzáadása határozottan javít egy szabályalapú rendszer teljesítményén.

Egy olyan módszer kiértékelésének, amely a kézi munkát hivatott automatikusan segíteni, egy másik lehetséges dimenziója az, ha az adott feladatra szánt kézi munka mennyiségét és az automatikus módszerrel megtámogatott munka mennyiségét vetjük össze. Ezt azt összehasonlítást tartalmazza a 4. táblázat.

	tokenszám (#)	becsült idő (h)	tényleges idő (h)	arány (%)
Máté	28 520	285,2	400	140,25
Márk	18 150	181,5	116	63,91
Lukács	30 805	308,05	96	31,16
János	23 435	234,35	80	34,13
ApCsel	28 631	286,31	100	34,92

4. táblázat. A teljesen kézi normalizálás és a gépi normalizálás utáni kézi javítás időigényének összevetése.

A kiértékelés alapjául a betűhű szöveg tokenszámát vettük, mert mind a kézi, mind a gépi normalizálásnak az a bemenete. A becsült idő az Ómagyar Korpusz annotálásának eddigi tapasztalataiból leszűrt idő, ami 100 token óránként – ezt tekintjük annak az időtartamnak, ameddig akkor tartott volna a normalizálás, ha teljes mértékben kézzel zajlott volna. A tényleges idő azt mutatja, hogy mennyi ideig tartott a manuális javítási munka, amikor az annotátorok a Normo által normalizált szöveget kapták meg bemenetként. Az arányszámok pedig a tényleges és a becsült idő arányát mutatják. A számokból azt láthatjuk, hogy a Máté evangéliumának normalizálása igényelte a legtöbb időt, és az erre kapott arányszám a legnagyobb. Ez egyrészt annak köszönhető, hogy ennek a

résznek az automatikus normalizálása a NORMO fejlesztésének egy korábbi szakaszában zajlott, így a kimenetben sokkal több kijavítandó hiba volt. Másrészt pedig annak, hogy a korábbi gyakorlattal ellentétben a NORMO kimenetében minden szónak van normalizált megfelelője – ha nem történik a bemenő szón semmi változás, akkor önmaga –, és ez nehezebben kezelhető volt az annotátor számára, mintha nem lett volna ott semmi. A NORMO fejlesztése során az egyre jobb minőségű kimenet egyre inkább megkönnyítette a kézi annotátorok dolgát, ami jól kirajzolódik a 2. ábrán látható diagramon.

2. ábra: A normalizáláshoz felhasznált idő változása az automatikus normalizálás bevezetésével. A vízszintes tengelyen a szövegek, a függőleges tengelyen a normalizálásuk ellenőrzéséhez felhasznált munkaidő látható órában kifejezve.

5. Összefoglalás

Cikkünkben egy automatikus normalizáló eszközt mutattunk be, amely egy memóriaalapú és egy szabályalapú modulból áll. A NORMO-val normalizált és a kézi annotátorok által ellenőrzött középmagyar szöveg megfelelő bemenet lesz a további nyelvfeldolgozó eszközök számára. A NORMO magas teljesítménye, valamint a gépi normalizálás utáni kézi javítás alacsony időigénye jól mutatja, hogy a NORMO használata megkönnyíti és meggyorsítja a normalizálást.

A fejlesztés jelenlegi fázisában az automatikus normalizálót a Károli-bibliára adaptáltuk, de a szótár és a szabályok módosításával további középmagyar szövegekre is alkalmazható. Ezért a későbbiekben szeretnénk más szövegekre (például a többi középmagyar bibliára) is alkalmazni az eszközt, valamint újabb normalizálási eljárásokat is (pl. Levenshtein-alapú normalizálást, gépi tanulási megoldásokat) kipróbálni és kiértékelni.

Hivatkozások

1. Kroch, A., Taylor, A.: The Penn-Helsinki Parsed Corpus of Middle English (PPCME2) (2000) URL: http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4.
2. Galves, C., Britto, H.: The Tycho Brahe Corpus of Historical Portuguese (2010) URL: http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html.
3. Simon, E.: Corpus building from Old Hungarian codices. In É. Kiss, K., ed.: The Evolution of Functional Left Perpiheries in Hungarian Syntax. Oxford University Press (2014) 224–236
4. McEnery, T., Hardie, A.: Lancester Newsbooks Corpus. (2003) URL: http://www.lancs.ac.uk/fass/projects/newsbooks/default.htm.
5. Novák, A., Gugán, K., Varga, M., Dömötör, A.: Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. Language Resources and Evaluation (2017)
6. Rayson, P., Archer, D., Baron, A., Culpeper, J., Smith, N.: Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In: Proceedings of the Corpus Linguistics Conference (CL2007), UK (2007)
7. Bollmann, M., Søgaard, A.: Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In: Proceedings of COLING 2016: Technical Papers, Osaka, Japan (2016) 131–139
8. Pettersson, E.: Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. PhD thesis, Uppsala University, Department of Linguistics and Philology (2016)
9. Brill, E., Moore, R.C.: An improved error model for noisy channel spelling correction. In: Proceedings of ACL 2000. (2000)
10. Oravecz, Cs., Sass, B., Simon, E.: Semi-automatic Normalization of Old Hungarian Codices. In: Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010), Lisbon, Portugal (2010) 55–60
11. Bollmann, M., Petran, F., Dipper, S.: Rule-based normalization of historical texts. In: Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria (2011) 34–42
12. Bollmann, M.: (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool. In: Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), Lisbon, Portugal (2012) 3–14

Hyphenation using deep neural networks

Gergely Dániel Németh and Judit Ács
neged.ng@gmail.com, judit@aut.bme.hu

Budapest University of Technology and Economics
Department of Automation and Applied Informatics

Abstract. Hyphenation algorithms are the computer based ways of syllabification and mostly used in typesetting, formatting documents as well as text-to-speech and speech recognition systems. We present a deep learning approach to automatic hyphenation of Hungarian text. Our experiments compare feed forward, recurrenct and convolutional neural network approaches.
Keywords: hyphenation, machine learning, feedforward neural network, recurrent neural network, long short-term memory, convolutional neural network

1 Introduction

Hungarian children learn the rules of syllabification in their early teens. Hyphenation rules are clearly defined in [1], and after years of practice most people use it naturally and this is seen as part of the common knowledge.

In Hungarian hyphenation depends mostly on the word itself and less so on context. Polysemous words may have different hyphenations in different senses which can only be derived from the context as shown in the example *me-gint* (again) and *meg-int* (warn). Although it is a known issue, most hyphenation algorithms hyphenate words with no regard to context and assign only one possible hyphenation for a word.

Commonly used hyphenation algorithms are based on the methods defined in the first version of TEX [2]. It is a pattern based hyphenation algorithm with thousands of manually chosen patterns. It's employed in the online syllabification system[1] created and maintained by RIL HAS[2]. [3] describes the collection of patterns used in this system.

While we have enormous amount of words in corpora, we lack a gold standard corpus of pre-hyphenated words. One way to create hyphenated words is to use already available hyphenation algorithms.

2 Related Work

The world of open-source software *de facto* uses the TEX's hyphenation algorithm. The TEXcurrently uses the Hunspell's hyphenation algorithm which based on Liang's pattern matching algorithm.

[1] http://helyesiras.mta.hu/helyesiras/default/hyph
[2] Research Institute for Linguistics of the Hungarian Academy of Sciences

2.1 Liang's algorithm

The basic concepts of Liang's algorithm are the hyphenation patterns. The process of hyphenating is the following: the algorithm finds all the matching patterns. These patterns have numbers in them (skipped while matching). The odd numbers predict hyphens. The method chooses the maximum of the numbers between each letters (zero if not given).

The word *hyphenation*'s patterns are the following: `hy3ph, he2n, hena4, hen5at, 1na, n2at, 1tio, 2io` [4, page 37]. Placing the patterns in the right position and inserting the numbers in the patterns between the letters we got Figure 1.[3]

```
. h y p h e n a t i o n .
  h y3p h
        h e2n
        h e n a4
        h e n5a t
           1n a
            n2a t
               1t i o
                 2i o
.0h0y3p0h0e2n5a4t2i0o0n0.
  h y-p h e n-a t i o n
```

Fig. 1: The hyphenation of 'hyphenation' by Liang's algorithm

2.2 Hunspell

Hunspell's hyphenation algorithm is currently used in TEX and OpenOffice. It is based on Liang's work with Sojka's non-standard hyphenation extensions [6] and was published by Németh in 2006 [5]. Hunspell supports non-standard hyphenation patterns such as the hyphenation of the German word *Zucker* – *Zuck-ker* (the exact pattern is `c1k/k=k`.

Different languages use different non-standard patterns whose sizes and types vary significantly.

2.3 Errors in the Hunspell's hyphenation

Most of the hyphenation errors come from the fact that the Hunspell's creators wanted to create a typesetting algorithm so they decided to not hyphenate one letter long syllables the begining and the ending of the words. However, when it comes to compound words, these one-letter parts can be in the middle of the word. There are some examples of hyphenation errors in Table 1.

[3] The visualization method comes from Németh's article [5].

Hyphenation by Hunspell:	Correct hyphenation:	Error type:
au-tó-val	a-u-tó-val	one-letter
szem-üveg-gel	szem-ü-veg-gel	one-letter
has-izom	has-i-zom	one-letter
messze	mesz-sze	no hyphen
fölül	föl-ül	no hyphen
top-ikok	to-pi-kok	no hyphen[4]
vi-deó	vi-de-ó	one-letter
geo-dé-zia	ge-o-dé-zi-a	no hyphen
diszk-ri-mi-na-tív	disz-kri-mi-na-tív	wrong place

Table 1. Hyphenation errors in Hunspell

2.4 Neural networks for hyphenation

[7] is the only paper we found specifically on automatic hyphenation using neural networks. Due to technical constraints they only tested on a very small training dataset. Our first approach (feedforward neural net) is very similar to the one they present.

3 Neural networks

The following summary of neural networks is based on the *Deep learning in neural networks: An overview* by Schmidhuber [8].

3.1 Feedforward neural network

A feedforward neural network approximates any given function f as $y = f(x, T)$ where T represents those parameters with which the model can learn to achieve the best approximation. These networks are called feedforward because the information flows through the function from x to inner (hidden) parts and finally to y. There are no directed cycles or loops in the network.

The simplest model is a single-layer perceptron which has a weight W and bias b, so for an input x can compute the $\hat{y} = Wx+b$ function where the learning method optimizes the weight W and the bias b. Later on researchers showed that adding a non-linear *activation function* to it can fasten the learning method and makes it usable in non-linear functions [9]. So from the $\hat{y} = Wx + b$ the function changed to $z = Wx + b$ and the prediction became $\hat{y} = g(z)$.

The deep neural network's name comes from that instead of a single-layer perceptron the output of the above equation $g(z)$ now called as h_1 is used as the input of the next layer and so for many-many layers. So for the first layer it became $z_1 = W_1x + b_1$ and $h_1 = g_1(z_1)$ and to the second layer: $z_2 = W_2h_1 + b_2$ and $h_2 = g_2(z_2)$ and so on, until the last, nth layer where the a_n became the prediction $h_n = \hat{y}$. Figure 2. summarizes idea of feedforward networks.

[4] This word is hyphenated incorrectly by the online hyphenation tool of the Research Institute for Linguistics.

Fig. 2: Basics of Feedforward Neural Networks (F(F)NN)

A single iteration of training consists of a forward step where the model predicts $\hat{y}$, the evaluation step where the model compare the $\hat{y}$ and y with some type of gradient descent [10] and lastly a backpropagation step where the model updates the weights [11].

3.2 Convolutional neural network

Convolutional neural networks were introduced to solve image recognition problems [12]. A convolutional neural network has a filter (or kernel) which is sliding around the input (image) and multiplying the values in the filter (the weights of the filter) with the original input values (pixels). Summing up these values the network get a single number for every position of the filter. This will be the output of the layer. The size of the output depends on the filter size and the parameter *strides* which defines the steps of the filter's sliding.

Let a_{ij} be the cell (pixel) of the input (image) in the ith row and jth column and f_{ij} the cell of the filter, while h_{ij} the output. Thus the first cell of convolutional layer's output is

$$h_{11} = \sum_{x=1..k, y=1..l} a_{xy} f_{xy},$$

(where k and l are the height and width of the filter respectively), and assuming that the stride is 1, the h_{ij} is:

$$h_{ij} = \sum_{x=i..(i+k), y=j..(j+l)} a_{xy} f_{(x-i+1),(y-j+1)}.$$

Figure 3. illustrates a convolutional network with a $(3, 3)$ kernel.

Kim, Jernite, Sontag and Rush showed a way of using 1 dimensional convolutional neural networks and LSTM networks in character sequences [13], where

Fig. 3: Basics of Convolutional Neural Network (CNN)

the filter size says that how many character should be included into the convolution.

3.3 Recurrent neural network - LSTM

A recurrent neural network (RNN) is suited for modelling sequential phenomena. At each time step t, an RNN takes the input vector $x_t \in \Re^n$ and the hidden state vector $h_{t-1} \in \Re^m$ and produces the next hidden state h_t by applying the following recursive operation: $h_t = f(Wx_t + Uh_{t?1} + b)$. In theory, an RNN can store all information in h_t, however learning long-range dependences with it is difficult due to vanishing/exploding gradients [14].

Long short-term memory (LSTM) [15] addresses the problem of learning long range dependencies by augmenting the RNN with a memory cell vector $c_t \in \Re^n$ at each time step. Concretely, one step of an LSTM takes as input x_t , $h_{t?1}$, $c_{t?1}$ and produces h_t , c_t via the following intermediate calculations:

$$
\begin{aligned}
i_t &= \sigma(W^i x_t + U^i h_{t-1} + b_i) \\
f_t &= \sigma(W^f x_t + U^f h_{t-1} + b_f) \\
o_t &= \sigma(W^o x_t + U^o h_{t-1} + b_o) \\
g_t &= \tanh(W^g x_t + U^g h_{t-1} + b_g) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

Here $\sigma(\cdot)$ and $\tanh(\cdot)$ are the element-wise sigmoid and hyperbolic tangent functions, $\odot$ is the element-wise multiplication operator, and i_t, f_t, o_t are referred to as *input*, *forget*, and *output* gates. At $t = 1$, h_0 and c_0 are initialized to zero vectors. Parameters of the LSTM are W^j , U^j, b^j for $j \in i, f, o, g$. See the visualisation of an LSTM unit in Figure 4.[5]

Bidirectional recurrent neural networks are based on the principle to split the neurons of a regular RNN into two directions, one for positive time direction (forward states), and another for negative time direction (backward states) [16]. In terms of characters in a word it means that the letters can affect their surroundings on both sides.

[5] Visualisation and more:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Fig. 4: Basics of Long short-term memory (LSTM)

4 Data and preprocessing

Hungarian Webcorpus is one of the largest Hungarian language corpora with over 600 million words and it is available in its entirety under a permissive Open Content license [17,18]. We used the 100000 most frequent words from the corpus.

4.1 Data preprocessing

The preparation methods used before the training:

Cleaning Numbers, punctuation characters are cleaned, setting all letters to lower-case.

Filtering non-standard hyphenation Because the models use character classification, they cannot handle letter addition or changing. The preprocessing method filters them however the models may find the position of the hyphenation.

Filtering special characters The models only accept the Hungarian alphabet and the padding characters.

Long word cut The CNN and LSTM networks wait for given length words so the longer words are dropped and the shorter ones are filled with padding characters.

Table 2. illustrates the preprocessing steps and the amount of dropped words.

	Origin	Cleaning	Non-standard	Special chars	Long words	Final
Words	100000	16322	1115	646	7	81910
% of the origin	100	16.32	1.11	0.65	0.01	81.91
% of the previous	-	16.32	1.33	0.78	0.01	
Words after	-	83678	82563	81917	81910	

Table 2. Data preprocessing

5 Experimental setup

We implemented and tested three neural network architectures described in the previous Section.

5.1 Character classification

Automatic hyphenation can be defined as a sequence labeling problem. We assign a binary label for each character:

- B: In the beginning of the syllables.
- M: Every other letter.

The word *leopárd* (leopard) hyphenated as `le-o-párd` and tagged as `BMBBMMM`.

There are more fine-grained labeling scemes such as the `BMES`, which denotes the middle of segments as well as single-character segments, they require a more complicated inference scheme so we decided to use the binary labeling system.

5.2 Feedforward neural network

In the feedforward neural network we want to decide to which class a specific character belongs. To do it we use the character and its surroundings coded in one-hot as the input layer of a fully connected network. The output is the class.

The network is summarized in Figure 5. The steps of the character classification.

Fig. 5: Feedforward neural network

Learning from the surroundings (Windowing) In an n-length window we use $(n-1)/2$ characters before the given one, itself and $(n-1)/2$ characters after. For example, if the character is the letter l in the word *leopárd*, and the window-length is 5, the window will be `^^LEO`.

One-hot encoding We used one-hot encoding to represent characters as vectors. We define a 37 length array for each letter in the word: each element of the array means one letter of the Hungarian characters (35 letters) or the beginning or ending characters. One-hot means that there is only one 1 in the array, the others are 0. For the letter a it is the first one and the character o it is the 18th.

Thus if there are 5 characters in a window, a $(5, 37)$ shaped two-dimensional array is given.

Flattening The final step before the training is flattening. From the two-dimensional array we create a reshaped one-dimensional. This means simply putting the characters after each other. So if we had the $(5, 37)$ array, now we have the $5 \cdot 37 = 185$ long 0,1 sequence with only 5 ones in it. And this is the training data.

In summary, from the *leopá* window we got a 185 long $0, 1$ as the training input and $[1, 0]$ as the training output.

The neural network The neural network is a fully connected feedforward neural network. In the hyperparameter optimization along the window length, the number of hidden layers and the units in each hidden layer had been optimized. The last layer has a softmax activation to approximate the values as probabilities of **B** and **M** tags.

The training process used *sigmoid* activation functions and *adam* optimizer [19].

5.3 Convolutional neural network

For the CNN and LSTM networks the data preparation steps are the same. First, using the Hunspell hyphenation to define the labels, then filling the words to a fixed size with padding (the filling label is **M**) and finally using the one-hot encoding method (Figure 6.).

Fig. 6: CNN and LSTM data preparation

The model has two parts. The first one is the convolutional network and the second is a feedforward network. The convolutional layers have a stride one so it convolves all the characters. The kernel size (it is the same dimension as the window-length in the FFNN model), filter and hidden layer numbers were optimized. The Keras' built-in padding was used to prevent problems at the endings of the word. The model currently uses ReLU activation functions.

The feedforward part is a softmax layer for every character to get back the **BM** one-hot probabilities.

Figure 7. summarizes the CNN model.

Fig. 7: CNN model summary

5.4 Long short-term memory

The LSTM network uses the same inputs as the CNN. The model uses BiLSTM and optimized by the unit and hidden layer numbers. At the output of the LSTM a softmax feedforward layer was inserted just as in the CNN one. The model summary of the LSTM network is in Figure 8.

5.5 Hyperparameter optimization

Hyperparameters were optimized via random search. Parameter values were uniformly sampled from the ranged listed in Table 3..

5.6 Results and error analysis

The best results are listed in Table 4., where the precision, recall and F-score values are for the **B** tagging while the word accuracy is the rate of successfully hyphenated words.

Fig. 8: Long Short-Term Memory network

	Window/Kernel	Hidden layers	Hidden units
FFNN	3-11	3-7	60-180 (step 10)
CNN	5-16	1-3	64-2048 (exponential scale)
LSTM	-	1-3	8-256 (exponential scale)

Table 3. Hyperparameter ranges

Model	Kernel	Layers	Hidden	Epochs	Trainable parameters	Precision	Recall	F-score	Word accuracy
FFNN	7	3	150	103	84,602	98.17%	99.11%	98.64%	93.57%
CNN	8	2	1024	12	8,695,810	98.37%	99.27%	98.81%	94.45%
LSTM	-	2	128	66	216,834	97.68%	99.16%	98.42%	93.13%

Table 4. The three model that performed the best.

All the errors of the three models were collected and 200 words were manually classified into the error categories. Figure 9. shows the category distribution of the errors for each model and among the errors. As we can see, over half of the errors are caused by using a Hungarian hyphenation algorithm on a non-Hungarian word.

Fig. 9: Category distribution in the 200 wrong words

We manually checked and grouped the both correctly and incorrectly hyphenated words into the following categories:

Non-hungarian word is which was recognized as a word but not Hungarian.
Compound word
Non-hyphenated part has a hyphen missing because of Hunspell's typesetting goals.
Wrong target Hunspell misses the hyphenation.
Not a words are mostly mistyped Hungarian words like *elol*.
Other words not filling the above categories.

Table 5. shows category distribution among all the words (correctly predicted words as well as incorrect ones). Note that one word can be in multiple categories.

6 Conclusion

In this paper we introduced a series of experiments on Hungarian hyphenation using deep neural networks. We compared three architecutres with varying hy-

Category	Distribution	Example (Hunspell's hyphenation)
Non-Hungarian word	11	obsta-c-les
Compound word	21	ak-ció-film
Non-hyphenated part	8	ak-ció-film
Wrong target	1	diszk-ri-mi-na-tív
Not word	4	el-er-ni
Others	59	

Table 5. Word categories among 100 randomly chosen words

perparameters: feedforward, recurrent and convolutional neural networks. We trained and tested on the 100000 most frequent words from the Hungarian Webcorpus. All models achieve over 95% word accuracy, occasionally correcting the errors made by Hunspell. Our error analysis suggests that foreign and compound words are the most challenging for our systems. Our source code and the best models are available on GitHub.[6]

References

1. Akadémia, M.T., Budapešt', M.: A magyar helyesírás szabályai. Akadémiai kiadó (1959)
2. Knuth, D.E.: TEX and METAFONT: New directions in typesetting. American Mathematical Society (1979)
3. Miháltz, M., Hussami, P., Ludányi, Z., Mittelholtz, I., Nagy, Á., Oravecz, C., Pintér, T., Takács, D.: Helyesírás. hu. (2013)
4. Liang, F.M.: Word hyphenation by computer. Department of Computer Science, Stanford University (1983)
5. Németh, L.: Automatic non-standard hyphenation in openoffice. org. COMMUNICATIONS OF THE TEX USERS GROUP TUGBOAT EDITOR BARBARA BEETON PROCEEDINGS EDITOR KARL BERRY (2006) 32
6. Sojka, P., et al.: Notes on compound word hyphenation in tex. TUGboat **16**(3) (1995) 290–296
7. Kristensen, T.: A neural network approach to hyphenating norwegian. In: Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on. Volume 2., IEEE (2000) 148–153
8. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61** (2015) 85–117
9. Fahlman, S.E.: An empirical study of learning speed in back-propagation networks. (1988)
10. Hadamard, J.: Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées. Volume 33. Imprimerie nationale (1908)
11. Hecht-Nielsen, R., et al.: Theory of the backpropagation neural network. Neural Networks **1**(Supplement-1) (1988) 445–448
12. Fukushima, K., Miyake, S.: Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets. Springer (1982) 267–285

[6] https://github.com/negedng/deep-hyphen

13. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In: AAAI. (2016) 2741–2749
14. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks **5**(2) (1994) 157–166
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780
16. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11) (1997) 2673–2681
17. Halácsy, P., Kornai, A., Laszlo, N., Andras, R., Szakadát, I., Viktor, T.: Creating open language resources for hungarian. (2004)
18. Kornai, A., Halácsy, P., Nagy, V., Oravecz, C., Trón, V., Varga, D.: Web-based frequency dictionaries for medium density languages. In: Proceedings of the 2nd International Workshop on Web as Corpus, Association for Computational Linguistics (2006) 1–8
19. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

Lemmi vagy nem lemmi

Novák Attila, Novák Borbála

Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar
MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Budapest, Práter u. 50/a.
{novak.attila, siklosi.borbala}@itk.ppke.hu

Kivonat A szövegfeldolgozással foglalkozó kutatások nagy része ma már a magasabb szintű elemzést megvalósító algoritmusok fejlesztésével foglalkozik, az alacsony szintű, leginkább az előfeldolgozás során használt eszközöket azok hiányosságaival együtt késznek, elfogadhatónak tekintve. Azonban ezeknek az eszközöknek a pontossága sok esetben meg sem közelíti a tökéletes eredményt, az ezek által az eszközök által bevezetett hibák így továbbterjednek a feldolgozási lánc magasabb szintű moduljaiba. Jelen kutatásban célunk egy olyan algoritmus létrehozása volt, ami az egyik ilyen alapeszköz, a lemmatizáló pontosságának javítására használható. A létrehozott rendszer igen magas pontossággal (92,13%) képes megítélni, hogy egy szóalakhoz automatikusan létrejött lemma az eredeti szóalakhoz rendelt szófajcímkével együtt helyes-e vagy sem.
Kulcsszavak: lemmatizálás, morfológia, hibajavítás, neurális hálózat

1. Bevezetés

A legtöbb automatikus szövegfeldolgozó lánc önálló modulok egymás után való alkalmazásából áll. Az egyes modulok kimenete az utána következő bemenete lesz. Egy ilyen architektúrában a feldolgozás elején keletkező hibák a rendszerben tovább terjednek, a magasabb szintű feldolgozást végző modulok a saját hibaarányukon túl a megelőző moduloktól kapott hibás bemenetre is hibás választ fognak adni. Ezért nagyon fontos, hogy a sztenderd alapfeldolgozást végző eszközök a lehető legnagyobb pontossággal működjenek.

Cikkünkben egy ilyen eszköz minőségének javítására teszünk javaslatot: a szófaji egyértelműsítéssel párhuzamosan történő lemmatizáló guesser által létrehozott lemmák helyességét értékelő algoritmust mutatunk be, melynek használatával jelentősen csökkenthető az automatikus elemzés során létrejövő hibás lemmák száma.

Kutatásunk során a PurePos [1] szófaji egyértelműsítő lemmatizáló algoritmusát vizsgáltuk (felmerült más lemmatizálók tesztelése is, ám ezek teljesítménye ránézésre is sokkal rosszabb volt a PurePosénál). A PurePos lehetőséget biztosít morfológiai elemző integrálására, így a lemmatizálás során a morfológiai elemzőt használja azoknak a szavaknak az elemzésére és lemmatizálására, amik benne vannak a morfológiai elemzőben. Azoknak a szavaknak a tövesítéséhez

viszont, amik nem szerepelnek a morfológiai elemző lexikonában, egy egyszerű lemmatizáló algoritmust használ. Ez az algoritmus a TnT tagger eredeti szuffix guesser algoritmusának [2] egy olyan kiegészítése, amely az adott szóalakból a lemmát levezető kódot is tartalmazza az adott szóhoz hozzárendelt morfoszintaktikai címkén kívül[1]. Ennek az algoritmusnak a pontossága viszonylag alacsony (89,01% sztenderd szövegen mérve, tehát minden 10. szóhoz hibás lemmát generál [1]).

Ugyanakkor, nem csak a morfológiai elemző számára ismeretlen szavakból jöhetnek létre hibás lemmák. Számos esetben a morfológiai elemző is olyan elemzéseket produkál, amit valamilyen produktív minta alkalmazásával hoz létre. Ezek azonban nem feltétlenül lesznek értelmes elemzések, a lemma sem lesz helyes.

A hibásan lemmatizált alakok fő forrását a rossz minőségű szövegek jelentik. Ha a szöveg sok hibás, elírt vagy nem sztenderd szóalakot tartalmaz, nem tud ezekre a szavakra a morfológiai elemző elemzést adni, így a rendszer a kevésbé pontos, végződés alapú guessert használja. Ezeknek a szavaknak az esetében azonban tulajdonképpen nincs is esélye az adott szóalakból a lemmát algoritmikusan kiszámolni próbáló eszközöknek arra, hogy helyes eredményt kapjanak: ehhez a szöveget előbb ki kellene javítani.

Kutatásunk során az volt a célunk, hogy egy olyan algoritmust hozzunk létre, amelyik a hibásan lemmatizált szavakat azonosítja, függetlenül attól, hogy a guesser vagy a morfológia hozta-e létre ezeket.

2. Módszer

Kutatásunkban a PurePos által létrehozott lemmáknak három osztályát definiáltuk:

1. Nem lemma (`notlem`): a létrejött szótőjelölt nem lemma, hanem ragozott alak. Pl. *nyelvészt[FN]*
2. Rossz elemzés (`badana`): a létrejött szótőjelölt a hozzárendelt szófajcímkével nem helyes. Pl. *vár[MN]*
3. Helyes lemma (`lem`): a létrejött szótőjelölt a hozzárendelt szófajcímkével együtt helyes. Pl. *vár[IGE]*

A lemmajelöltek e három osztályba való besorolására két módszert próbáltunk ki. Először egy karakteralapú rekurrens neurális hálózatot tanítottunk be, majd egy SVM (Support Vector Machine) osztályozót hoztunk létre. Az SVM bemenetére a lemmajelölteknek egy olyan reprezentációja került, ami az eredeti szóalak és a lemma szóbeágyazási vektorából és a szófajcímkéből áll. A következő alfejezetekben a rendszer létrehozásának és tanításának lépéseit ismertetjük.

2.1. A tanítóanyag létrehozása

A tanítóanyag előállításához egy 1,2 milliárd tokenből álló magyar nyelvű webkorpuszt használtunk [3]. Mivel a webről gyűjtött korpuszban elég sok a zajos,

illetve nem sztenderd nyelvhasználattal írt szövegrész (fórumok, kommentek, blogok, stb.), ezért ez a típusú forrás elég sok hibásan elemzett és lemmatizált szót tartalmazott. A korpuszt a Humor morfológiai elemzővel [4,5] kiegészített PurePos eredeti verziójával elemeztettük, majd kigyűjtöttük belőle azokat a szavakat, amik legalább ötször előfordultak a korpuszban. A létrejött listában szereplő lemmák vagy az integrált morfológiai elemzőből származnak, vagy a guesser algoritmus hozta őket létre. A kapott szólistán ezért lefuttattuk a morfológiai elemzőt, és azokat a szóalakokat, amik a Humor lexikonában lemmaként szerepelnek, tehát a morfológia felismerte, az adott szófajjal helyes lemmának tekintettük (`lem`). Azokat a szavakat, amik a morfológia szerint helyes szóalakok, de nem lemmák, hanem egy szónak valamilyen ragozott alakjai, azokat a „nem lemma" osztályba soroltuk (`notlem`).

A harmadik osztályba pedig olyan szavakra kellett példákat gyűjtenünk, amiknek hibás volt az elemzése, amik biztosan nem helyes lemmák az adott szófajcímkével. Az ilyen jellegű rosszul lemmatizált szóalakok igen jellemzőek egy bizonyos típusú hibát nagy mennyiségben tartalmazó szövegtípusra, mint például ékezet nélküli szövegrészek, rosszul tokenizált szövegrészek, stb. Ezek tehát a tanítóanyag létrehozásához használt korpusznak jellemzően ilyen részleteiből kerültek ki, bár nem szűkítettük le a korpuszt ilyen részkorpuszokra a hibás szavak gyűjtésekor sem. Az ebbe a csoportba tartozó szóalakoknak a kigyűjtését jobb híján úgy oldottuk meg, hogy először kerestünk olyan szavakat, amik egyrészt lemmaként szerepeltek az elemzett korpuszban és volt is olyan szófajú elemzésük, amilyen szófajúként az adott szó az adott alakban ténylegesen szerepelt, másrészt viszont volt egy másik elemzésük más szófajjal, amilyen szófajjal az adott szó az adott alakban nem szerepelt a korpuszban. Például a *vár* lemma főnévi (FN) és igei (IGE) elemzése mellett melléknévként (MN) is megjelenik az elemzés után, mint melléknévi lemma. Ezek közül viszont csak a *vár[FN]* és a *vár[IGE]* alakok szerepelnek önmagukban is, a *vár[MN]* elemzést a PurePos lemmatizáló guessere állította elő valamilyen más szóalak egyébként hibás elemzéseként. Néhány ilyen szóalakból kiindulva feltérképeztük a korpuszból épített szóbeágyazási modellben (lásd a 2.2. részt) az ezek környezetében szereplő szavakat, kihasználva azt a korábbi megfigyelést, hogy a szóbeágyazási térben a hasonló jellegű hibát tartalmazó szóalakok egymáshoz közel helyezkednek el [6]. A 1. táblázatban néhány ilyen hibásan lemmatizált szó és legközelebbi szomszédaik láthatók. Néhány ilyen szóalak szomszédsági listáját kézzel ellenőrizve hoztuk létre a hibás lemmák osztályába (`notlem`) tartozó tanítómintát.

Végül 11066 rosszul elemzett szó, 82582 helyes lemma és 80000 ragozott szóalak alkotta a tanítóanyagot, amiből mindhárom kategóriából 1000-1000 véletlenszerűen kiválasztott elemet különítettünk el a teszteléshez.

2.2. Szóbeágyazási modellek

A neurális szóbeágyazási modellek a tanítóanyagban (korpuszban) szereplő szavakat egy néhány száz dimenziós térben helyezik el, ahol a hasonló jelentésű, és/vagy hasonló szintaktikai, grammatikai, stiláris, stb. tulajdonságokkal rendelkező szavak egymáshoz közel, a kevésbé hasonlóak távolabb helyezkednek el

vár[MN]	alma[IGE]	szép[IGE]
tud[MN]	funda[IGE]	ilo[IGE]
kér[MN]	mangus[IGE]	nari[IGE]
jé[MN]	manda[IGE]	evian[IGE]
fogad[MN]	spiritualité[FN]	insa[IGE]
cso[MN]	Tibi[IGE]	manam[IGE]

1. táblázat. Példa hibásan lemmatizált és elemzett szavakra, illetve a hozzájuk legközelebb eső elemek az elemzett szövegből épített szóbeágyazási modellben

[7,8]. A már korábban létrehozott modelljeink közül [9] kettőt használtunk fel ebben a kutatásban. Az egyiket a nyers korpuszon tanítottuk, így a szavak eredeti alakjukban szerepelnek benne. A másik modell tanításakor pedig a korpuszra szófaji egyértelműsítést, illetve morfológiai elemzést alkalmaztunk, majd a tanítás előtt a szótöveken csak a fő szófajcímkét hagytuk meg, a címkék további morfoszintaktikai információt tartalmazó részét leválasztottuk, és az önálló tokenként szerepelt a tanítóanyagban. A modellek részletes összehasonlítását lásd [9]-ben. Mindkét modell esetében a tanítás során a reprezentációt létrehozó neurális hálózat a CBOW architektúrát valósította meg 5 token sugarú ablakméret alkalmazásával 300 dimenziós vektorokat hozva létre.

2.3. Karakteralapú rekurrens neurális hálózat használata az osztályozási feladatra

Az utóbbi évek egyik legeredményesebb kutatási iránya a neurális hálózatok, illetve a mélytanulás módszereinek alkalmazása számos, köztük szövegfeldolgozási, területen előforduló problémákra. Ezért első megoldásként egy rekurrens neurális hálózatot tanítottunk be a létrehozott tanítóanyag alapján. Ez a modell egyáltalán nem használja az előző részben bemutatott szóbeágyazási modelleket. A neurális hálózat bemenetén a vizsgálandó szó karakterei egymást követve jelennek meg, végül pedig a szófajcímke. Az egyes szavakat alkotó karaktereket one-hot vektor formában reprezentáltuk. Ennek a vektornak a hossza megegyezik a tanító- és tesztanyagban szereplő különböző karakterek és szófajcímkék számával. Az egyes pozíciókon pedig 0-k szerepelnek, kivéve az aktuális karakter/címke indexének megfelelő pozíciót, ahol 1-es áll.

A létrehozott rekurrens neurális hálózat (RNN) egy kétrétegű hálózat, egy rejtett állapottal. Minden lépésben a következő bemenetet (karakter, illetve címke), illetve az előző lépésben kapott eredményt használja fel (az első lépésben ez nulla). Az egyes lépések kimenete pedig egy, a három osztálynak megfeleltethető háromelemű vektor, ami a bemenet addigi feldolgozása alapján az egyes osztályokba való tartozás valószínűségét tartalmazza. Ezt a részeredményt használja rejtett állapotként is, amit a bemenetre visszacsatolva a következő lépésben felhasznál a következő karakter vagy szófajcímke beolvasása alapján adott predikció során. A kimeneti vektor normalizálására egy LogSoftmax réteget használtunk. A

neurális hálózat architektúrája az 1. ábrán látható, aminek az implementációjához a PyTorch keretrendszert[1] használtuk a következő paraméterbeállításokkal: 256 volt a rejtett réteg dimenziószáma, a rendszert 200000 epochon keresztül tanítottuk, a tanulási faktor 0,005 volt.

A vizsgált `szóalak[CÍMKE]` formájú bemenet feldolgozása során tehát az RNN megtippeli, hogy az adott elem mekkora valószínűséggel tartozik az egyes osztályokba. Bár a tanítóanyagban egy-egy elem akár több osztályban is szerepelt, a kiértékelés során az adott szóhoz legnagyobb valószínűséggel rendelt osztályt vettük csak figyelembe.

1. ábra: A rekurrens neurális hálózat sematikus ábrája

2.4. Support Vector Machine alapú osztályozó

Egy másik kísérletben azt vizsgáltuk, hogy az említett szóbeágyazási modellek felhasználásával egy hagyományos osztályozó algoritmus milyen pontossággal oldja meg az adott osztályozási feladatot. A support vector machine (SVM) algoritmus egy olyan felügyelt tanítási módszert valósít meg, ami minden egyes elemet egy n-dimenziós tér egy pontjának tekint és ebben a térben keresi meg azokat a hipersíkokat, amik a különböző osztályokba tartozó elemeket egymástól elkülönítik. A használat során pedig minden egyes új bemenetet ugyanebbe a térbe helyez el, az adott elem pozíciójának megfelelő osztály lesz a rendszer által hozzárendelt besorolás. Ebben az esetben tehát minden elemhez egyetlen osztályt (és nem az összes osztályba való tartozás valószínűségeinek vektorát) kapjuk eredményként.

Az SVM bemenete tehát egy n-dimenziós vektor, ami esetünkben három részből állt:

[1] http://pytorch.org/

- a vizsgált szónak (szófajcímke nélkül) a szavak felszíni alakjából épített szóbeágyazási modellben szereplő 300 dimenziós vektora. Ha nem szerepelt a szóalak a modellben, akkor egy nullákból álló 300 dimenziós vektort rendeltünk hozzá. A reprezentációnak ez a része azzal kapcsolatos információt tárol, hogy a lemmajelölt szerepel-e a modellben (illetve a korpuszban), és ha igen, akkor a kapott vektor jellemzi az adott szóalak grammatikai tulajdonságait. Lemmaként normális esetben a szó valamelyik ténylegesen létező alakját használjuk, és hogy konkrétan melyiket, az a szó szófajától függ. Ez az összetevő tehát egyrészt arra a feltevésre épül, hogy a toldalékolt szavak lemmája is nagy valószínűséggel előfordul a nyers korpuszban,[2] másrészt hogy mind a valódi lemmák, mind a ragozott szóalakok esetében a szófajcímke és a lemmajelölt vektorreprezentációja között rendszeres összefüggést feltételezhetünk.
- a vizsgált szónak (szófajcímkével együtt) a lemmákból és morfoszintaktikai jegyekből épített szóbeágyazási modellben szereplő 300 dimenziós vektora. Ha nem szerepelt a szóalak a modellben, akkor egy nullákból álló 300 dimenziós vektort rendeltünk hozzá.
- a szófajcímkék one-hot vektora, amit az RNN-ben használt reprezentációhoz hasonlóan úgy hoztunk létre, hogy egy, az összes különböző szófajcímke számával megegyező hosszúságú vektort nullákkal töltöttünk fel, csupán az aktuális szó szófajához tartozó pozícióban szerepelt 1-es. Az általunk használt tanítóanyagban és tesztanyagban összesen 23 különböző szófajcímke szerepelt, tehát ez az összetevő egy 23 dimenziós vektor.

Így minden szóhoz egy 623 dimenziós vektor állt elő a fenti összetevők alapján, ami az SVM bemeneteként szerepel. Azért, hogy ne csupán lineáris szeparáció jöhessen létre a tanítóminták alapján, a tanítás során RBF kernelt használtunk.

3. Eredmények

A kiértékelés során mindhárom osztályból 1000-1000, a tanítóanyagban nem szereplő mintát használtunk tesztanyagként. Mivel a tanítóanyagban egy-egy szó akár több osztályban is szerepelhetett, a véletlenszerűen választott mintában előfordultak ismétlődések, ezeket a mérések pontossága érdekében kitöröltük, így végül 2974 szón értékeltük ki a modelleket. A számszerű eredményeket a 2. táblázat tartalmazza.

A rendszerek minőségét két szempontból vizsgáltuk. Az első esetben a három osztályra vonatkozó pontosságot mértük, csak azt az eredményt tekintve helyesnek, amikor a ténylegesen helyes osztályt eltalálta a rendszer (lásd *Pontosság (3 osztályra)*). Ez az RNN esetében 74,37% volt, míg az SVM esetében 87,86%. Mindkét esetben tehát a vizsgált szavaknak legalább a háromnegyedéhez a helyes osztály került hozzárendelésre, de az SVM esetén ez az arány majdnem elérte a

[2] Ez alól van néhány kivétel, mint pl. az *ága-boga* szó, amelynek esetében a lemmatizáló az *ág-bog* alakot állítja elő lemmaként, illetve más hasonló kifejezések.

90%-ot is. A fő célunk azonban nem a pontos besorolás volt, hanem egy olyan algoritmus létrehozása, ami egy lemmajelöltről el tudja dönteni, hogy a hozzárendelt szófajjal együtt helyes-e vagy sem. Ezért a rendszerek teljesítményét olyan formában is kiértékeltük, hogy azt nem vettük figyelembe, hogy a `badana` és a `notlem` osztályokat meg tudják-e különböztetni. Így az RNN 80,53%, míg az SVM 92,13% pontossággal tudta megítélni, hogy egy lemmajelölt helyes-e vagy sem. Mindkét kiértékelés során látható tehát, hogy beágyazási vektorokat használó SVM felülmúlta a kizárólag a karaktersorozat alapján dolgozó RNN alapú rendszert. Bár a szóvégződés igen meghatározó tényező, az utóbbinak azonban semmilyen azzal kapcsolatos információ nem állt rendelkezésére, hogy az adott szó milyen kontextusokban szokott megjelenni.

	Pontosság (3 osztályra)	Pontosság (2 osztályra)
RNN	74,37%	80,53%
SVM	87,86%	92,13%

2. táblázat. A két rendszer pontossága 3, illetve 2 osztályra vizsgálva (`lem` vs. {`notlem`|`badana`})

Az eredmények részletesebb vizsgálata során a konkrét tévesztéseket is számszerűsítettük, amit a 3. táblázatban összesítettünk. Látható, hogy az RNN sokkal több szót sorolt a `badana` osztályba, ezeknek egy jelentős része (31,64%) azonban valójában vagy szófajilag helyesen osztályozott ragozott szó, vagy helyes lemma. Az SVN azonban – bár alacsonyabb fedéssel találta el ennek az osztálynak az elemeit – csupán 5 olyan szót sorolt ebbe az osztályba, amik valójában helyes lemmák az adott szófajcímkével (az 5 példa: *sziasztok[MSZ], Salaam[FN], szí[IGE], jobbul[IGE], Karmelina[FN]*). Elmondható tehát, hogy az SVM alapú megoldás igen pontos, amelyik lemmajelöltet hibásnak jelzi, az nagy valószínűséggel tényleg hibás. A `notlem` és `lem` osztályok közötti tévesztések számát tekintve is sokkal kisebb a hibák száma az SVM esetén. Azoknak a lemmajelölteknek, amik az SVM alapú rendszer szerint nem lemmák, 97,8%-a valóban nem az. Sajnos az SVM-es rendszer által jó lemmaként azonosított jelöltek pontossága alacsonyabb: ez csak 83,3%. A két rendszert összevetve csupán a `badana` osztály fedését tekintve mutat az RNN valamivel jobb teljesítményt, az ebbe az osztályba tartozó tesztesetekből az SVM többet sorolt a másik két osztály valamelyikébe. A két rendszer kombinálásával feltehetőleg javítani lehetne a hibásként azonosított lemmák fedését a pontosság jelentős romlása nélkül.

Az eredmények vizsgálata során az is látható volt, hogy sok esetben a hibás lemmatizálás forrása helyesírási hiba, elírás volt. Ezeket a szavakat, amiknek így a lemmája is hibásan szerepelt, a `badana` osztályba soroltuk mind a tanításnál, mind a kiértékelésnél akkor is, ha az eredeti szó helyes alakja esetén jó lenne az elemzés a helyes lemmával. Például a *butít* szó rövid *i*-s *butit* alakja a *butit[IGE]* elemzéssel szerepel, ami egyébként helyes lenne a hosszú *í*-s változat esetén. A létrehozott algoritmus ezeket az alakokat gyakran helyes lemmának és

RNN		eredmény		
		badana	notlem	lem
cél	badana	713	93	156
	notlem	90	840	70
	lem	240	113	659

SVM	eredmény		
	badana	notlem	lem
badana	646	127	189
notlem	0	994	6
lem	5	34	973

3. táblázat. A két rendszer tévesztési mátrixa a vizsgált tesztanyagon mérve

elemzésnek tekinti, ami tulajdonképpen zajos szövegek feldolgozása esetén akár elfogadható eredmény is lehet.

4. Összegzés

Cikkünkben bemutattunk két olyan algoritmust, amik annak eldöntésére képesek, hogy egy szóalakhoz automatikusan létrejött lemma az eredeti szóalakhoz rendelt szófajcímkével együtt helyes-e vagy sem. A rekurrens neurális hálózatot, illetve a support vector machine módszert megvalósító rendszerek közül az utóbbi teljesítménye lett kiemelkedően jó, 92% fölötti pontosságot érve el. Az algoritmusban használt modell betanításához egy automatikusan létrehozott és kézzel ellenőrzött tanítóanyagban szereplő szavaknak két különböző jellegű, magyar nyelvű szóbeágyazási modellből származó reprezentációját, illetve a szó eredeti szófaját használtuk fel. Bár a bemutatott algoritmus nem képes arra, hogy a hibásnak ítélt lemmák esetén javaslatot tegyen a helyes alakra, a szófaji egyértelműsítő rendszerbe való integrálásával akár a helyes lemma előállítása is lehetséges lenne. A beépített lemmatizáló ugyanis a morfológiai elemző által nem ismert szavakhoz több lehetséges lemmát is generál, ezek közül választva ki azt, amelyikhez a legnagyobb valószínűséget rendelte. A bemutatott rendszer azonban alkalmas lehetne arra, hogy az ebben a listában esetleg nagyobb valószínűséggel szereplő, de nem helyes lemmákat elvesse, így a kisebb „valószínűségű”, de helyes lemmák is megjelenhetnének az eredményekben.

Köszönetnyilvánítás

Jelen kutatás az FK 125217 és a PD 125216 számú projekt keretében a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással az FK 17 és a PD 17 pályázati program finanszírozásában valósult meg.

Hivatkozások

1. Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, Incoma Ltd. Shoumen, Bulgaria (2013) 539–545

2. Brants, T.: Tnt: A statistical part-of-speech tagger. In: Proceedings of the Sixth Conference on Applied Natural Language Processing. ANLC '00, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 224–231
3. Endrédy, I., Prószéky, G.: A pázmány korpusz. Nyelvtudományi Közlemények **112** (2016) 191–206
4. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
5. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. ACL '99, Stroudsburg, PA, USA, Association for Computational Linguistics (1999) 261–268
6. Novák, A.: Improving corpus annotation quality using word embedding models. Polibits **53** (2016) 49–53
7. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. (2013) 746–751
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. (2013) 3111–3119
9. Novák, A., Novák, B.: Magyar szóbeágyazási modellek kézi kiértékelése. In: XIV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2018)

IV. Beszédtechnológia

Beszédfelismerők mély neuronhálós állapotkapcsolási algoritmusainak kísérleti összehasonlítása

Tóth László[1], Grósz Tamás[1,2], Gosztolya Gábor[2]

[1]Szegedi Tudományegyetem, Informatikai Intézet
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
{ tothl, groszt, ggabor } @ inf.u-szeged.hu

Kivonat Az utóbbi években a rejtett Markov-modelles beszédfelismerőkben a kibocsátási eloszlások becslésére használt Gauss-keverékmodelleket a mély neuronhálók váltották fel. Azonban az elmúlt három évtizedben a Gauss-alapú modellezés javítására számos algoritmikai finomítást vezettek be, amelyeknek a mély neuronhálós környezetbe való átültetése nem mindig triviális. Egy ilyen példa a környezetfüggő beszédhangmodellek létrehozására kitalált döntési fa-alapú állapotkapcsolási eljárás. Jelenleg mindenki a régi, jól bevált algoritmust használja a mély hálós felismerőkben is, annak ellenére, hogy az algoritmus speciálisan a Gauss-görbék illeszkedését használja ki, így optimalitása egy mély hálós rendszerben megkérdőjelezhető. Az utóbbi időben azonban több olyan állapotklaszterező algoritmust is publikáltak, amelyek megkísérlik a korábbi eljárást a mély neuronhálós modellezéshez igazítani. Jelen cikkben négy ilyen algoritmust hasonlítunk össze egy angol nyelvű adatbázison, majd a legjobbnak bizonyuló módszert egy magyar korpuszon is kiértékeljük. Eredményeink azt mutatják, hogy ezek az új algoritmusok szignifikánsan jobb eredményt adnak, mint a régi, Gauss-alapú megoldás, köszönhetően annak, hogy működésük a neuronhálók kimenetéhez van igazítva.
Kulcsszavak: beszédfelismerés, mély neuronháló, környezetfüggő modellezés, állapotkapcsolás

1. Bevezetés

A gépi beszédfelismerésben használt rejtett Markov-modellek (Hidden Markov Model, HMM) legfontosabb komponense az akusztikus jellemzők eloszlását leíró valószínűségi modell, mely célra hagyományosan Gauss-eloszlások keverékét (Gaussian Mixture Model, GMM) szokták alkalmazni. Az utóbbi években azonban a mély neuronhálók (Deep Neural Network, DNN) annyival jobbnak bizonyultak ezen a téren, hogy a hagyományos, GMM-alapú rendszereket (röviden HMM/GMM) teljesen kiszorította a DNN-alapú hibrid megoldás (röviden HMM/DNN hibrid). A HMM/GMM modellek mögött azonban több évtizednyi fejlesztés van, számos olyan algoritmikai trükkel, amelyeket nem feltétlenül lehet a hibrid HMM/DNN sémára triviális módon átültetni. Egy ilyen példa a

beszédfelismerők betanítása részletes szegmentálási annotáció nélkül (angol szakzsargonban "flat start" tanítás), egy másik fontos feladat pedig a környezetfüggő (context-dependent, CD) beszédhang-modellek automatikus kialakítása. A hagyományos HMM/GMM keretben mindkettőre jól bevált algoritmusok léteznek, így továbbra is ezeket a módszereket használjuk a hibrid rendszerek inicializálása során. Ez gyakorlatilag azt jelenti, hogy a HMM/DNN modell betanítása egy HMM/GMM modell betanításával kezdődik, amelyet a fenti két részfeladat megoldása után eldobunk. Ez a módszer ugyan működőképes, de egyáltalán nem gazdaságos, és valószínűleg nem is optimális.

A flat start tanítás lényege a modellek iteratív újratanítása és újraillesztése, ami mögött a HMM/GMM rendszerek esetében jól kidolgozott matematikai háttér és komoly gyakorlati tapasztalat van. Ugyanezt mély hálós rendszerekkel megvalósítani nem magától értetődő, bár az utóbbi években többen megmutatták, hogy megfelelő odafigyeléssel nem lehetetlen [1,2,3,4]. A környezetfüggő modellezés mély hálós megvalósítása még kevésbé megoldott. A hibrid modellek feltalálói évtizedekig ragaszkodtak a környezetfüggetlen (context-independent, CI) modellezéshez [5], azóta azonban kiderült, hogy a környezetfüggő modellek a mély neuronhálós HMM/DNN hibridek pontosságát is szignifikáns mértékben növelik [6,7]. Ebből kifolyólag fontos lenne hatékony algoritmusokat találni a CD modellek automatikus kialakításához. Jelenleg mindenki a HMM/GMM rendszerek jól bevált döntési fa-alapú állapotklaszterező eljárását használja e célra [8]. Ez a módszer Gauss-eloszlásokat illeszt az egyes állapotokhoz tartozó adatpontokra, majd az illeszkedés pontossága alapján tart egyben vagy oszt szét egyes fonéma-állapotokat (a későbbi osztályokat). Ezt a lépést iteratívan ismételgetve az eljárás egy fa-jellegű hierarchikus klaszterezést eredményez. A módszer Gauss-alapú modellezés esetén remekül bevált, azonban erősen kétséges, hogy a Gauss-görbék illeszkedése mennyire jó kritérium, amikor a végső modell egy teljesen más reprezentáció, például egy neuronháló lesz.

További probléma, hogy a GMM- és a DNN-alapú rendszerek esetén az input reprezentálási módja is jelentősen eltérhet. HMM/GMM rendszerek esetében jól bevált az ún. mel-frekvenciás kepsztrális együtthatók (mel-frequency cepstral coefficient, MFCC) használata. Úgy tűnik azonban, hogy a HMM/DNN rendszerek némileg jobban működnek olyan primitívebb jellemzőkkel, mint például a mel-frekvenciasávok energiái [9]. A HMM/GMM modelleket azonban ezeken a jellemzőkön nem lehet jól betanítani. A jelenlegi módszer szerint tehát az MFCC jellemzőket is pusztán az állapotklaszterező eljárás kedvéért generáljuk le, majd utána eldobjuk. Ennél nyilván léteznie kell jobb megoldásnak is.

Az utóbbi néhány évben több szerző is megpróbálkozott a régi klaszterező eljárás DNN-ekhez való igazításával. A jellemzőkészletek eltéréséből eredő probléma elkerülésére azt a megoldást javasolták, hogy a klaszterezést a DNN kimenetén futtassuk le, ne pedig a bemenő jellemzőkön. Ezzel az egyszerű módszerrel többen is próbálkoztak [1,2,10,11]. Bár ez a huszárvágás kikerüli az eltérő jellemzőkészletek problémáját, a klaszterező eljárás ugyanaz marad, így a Gauss-os eloszlási feltevéstől továbbra sem szabadultunk meg.

Más szerzők a klaszterező eljárás döntési kritériumát is módosítják oly módon, hogy az jobban illeszkedjen a neuronhálós eloszlás-modellezéshez. Mivel a neuronháló kimenete egy diszkrét valószínűségi eloszlás közelítéseként értelmezhető, Gosztolya és tsai egy korábbi cikkükben döntési kritériumként a Kullback-Leibler divergencia használatát javasolták [12]. Módszerük alapötletét Imseng és tsai algoritmusa adta, amelyet eredetileg az ún. KL-HMM modellek betanításához találtak ki [13]. Zhu és tsai egy entrópia-alapú faépítő kritériumot javasoltak [14]. Végezetül, Wang és tsai egy speciális hálózaton alapuló klaszterezési megoldással álltak elő, amely mély kanonikus korrelációelemzésre (Deep Canonical Correlation Analysis) van optimalizálva [15].

A fenti szerzők mindegyike a felismerési szóhiba (word error rate, WER) csökkenését tapasztalta az új algoritmusok alkalmazásával a hagyományos GMM-alapú eljáráshoz képest. Azonban egyik szerzőcsoport sem hasonlította össze a cikkét a többi hasonló módszerrel. Ráadásul mindegyikük más-más adatbázist használt, ami az egyes módszerek egymással való összevetését teljesen lehetetlenné teszi. Jelen cikkben négy ilyen algoritmust hasonlítunk össze ugyanazon a nagyszótáras felismerési feladaton, ráadásul ugyanaz a neuronháló fogja szolgáltatni a klaszterező algoritmusok inputját. Megjegyezzük, hogy a kiindulási, környezetfüggetlen címkék illesztését a korábbiakban javasolt neuronhálós "flat-start" módszerrel végeztük [3], így ezen lépés során sem használtunk GMM modelleket. Ebből kifolyólag az összehasonlított rendszerek közül az a kettő, amelyikben a döntési kritérium DNN-ekhez van igazítva (azaz Gosztolya és tsai [12] valamint Zhu és tsai [14]) teljesen "GMM-mentes" megoldást ad a teljes tanítási folyamatra.

2. Döntési fa-alapú állapotkapcsolás

A döntési fa-alapú állapotkapcsolási algoritmus Young és tsai nevéhez fűződik [8], és mára a nagyszótáras beszédfelismerő rendszerek készítésének elengedhetetlen komponensévé vált. A módszer alapötlete, hogy az adott beszédhang (pontosabban modell-állapot) kontextusfüggő variánsait összevonjuk, majd a halmazt lépésről lépésre részhalmazokra osztjuk, ami egy hierarchikus, faként reprezentálható klaszterezést eredményez. Az egyes lépésekben az előre definiált kérdésekkel megadott felosztási lehetőségek közül az algoritmus mindig azt választja, amelyre az egyben hagyott, illetve a szétosztott halmazok maximálisan különböznek. Ennek a különbözőségnek, pontosabban a kiindulási $\mathcal{S}$ állapothalmaz szétosztásából származó nyereségnek a mérésére az algoritmus egy valószínűségi alapú döntési kritériumot használ. Ez a módszer annyira bevált, hogy az elmúlt húsz évben csak apró módosításokat javasoltak rajta (például a faépítést vezérlő, fonetikai ismeretet igénylő kérdések automatikus generálását [16]).

2.1. A valószínűségi (likelihood) alapú döntési kritérium

Tegyük fel, hogy adott állapotok egy $\mathcal{S}$ halmaza, amelyeket klaszterezni (majd ez alapján összekapcsolni) szeretnénk Young és tsai módszerével. A döntési fa építése során minden csomópontnál két nem átfedő részhalmazra osztjuk az aktuális

$\mathcal{S}$ halmazt az illeszkedő kérdések alapján, a kérdésre adott válasznak megfelelően. Az optimális szétosztás meghatározására Odell egy maximum likelihood alapú döntési kritériumot definiált [17], majd egy hatékony algoritmust is javasolt, amely a kritériumot az alábbi módon közelíti:

$$L(\mathcal{S}) \simeq -\frac{1}{2}\big(\log[(2\pi)^K|\Sigma(\mathcal{S})|] + K\big)\sum_{s\in\mathcal{S}} N(s), \tag{1}$$

ahol $s \in \mathcal{S}$ jelöli az egyes állapotokat, $\Sigma(\mathcal{S})$ az $\mathcal{S}$-be eső adatok szórását, $N(s)$ pedig az s halmazba eső példák (adatvektorok) számát. A fenti formula alapján minden lépésben azt a q kérdést fogjuk választani, amely maximizálja a likelihood érték $\Delta L(q|\mathcal{S})$ változását:

$$\Delta L(q|\mathcal{S}) = \big(L(\mathcal{S}_y(q)) + L(\mathcal{S}_n(q))\big) - L(\mathcal{S}), \tag{2}$$

ahol $\mathcal{S}_y(q)$ és $\mathcal{S}_n(q)$ a $\mathcal{S}$ halmaz két részhalmaza, amelyek a q kérdés alkalmazásával állnak elő. Megjegyezzük, hogy a likelihood értékek nem függnek direkt módon az egyes tanítópéldáktól, csupán az egyes állapotokra eső tanítópéldák szórásától, valamint a példák számától.

3. Neuronháló-alapú állapotkapcsolás

A fent ismertetett kritérium a példák szórására épül, ami Gauss-alapú modellezés esetén ésszerűen hangzik, hiszen a Gauss-görbék egyik fő paramétere a szórás. Ha azonban a GMM-alapú modellek helyett neuronhálóra épülő modelleket készítünk, akkor a fenti formula használata kérdésessé válik. Ehhez jön még a jellemzőkészletek esetleges eltéréséből fakadó korábban ismertetett dilemma.

E két problémára több szerző is azt a megoldást javasolta, hogy Gauss-görbék illesztése helyett tanítsunk be egy neuronhálót a kiindulási környezetfüggetlen címkék osztályozására (ez lehet hagyományos vagy mély háló is). Ezt a hálót a továbbiakban "segéd" neuronhálónak fogjuk nevezni, és röviden CI-NN jelöléssel fogunk rá hivatkozni. A betanításához szükséges CI címkék időbeli illesztését akár egy hagyományos HMM/GMM rendszer betanításával, akár az utóbbi időben javasolt DNN-alapú 'flat-start' technikák segítségével (pl. [2,3,4]) is elvégezhetjük. A lényeg jelen esetben, hogy az állapotkapcsoláshoz szükséges klaszterezést nem a bemenő jellemzőkön, hanem a segéd-neuronháló kimenetén végezzük el. Ily módon a klaszterezés során nem lesz jelentősége annak, hogy a segéd-neuronháló milyen jellemzőkön tanult. Ki kell azonban találnunk, hogy a klaszterező algoritmust miként igazítsuk a segédháló kimeneti értékeihez. A továbbiakban erre ismertetünk négy módszert.

3.1. A segéd-neuronháló kimenő vektorainak klaszterezése

Ez a Senior és tsai által javasolt megoldás a lehető legegyszerűbb. Ők közvetlenül a segédháló kimenetén futtatják le a klaszterző algoritmust, míg magát az

algoritmust semmilyen módon nem módosítják. A módszert ismertető cikkükben a szószintű hiba kis javulását tapasztalták [10]. Bár cikkük hangzatosan a "GMM-mentes" címet viseli, magán a Gauss-görbék illesztésén alapú klaszterezési módszeren semmit sem változtatnak, így véleményünk szerint a címadás legalábbis megtévesztő.

3.2. A DNN rejtett rétegének kimenetén végzett klaszterezés

Senior és tsai munkájával párhuzamosan Bacchiani és Rybach azt javasolta, hogy a klaszterezést a segédhálónak ne a kimenetén, hanem a legutolsó rejtett rétegén végezzük el [11]. Magán a klaszterezési eljáráson nem változtattak, ők is a jól bevált Young-féle algoritmust alkalmazták. Úgy találták, hogy kisebb állapotszámok esetén ez a megoldás jobban működött, mint a standard eljárás, habár nagy állapotszám esetén ez megfordult. Ezt a keretszintű címkék pontatlan illesztésével próbálták magyarázni, az illesztést ugyanis egy hagyományos HMM/GMM rendszerből vették át.

3.3. Kullback-Leibler divergencián alapuló klaszterezés

Az eddig ismeretett két módszer egyike sem változtatta meg a klaszterezés során használt döntési kritériumot, csupán az inputtal játszottak. Gosztolya és tsai viszont egy korábbi cikkükben javaslatot tettek egy neuronháló-kimenetekhez igazított döntési kritériumra [12], melynek alapötletét Imseng és tsai Kullback-Leibler HMM-ekhez kifejlesztett módszere adta [13]. Ez a módszer teljesen GMM-mentes, ahogy az alábbiakban röviden összefoglaljuk, [12], [18] és [19] alapján.

A neuronhálók kimenete egy diszkrét valószínűségi eloszlásként értelmezhető. Ilyen eloszlások összehasonlítására használhatjuk a Kullback-Leibler divergenciát, amely egy z_t és egy y_s vektor fölött így definiálható:

$$D_{KL}(y_s||z_t) = \sum_{k=1}^{K} y_s(k) \log \frac{y_s(k)}{z_t(k)}, \tag{3}$$

ahol $k \in \{1, \ldots, K\}$ a vektor dimenziója [20]. Ez alapján a likelihood érték maximalizálása helyett a KL-divergencia minimalizálására fogunk törekedni:

$$D_{KL}(\mathcal{S}) = \sum_{s \in S} \sum_{f \in F(s)} \sum_{k=1}^{K} y_{\mathcal{S}}(k) \log \frac{y_{\mathcal{S}}(k)}{z_f(k)}, \tag{4}$$

ahol $\mathcal{S}$ az s állapotok egy halmaza, $F(s)$ pedig az adott állapothoz tartozó tanítópéldák. Az $\mathcal{S}$-hez tartozó példákat reprezentáló vektor (azaz $y_{\mathcal{S}}$) úgy kapható meg, mint az $\mathcal{S}$-be eső példák mértani közepe, bár Imseng eredeti cikkében megemlíti, hogy a számtani közép is alkalmazható e célra [21]. Kibontás és egyszerűsítés után azt kapjuk, hogy

$$D_{KL}(\mathcal{S}) = -\sum_{s \in \mathcal{S}} N(s) \log \sum_{k=1}^{K} y_{\mathcal{S}}(k), \tag{5}$$

vagyis az állapotok egy $\mathcal{S}$ halmazának KL-divergenciája az egyes állapotokhoz tartozó y_s és $N(s)$ statisztikai értékekből kapható meg [18]. A $\mathcal{S}$ állapothalmaz szétosztásakor magától értetődően azt a kérdést érdemes választani, amelyre a KL-divergencia $\Delta D_{KL}(q|\mathcal{S})$ változása maximális:

$$\Delta D_{KL}(q|\mathcal{S}) = D_{KL}(\mathcal{S}) - \big(D_{KL}(\mathcal{S}_y(q)) + D_{KL}(\mathcal{S}_n(q))\big). \tag{6}$$

3.4. Entrópián alapuló állapotklaszterezés

Zhu és tsai egy negyedik megoldási lehetőséget javasoltak [14]. Ez szintén az Odell-féle kritérium (Eq. (1)) egy Gauss-görbék illeszkedésétől független formulával való helyettesítésén alapszik, konkrétan az összevont klaszterekbe tartozó példák hasonlóságát az entrópiájuk alapján formalizálja. A példavektorok alapján egy K-dimenziós eloszlás entrópiája az alábbi módon becsülhető:

$$H(p) = \sum_{i=1}^{K} p(i) \log p(i). \tag{7}$$

A kiinduló állapotokat leíró eloszlások (azaz az y_s vektorok) becslésére pedig az adott állapotokhoz tartozó neuronháló-kimenetek átlagát fogjuk használni. Továbbmenve, állapotok egy $\mathcal{S}$ halmaza esetén a prototípus valószínűségi vektort ($y_{\mathcal{S}}$) a halmazba eső állapotok y_s vektorainak súlyozott átlagaként állítjuk elő, ahol a súlyozás a $N(s)$ példaszámok alapján történik. Maga a döntési kritérium pedig az entrópiafüggvény alapján formalizálható az alábbi módon:

$$D_E(\mathcal{S}) = -\sum_{k=1}^{K} N(\mathcal{S}) y_{\mathcal{S}}(k) \log y_{\mathcal{S}}(k). \tag{8}$$

4. DNN Flat Start

Az állapotklaszterezés előtt el kell végeznünk a CI címkék időbeli illesztését, majd a címkék osztályozására be kell tanítanunk a segéd-neuronhálót. Az illesztést elvégezhetjük hagyományos módon, egy HMM/GMM rendszer betanításával is. Bacchiani and Rybach azonban arra figyelmeztetnek, hogy a kapott illesztés nem feltétlenül lesz optimális a DNN tanítása céljára [11]. Ebből kifolyólag mi egy elegánsabb megoldást választottunk, amely nem igényli GMM-alapú modellek betanítását. A CI-NN neuronhálónkat 'flat start' módon tanítottuk be a Maximum Mutual Information (MMI) tanítási kritérium segítségével, Gosztolya és tsai korábbi cikkét követve [3]. A hatékonyság növelésére a tradicionális MMI tanítási eljáráson az alábbi módosításokat végeztük. Elsőként, a tanítási célvektorokat a HMM-ek forward-backward algoritmusával állítottuk elő. Másodsorban, amíg a szokványos eljárás az MMI-tanítást csak a tanítási folyamat végén, szószintű címkékkel végzi (lásd pl. [22,23]), mi puszta fonetikai címkesorral dolgoztunk, nyelvi modell nélkül. Ezeknek az egyszerűsítéseknek köszönhetően

a fonetikai címkék újraillesztése minden mondat (input fájl) után elvégezhető, gyors konvergenciát eredményezve. Ezzel a módszerrel a DNN tanítása nem csupán gyorsabb lett más módszerekhez képest (pl. [2]) hanem némileg jobb eredményt is adott.

5. Kísérleti beállítások

Mély neuronhálón alapuló felismerőnket a 81 órányi angol olvasott szöveget tartalmazó Wall Street Journal (WSJ) korpusz `si-284` jelű részhalmazán [24] tanítottuk be. A kiértékelést az `eval92` és `eval93` elnevezésű teszthalmazokon végeztük, az ún. „open-vocabulary" (azaz hatvanezer szót tartalmazó) szótárral, a szokványos szűkített trigram nyelvi modellel. Ezek közül fejlesztői (development) halmazként az `eval93` elnevezésű rész-korpusz szolgált, azaz ezen optimalizáltuk az olyan meta-paramétereket, mint a nyelvi modell súlya, a szó-beszúrási büntetés, illetve a kapcsolt állapotok száma. A tesztelés az ily módon optimalizált paraméterértékek mellett az `eval92` nevű részhalmazon történt.

A modellezéshez a saját belső fejlesztésű neuronhálós csomagunkat használtuk. Input jellemzővektorként 40 mel-skálás energiaegyüttható szolgált, az ezekhez tartozó első és második derivált-jellegű jellemzőkkel kiegészítve. A neuronháló bemenetét egy 15 szomszédos jellemzővektorból álló adatblokk képezte. Mind a CI segédháló, mind a végleges CD neuronháló öt rejtett réteget tartalmazott, rétegenként 1000-1000 neuronnal, az ún. ReLU aktivációs függvényt használva [25]. A kimenő rétegben a szokványos módon szoftmax aktivációt használtunk. A dekódolás és kiértékelés a HTK programcsomag neuronhálókhoz igazított változatával történt [26]. A klaszterező algoritmus négy változatát négy-négy paraméterbeállítás mellett futtattuk le, ezekben a klaszterezést vezérlő küszöböt úgy igyekeztünk belőni, hogy körülbelül 1800, 2400, 3000, illetve 3600 kapcsolt állapotot kapjunk.

A 1. ábra mutatja az állapotkapcsolás után előállt címkék darabszámának eloszlását a közelítőleg 2400 állapotot eredményező beállítások mellett. Amellett, hogy az eloszlás minden esetben nagyon hasonló, azt is megfigyelhetjük, hogy a leginkább kiegyensúlyozott eloszlást a Senior-féle megoldás (azaz a klasszikus klaszterező algoritmus DNN-kimeneten való lefuttatása) adta.

6. Eredmények és diszkusszió

Az 1. táblázat mutatja a development halmazon elért legjobb szóhiba-értékeket az egyes algoritmusokkal, valamint az ugyanazon paraméter-beállításokkal a teszthalmazon kapott hibát. Láthatóan a legegyszerűbb algoritmus működött a legroszabban, azaz a CI-DNN segédháló outputjain a klasszikus klaszterező algoritmust futtatva 8,7%-os szóhibát értünk el a development halmazon, míg a teszthalmazon 6,47%-ot. Ugyanazt a standard klaszterező algoritmust a háló legutolsó rejtett rétegének kimenetén futtatva (Bacchiani és Rybach cikkét követve) a development halmazon ugyan némileg nagyobb hibát kaptunk, a teszthalmazon viszont szignifikánsan javult az eredmény az előző módszerhez képest.

1. ábra. A tanító adatvektorok számának eloszlása a különböző klaszterező eljárásokkal kapott címkék esetén, kb. 2400 kapcsolt állapot mellett.

Klaszterezés inputja (DNN réteg)	Klaszterezés döntési kritériuma	WER	
		Dev.	Teszt
Kimeneti	Likelihood (Senior, [10])	8,70%	6,47%
Rejtett	Likelihood (Bacchiani, [11])	8,85%	6,04%
Kimeneti	KL-divergencia (Gosztolya, [12])	8,06%	5,72%
	Entrópia (Zhu, [14])	8,03%	5,92%

1. táblázat. A development halmazon elért legjobb szóhiba-arányok (WER) összegzése a hozzájuk tartozó teszthalmazon elért hibával, a különféle állapotkapcsoló algoritmusok esetén.

A további két tesztelt módszer nem a standard, Gauss-alapú klaszterezést alkalmazta, hanem egy új, DNN-ekhez igazított döntési kritériumot, és ez látszik is a hatékonyságukon. A development halmazon a két módszer gyakorlatilag megegyező szóhiba-arányt ért el (8,06% a Kullback-Leibler, illetve 8,03% az entrópia-alapú kritérium esetén); a teszhalmazon kicsit nagyobb különbséget mértünk, de ez még mindig nem statisztikailag szignifikáns. Összességében a Kullback-Leibler távolságon alapuló döntési kritériummal 0,8% abszolút hibacsökenést értünk el a Senior-féle triviális módszerhez képest, ami relatív skálán 12% hibacsökkenésnek felel meg.

A 2. és a 3. ábra mutatja a részletes eredményeket, azaz a négy algoritmus által a különböző állapotszámok esetére adott szóhiba-arányokat. Az ábrán szembetűnő módon szétválik a két standard hibakritériumot, valamit a két DNN-ekhez igazított hibakritériumot alkalmazó módszer, mivel az utóbbi csoport állapotszámtól függetlenül konzisztensen kisebb hibaértékeket ért el. Álláspontunk

2. ábra. A négy klaszterező algoritmus szóhibaaránya a development halmazon.

3. ábra. A négy klaszterező algoritmus szóhibaaránya a teszthalmazon.

szerint ez az eredmény nyilvánvaló módon azt tükrözi, hogy az állapotklaszterező algoritmus inputjának DNN-ekhez való igazítása mellett magát az algoritmust vezérlő döntési kritériumot is érdemes a neuronhálókhoz szabni, mivel így meggyőzően jobb eredményeket kaphatunk.

Az ábrákon látható görbék alapján a kapcsolt állapotok számának növelése javítja a klasszikus, likelihood-alapú klaszterezést használó módszerek teljesítményét. A két módosított kritériumot alkalmazó algoritmus esetén azonban az optimumpont valahol 2400 állapot környékén van a development halmazon. A teszthalmazon ezzel szemben nem lehet egyértelmű tendenciákat megfigyelni, mind a négy módszer relatíve érzéketlen az állapotszám növelésére. E téren

Klaszterezési algoritmus	WER	
	Dev.	Teszt
GMM + Likelihood	17,8%	17,3%
DNN + KL-divergencia	16,3%	15,9%
DNN + Entrópia	17,2%	16,3%

2. táblázat. A két legjobb módszer eredményei a magyar nyelvű korpuszon

meggyőző tanulságok levonásához valószínűleg még egy nagyságrenddel több tanítóadat lenne szükséges. Végezetül megjegyezzük, hogy a kapott optimális állapotszámok kisebbek, mint amekkorával a SWJ korpuszon rutinszerűen dolgozni szoktak. Ezért már magában a kisebb számításigény miatt is érdemesebb az általunk javasolt algoritmust használni.

7. Magyar nyelvű eredmények

A korábbiakban ismertetett kiértékelést az angol nyelvű Wall Street Journal beszédkorpuszon végeztük, egyrészt a nemzetközi publikálhatóság miatt, másrészt mivel nagyobb méretű, mint a rendelkezésünkre álló magyar adatbázisok. Magyar szerzőként azonban a magyar nyelvű beszédfelismerés fejlesztésében vagyunk érdekeltek, így a két legjobban teljesítő modell összevetését magyar nyelvű korpuszon is elvégeztük. E kísérletben a "Szeged" fantázianevű, 28 órányi híradófelvételt tartalmazó korpuszt használtuk. Mivel mind a korpusz, mind a modellezési paraméterek tökéletesen megegyeztek a korábbi publikációkban használatakkal, így az adatok és a rendszer részletes ismertetéstől eltekintük (megtalálható pl. [27]-ben).

A 2. táblázat mutatja a magyar nyelvű adatábázison kapott legjobb eredményeket a két módosított kritériumot alkalmazó (azaz a KL-divergencián, illetve az entrópián alapuló) algoritmus esetében. Viszonyítási pontként a GMM-alapon elvégzett címkeillesztést és állapotklaszterezést alkalmazó algoritmust tüntettük fel. Az eredményekből az látszik, hogy a KL-divergencián alapuló megoldás mind a development, mind a teszthalmazon némileg jobban teljesített, mint az entrópia-alapú kritérium, a relatív hibacsökkenés pedig a GMM-alapú technológiához képest kb. 8%.

8. Összegzés

Jelen cikkben négy, környezetfüggő állapokészlet kialakítására szolgáló klaszterező módszer teljesítményét hasonlítottuk össze mély neuronhálós környezetben. A négy algoritmus mindegyike egy segéd-neuronháló kimenetén (illetve rejtett rétegének kimenetén) végezte a klaszterezést. Kísérleteink azonban azt mutatták, hogy a klaszterezés bemenő jellemzői mellett érdemes az algoritmus döntési kritériumát is a neuronháló használatához igazítani, a módosított kritériummal

dolgozó algoritmus-változatok ugyanis minden esetben jobban teljesítettek, mint a hagyományos képletet használók. A kiértékelést az angol nyelvű Wall Street Journal beszédkorpuszon végeztük, de a legjobban teljesítő két módszert egy magyar nyelvű híradós beszédadatbázison is összevetettük. Mivel a kísérletek során a korábban javasolt MMI-alapú 'flat-start' tanítási módszert használtuk az inicializáláshoz, a Kullback-Leibler divergencián alapú megoldásunk sem az címkék illesztése, sem a környezetfüggő címkék kialakítása során nem igényli Gauss-görbék, azaz a "régi technológia" alkalmazását (a teljes GMM-mentes illesztési megoldás részletes ismertetése megtalálható [27]-ban).

Köszönetnyilvánítás

Tóth László munkáját az MTA Bolyai János Kutatási Ösztöndíja támogatta. Grósz Tamást az Emberi Erőforrások Minisztériuma UNKP-17-3 kódszámú Új Nemzeti Kiválóság Programja támogatta. A cikk elkészítéséhez használt Titan-X grafikus kártyát az NVIDIA Corporation adományozta.

Hivatkozások

1. Senior, A., Heigold, G., Bacchiani, M., Liao, H.: GMM-free DNN training. In: ICASSP. (2014)
2. Zhang, C., Woodland, P.: Standalone training of context-dependent Deep Neural Network acoustic models. In: ICASSP. (2014) 5597–5601
3. Gosztolya, G., Grósz, T., Tóth, L.: GMM-free flat start sequence-discriminative DNN training. In: Interspeech, San Francisco, USA (2016) 3409–3413
4. Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S.: Purely sequence-trained neural networks for ASR based on lattice-free MMI. In: Interspeech, San Francisco, USA (2016) 2751–2755
5. Bourlard, H., Morgan, N.: Connectionist Speech Recognition – A Hybrid Approach. Kluwer Academic (1994)
6. Yu, D., Deng, L., Dahl, G.: Roles of pretraining and fine-tuning in context-dependent DNN-HMMs for real-world speech recognition. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning. (2010)
7. Dahl, G., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained Deep Neural Networks for large vocabulary speech recognition. IEEE Trans. ASLP **20**(1) (2012) 30–42
8. Young, S.J., Odell, J.J., Woodland, P.C.: Tree-based state tying for high accuracy acoustic modelling. In: HLT. (1994) 307–312
9. Mohamed, A., Dahl, G.E., Hinton, G.: Acoustic modeling using Deep Belief Networks. IEEE Trans. ASLP **20**(1) (2012) 14–22
10. Senior, A., Heigold, G., Bacchiani, M., Liao, H.: GMM-free DNN acoustic model training. In: ICASSP. (2014) 5639–5643
11. Bacchiani, M., Rybach, D.: Context dependent state tying for speech recognition using deep neural network acoustic models. In: ICASSP. (2014) 230–234
12. Gosztolya, G., Grósz, T., Tóth, L., Imseng, D.: Building context-dependent DNN acousitc models using Kullback-Leibler divergence-based state tying. In: ICASSP, Brisbane, Ausztrália (2015) 4570–4574

13. Razavi, M., Rasipuram, R., Magimai-Doss, M.: On modeling context-dependent clustered states: Comparing HMM/GMM, hybrid HMM/ANN and KL-HMM approaches. In: ICASSP. (2014)
14. Zhu, L., Kilgour, K., Stüker, S., Waibel, A.: Gaussian free cluster tree construction using Deep Neural Network. In: Interspeech, Drezda, Németország (2015) 3254–3258
15. Wang, W., Tang, H., Livescu, K.: Triphone state-tying via Deep Canonical Correlation Analysis. In: Interspeech, San Francisco, USA (2016) 3444–3448
16. Beulen, K., Ney, H.: Automatic question generation for decision tree based state tying. In: ICASSP. (1998) 805–808
17. Odell, J.: The Use of Context in Large Vocabulary Speech Recognition. PhD thesis, University of Cambridge (1995)
18. Imseng, D., Dines, J.: Decision tree clustering for KL-HMM. Technical Report Idiap-Com-01-2012, Idiap Research Institute (2012)
19. Imseng, D., Dines, J., Motlicek, P., Garner, P., Bourlard, H.: Comparing different acoustic modeling techniques for multilingual boosting. In: Interspeech. (2012)
20. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Statist. **22**(1) (1951) 79–86
21. Imseng, D.: Multilingual speech recognition A posterior based approach. PhD thesis, École Polytechnique Fédérale de Lausanne (2013)
22. Kingsbury, B.: Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In: ICASSP. (2009) 3761–3764
23. Veselý, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: Interspeech. (2013) 2345–2349
24. Paul, D.B., Baker, J.M.: The design for the Wall Street Journal-based CSR corpus. In: HLT, Stroudsburg, PA, USA, Association for Computational Linguistics (1992) 357–362
25. Tóth, L.: Phone recognition with deep sparse rectifier neural networks. In: ICASSP. (2013) 6985–6989
26. Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, Cambridge, UK (2006)
27. Grósz, T., Gosztolya, G., Tóth, L.: Mély neuronhálós beszédfelismerők GMM-mentes tanítása. In: MSZNY, Szeged (2017) 170–180

Televíziós feliratok írásjeleinek visszaállítása rekurrens neurális hálózatokkal

Tündik Máté Ákos, Tarján Balázs, Szaszák György

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
e-mail:{tundik,tarjanb,szaszak}@tmit.bme.hu

Kivonat Az automatikus beszédfelismerő rendszerek (ASR) kimenete általában nem tartalmaz írásjeleket, pedig ezek döntően befolyásolják egy szöveg értelmét és értelmezhetőségét. Az írásjel-visszaállítás problémájára a közelmúltban eredményesen alkalmaztak rekurrens neurális hálókat (RNN). A valós idejű ASR kimenetekben (pl. televíziós feliratok) történő írásjel-visszaállítás azonban nagy kihívást jelent, mivel az alacsony késleltetés érdekében időben előre csak kis mértékben tekinthet a rendszer, illetve ASR-hibákkal is számolni kell, különösen informális vagy spontán beszéd esetén. Cikkünkben Maximum Entrópia (MaxEnt) és RNN írásjelező modelleket hasonlítunk össze magyar nyelvű, televíziós feliratok valós idejű és off-line feldolgozása során. A várakozásoknak megfelelően az RNN-alapú modellek teljesítménye felülmúlta a MaxEnt baseline rendszerét, illetve az ASR-hibák is jelentősen befolyásolták az írásjelek minőségét. Ezzel szemben az előretekintés mértékének korlátozása csak kisebb pontosságcsökkenést eredményezett. A hibák okainak feltárása érdekében műsortípusonként is kiértékeltük az írásjelező modellünket. Végezetül a modellünket egy angol nyelvű adatbázis segítségével más nemzetközi megoldásokkal is összehasonlítottuk.
Kulcsszavak: írásjel-visszaállítás, rekurrens neurális hálózatok, LSTM, maximum entrópia, ASR, feliratozás

1. Bevezetés

Az írásjelek fontossága nem csak abban rejlik, hogy az emberek számára olvashatóbbá teszik a szöveget, hanem a szövegfeldolgozó rendszerek is építenek rájuk a szintaktikai elemzés során [1,2]. Azaz hiányuk nem csak az emberi, de a gépi feldolgozást is megnehezíti. Az ASR-alapú diktáló rendszerek ezt a problémát általában úgy kerülik meg, hogy elvárják az írásjelek bemondását is, azonban könnyű belátni, hogy a beszélő ilyen fokú együttműködése nem várható el egy televíziós feliratozó-rendszer esetén.

Az írásjelek visszaállítása az ASR rendszerek kimenetén ismert probléma a beszédtechnológiában. Két alapvető megközelítést lehet megkülönböztetni: a prozódia alapút és a szövegalapút, bár gyakran kombinálva is használják őket. Általánosságban a prozódia alapú megközelítések számításigénye alacsonyabb és robusztusabbak az ASR hibákra, viszont a tanításukhoz szükséges címkézett,

akusztikai adatbázisok nehezebben hozzáférhetőek. Ezzel szemben a szövegalapú megközelítések többnyire pontosabb írásjelezést tesznek lehetővé, hála a nagy mennyiségű, aránylag könnyen hozzáférhető szövegkorpuszoknak, ugyanakkor érzékenyebbek az ASR-hibákra és több számítást igényelnek.

A korai írásjel-visszaállító megoldások az ASR-rendszerek n-gram nyelvi modelljébe épített rejtett eseményként tartalmazták az írásjeleket [3]. Ezek a modellek azonban hatalmas korpuszok bevonását igénylik, hogy csökkentsék az adatelégtelenségből fakadó problémát. Később megjelentek bonyolultabb szekvenciamodellezési megközelítések is: a transzdúcer-elvű keretrendszerek esetén egy írásjelezetlen szöveg kerül a bemenetre, melybe conditional random field (CRF), rejtett Markov-modellek (HMM), vagy maximum entrópia (MaxEnt) modellek segítségével helyezik el az írásjeleket [4,5]. A MaxEnt-alapú írásjelezési modellek lehetővé teszik a szöveges és prozódiai jellemzők hatékony kombinációját [6]. Egy átfogó tanulmányban [7] különböző jellemzőket hasonlítottak össze az írásjelezésre gyakorolt hatásuk tekintetében, MaxEnt-modellt használva. A legerősebb prediktív szöveges jellemzők közé a szóalakok és a szófaji (POS) címkék tartoztak, míg a legjobb prozódiai jellemző a szavak közötti szünetek időtartama volt. A Cho és társai által javasolt megoldás az írásjelezésre mint nyelven belüli fordítási feladatra tekint, írásjelezetlenből írásjeleket tartalmazó szószekvenciákat modellezve, mellyel jelentősen csökkentette az időkésleltetést [8].

A közelmúltban megjelentek a mély neurális hálózatokon alapuló megoldások: adott egy relatíve hosszú szó-kontextus, melyben a szavaknak egy beágyazási (Embedding) réteg révén átalakított alacsony dimenziójú vektor-reprezentációját vesszük, majd ezeket átvezetve kétirányú, rekurrens neurális hálózaton (RNN) [9] az írásjelek predikcióját megkapjuk mindegyik szóhoz. Az RNN-eket sokféle szekvenciacímkézési feladathoz alkalmazták már sikeresen, mivel képesek nagy kontextusok modellezésére, valamint a szavakból olyan jellemzőket tudnak kinyerni, amivel az adatelégtelenség problémája áthidalható. Tilk és Alumäe kétirányú RNN modelljében GRU [10] cellákat alkalmazva, ún. attention (figyelmi) mechanizmussal kiegészítve - mely segít még jobban fókuszálni az írásjelek szókörnyezetére közvetlen összeköttetések révén -, felülmúlta a korábbi észt és angol IWSLT eredményeket [11]. Egy nemrégiben készült tanulmányban [12] a kapitalizáció és az írásjelek helyreállítását egymással korreláló, többszörös szekvenciacímkézési feladatként kezelték, kétirányú RNN modellekkel. A [13]-ban a szerzők egy olyan prozódiai jellemzőket használó RNN-alapú központozási megoldást javasoltak, melynek hátterében a megnyilatkozások fonológiai frázisokra történő bontása és az elhelyezendő írásjelek közötti kapcsolat megállapítása állt.

A televíziós társaságok az ASR technológiát széles körben használják a feliratok készítéséhez, különösen az élő programokhoz, amelyeknek közel valós idejű feldolgozása kis késleltetést igényel [14]. Cikkünkben ehhez a feladathoz illeszkedő, alacsony késleltetésű, szövegalapú automatikus írásjelező modellek teljesítményét vizsgáljuk meg. Bemutatunk egy egyszerű felépítésű, RNN-alapú írásjel-visszaállító modellt, kétirányú LSTM cellákkal és a szóbeágyazást elvégző Embedding réteggel, majd összehasonlítjuk teljesítményét egy MaxEnt-alapú baseline megoldással. Különös figyelmet fordítunk az alacsony késleltetésű megoldá-

sokra. Mindkét rendszer teljesítményét automatikus és manuális átiratokon is kiértékeljük, megkülönböztetve az on-line és off-line működtetéshez alkalmas beállításokat. Magyar nyelvű televíziós műsorszórásból származó anyagokon kívül, az IWSLT angol nyelvű adathalmazán [15] elvégzett kísérleteink eredményeit is bemutatjuk, így összehasonlítva modellünk teljesítményét a legmodernebb rendszerekkel. A [13] által ismertetett tisztán prozódia alapú megközelítésen kívül nem ismerünk semmilyen előzetes munkát a magyar nyelvű beszédátiratokon történő írásjelek visszaállítására.

Cikkünk az alábbi struktúra szerint épül fel: először bemutatjuk az általunk használt adatbázisokat, illetve a kísérletekhez használt modelleket. Ezt követően áttérünk a magyar és angol nyelvű írásjelezési kísérletek ismertetésére és diszkussziójára. Végezetül a tanulságok levonása után néhány jövőbeni lehetséges kutatási irányt vázolunk fel.

2. Adatbázisok

2.1. Magyar nyelvű adatbázis

Az írásjel-visszaállítási kísérleteinkhez használt magyar nyelvű adatbázist a Médiaszolgáltatás-támogató és Vagyonkezelői Alap (MTVA) bocsátotta rendelkezésünkre. Az adatbázis különböző TV-műfajokhoz tartozó, kézzel készített feliratokat tartalmaz, amelyek lehetővé teszik számunkra, hogy különböző társalgási formákat tartalmazó műsorokon, például időjárás-előrejelzéseken, hírműsorokon, hírháttér-beszélgetéseken, magazinokon, sporthíreken és sportmagazinokon értékeljük ki az írásjelező megoldásunkat. A leggyakoribb és egyben a szöveg érthetősége szempontjából legfontosabb írásjeleket állítjuk vissza: a vesszőt, a mondatvégi pontot, a kérdőjelet és a felkiáltójelet. A kettőspontokat és a pontosvesszőket vesszővel helyettesítettük, minden más írásjeltől eltekintettünk. A tanítóanyag 20% -át validációs célból leválasztottuk, valamint egy külön, reprezentatív tesztkészletet használunk, amelynek nincs átfedése sem a tanító, sem a validációs halmazzal. Az adatokhoz kapcsolódóan további statisztikák a 1. táblázatban olvashatók.

1. táblázat. A magyar nyelvű adatbázis statisztikái

Műfajok	Tanító és validációs halmaz					Teszthalmaz					
	#Szavak	#Vessző	#Pont	#Kérdő	#Felki	#Szavak	#Vessző	#Pont	#Kérdő	#Felki	WER
Időjárás	478K	40K	31,5K	30	730	2,4K	250	200	0	20	6,8
Híradó	3493K	279K	223K	3,5K	4,6K	17K	1,5K	1K	20	50	10,1
Sporthírek	671K	55K	39,5K	280	2K	6K	500	400	2	30	21,4
Hírháttér	4161K	533K	225K	26,5K	4K	46,8K	6,3K	2,6K	250	130	24,7
Sportmagazin	-	-	-	-	-	22,7K	2K	1,4K	100	50	30,3
Magazin	4909K	732K	376K	72K	36K	10,4K	1,5K	700	150	70	38,7
Vegyes	1526K	187K	102K	11K	11,4K	30,7	4K	1,7K	280	150	-
Összesen	**15238K**	**1826K**	**997K**	**113K**	**58,8K**	**136K**	**16K**	**8K**	**800**	**500**	**24,2**

A kézzel készített feliratok mellett automatikus leiratokat is használtunk kísérleteinkben. Ezek egy, a televíziós műsorok élő feliratozásához optimalizált

ASR rendszerrel készültek [16]. Az ASR nyelvi modelljének tanítóhalmaza megegyezett az írásjelezési modellével, emellett hozzávetőlegesen 500 órányi beszédet használtunk fel az akusztikai modell tanításához. Az automatikus átiratok átlagos szóhibaaránya (WER) körülbelül 24% volt, azonban a műfajtól függően nagy változatosságot mutatott (lásd 1. táblázat). A "Vegyes" kategóriában nem állt rendelkezésre hanganyag az adatbázisban.

2.2. Angol nyelvű adatbázis

Az IWSLT adathalmaz angol nyelvű TED előadások átirataiból áll, és a közelmúltban az angol írásjel-visszaállítási kísérletek egyik népszerű feladatává vált [11,12,15,17]. Kísérleteink során ugyanazt a tanító, validációs és teszthalmazt használtuk, mint a fenti tanulmányok. Ezek rendre 2,1M, 296K és 13K szót tartalmaznak. Az adatbázis csak vesszőt, pontot és kérdőjelet tartalmaz, így csak ezek visszaállítására nyújt lehetőséget.

3. Írásjel-visszaállító módszerek

3.1. Maximum Entrópia (MaxEnt) Modell

A maximum entrópia (MaxEnt) modellt eredetileg Ratnaparkhi javasolta a szófaji címkézésre [18]. Minden mondatot tokenek (szavak) szekvenciájaként írunk le. Minden egyes tokenhez egy sor egyedi jellemző rendelhető, melyek segítenek a kimeneti címkék (jelen esetben az írásjelek) meghatározásában. Lévén, hogy ez egy felügyelt tanítási technika, a kimeneti címkék a token-sorozathoz hozzá vannak rendelve. A jellemzők meghatározásához a MaxEnt modell meghatároz a rendelkezésre álló címkék és az aktuális kontextus között egy együttes eloszlást, amely egy rádiusz paraméterrel szabályozható.

Maxent írásjelezési modellünkben csak a kisbetűs szóalakokat használtuk fel jellemzőként mind magyar, mind angol nyelven. Ehhez a Huntag nyílt forráskódú, nyelvfüggetlen, Markov modell-alapú, MaxEnt szekvenciacímkéző programot használtuk fel [19].

Mint említettük, a MaxEnt modell rádiusz paramétere határozza meg a figyelembe vett kontextus méretét. Alapértelmezés szerint a múltbeli és a jövőbeli kontextust egyenlő mértékben veszi figyelembe. Erre a beállításra *off-line módként* fogunk hivatkozni. A jövőbeli kontextus figyelembevétele azonban a késleltetést növeli; ezt limitálva, az *on-line módú* modellek teljesítményét is kiértékeltük. Cikkünkben az alábbi, kerek zárójeles jelöléssel utalunk a múltbeli és jövőbeli kontextus méretére; (5,1) azt jelenti, hogy valójában az adott modellben maximum 5 múltbeli és 1 jövőbeli tokent veszünk figyelembe az adott szóhoz tartozó írásjel predikciójához.

3.2. Rekurrens neurális háló

A tanító, a validációs és a teszthalmazt rövid, fix hosszúságú szekvenciákra osztjuk fel, köztes átfedések nélkül. A különböző szóalakok számát limitáljuk, a tanítóhalmaz k-leggyakoribb szavából szótárt képezve, a kieső szavakat pedig egy

közös *Ismeretlen* címkével látjuk el. A modellhez saját szóbeágyazási mátrixot képzünk az előre tanított beágyazási modell és a szótárban szereplő szavak segítségével.

Kísérleteinkben egy egyirányú és egy kétirányú RNN modell teljesítményét vizsgáljuk meg. A modellben az aktuális szót megelőző időpillanatra jósoljuk az írásjelet. A kísérletekhez használt RNN-architektúrákat az 1. ábrán mutatjuk be.

Az RNN-modellek (WE-LSTM és WE-BiLSTM, a Word Embedding (szóbeágyazás) rövidítéséből) a következőképpen épülnek fel: a szóbeágyazási mátrix alapján a modellnek átadott szószekvenciák a szóbeágyazási térbe (x_t reprezentálja az x szóhoz tartozó n-dimenziós szóbeágyazási vektort t időpillanatban). Ezek a reprezentációk a következő, rejtett rétegbe kerülnek, amely LSTM vagy BiLSTM rejtett cellákból áll, melyek a x_t kontextus rögzítéséért, az információ kinyeréséért felelősek. A kimenetet egy $softmax$ aktivációs függvény használata után kapjuk meg, mely az y_t kimeneti címkék eloszlását a jelenlegi szó x_t előtti időpillanatra (slot-ra) adja meg. Láthatjuk, hogy ez egy egyszerű felépítésű modell, ezáltal alacsony késleltetést, valós idejű működést tesz lehetővé.

1. ábra: A WE-BiLSTM (bal oldalon) és a WE-LSTM (jobb oldalon) RNN modell szerkezete

A magyar írásjelező modelleket a tanítókorpusz 100 000 leggyakoribb szavával tanítottuk, a kimaradt szavakhoz egy közös "*Ismeretlen*" szimbólumot rendeltünk. Az RNN-alapú írásjel-helyreállítási modellekhez egy 600 dimenziós, előre tanított magyar nyelvű szóbeágyazási modellt használtunk [20]. Angol nyelvű RNN-modelljeinkben egy 100 dimenziós, népszerű szóbeágyazási modellt használtunk, a "GloVe"-ot. A tanítás során az RNN modell súlyait a kategorikus keresztentrópia költségfüggvény alapján módosítjuk, valamint minden egyes epoch-ban frissítjük a szóbeágyazásokat is.

Szisztematikus, kimerítő keresés (grid search) alapú optimizációt hajtottunk végre az RNN-ek hiperparaméterein, a validációs halmaz elemeit értékelve. A szekvenciák hosszát, a szótár méretét, a rejtett állapotok számát, a mini-batch méretét, és az optimizáló típusát változtattuk. Korai leállítást (early stopping, *Patience*) is használunk a túltanítás elkerülése érdekében. A 2. táblázat összefoglalja a magyar és az angol WE-BiLSTM és WE-LSTM modellekben használt

hiperparaméterek végső értékeit, beleértve azokat is, amelyeket a [11] cikkből vettünk át, a minél jobb összehasonlíthatóság érdekében.

2. táblázat. A WE-BiLSTM és WE-LSTM modellek hiperparaméterei

Nyelv	Model	Szekv. Hossza (#szavak)	Szótár Mérete (#szavak)	Szó-beágyazás Dim.	#Rejtett állapotok	Batch mérete	Optimalizáló	*Patience*
HUN	WE-BiLSTM	200	100 000	600	512	128	RMSProp	3
HUN	WE-LSTM				256			2
EN	WE-BiLSTM	200	27 244	100	256			2
EN	WE-LSTM	250	([11] révén)	([11] révén)				

A MaxEnt-hez hasonlóan az RNN modellnél is megkülönböztetjük az alacsony késleltetéshez adaptált on-line módot és a robusztus off-line módot, attól függően, hogy figyelembe vesszük-e a jövőbeli kontextust. Az RNN-alapú írásjelező rendszerek implementálásához a Keras keretrendszert [21] használtuk, a tanítást GPU-n végeztük el. Az RNN-modellek forráskódja nyilvánosan elérhető [1].

Vizsgálataink előkészítő fázisában a szóalakok mellett más szöveges jellemzők használatát is fontolóra vettük (lemmák, szófaji címkék (amit [22] is javasolt) és morfológiai elemzésből származó címkék), ezekkel azonban nem értünk el jobb eredményeket, így további kísérleteink során csak a szóalakokra támaszkodtunk.

4. Kísérleti eredmények

A következő fejezetben bemutatjuk a magyar és angol nyelvű írásjelezési kísérleteink eredményeit. A kiértékeléshez standard információ-visszakeresési mutatókat használtunk: Pontosság (Pr), Felidézés (Rc) és F1-érték (F1). Ezenkívül megadtuk a Slot Error Rate (SER) [23] értéket is, amely egy metrikában egyszerre tükrözi az írásjel-visszaállításhoz kapcsolódó hibák minden lehetséges típusát - beszúrásokat (Ins), helyettesítéseket (Sub) és törléseket (Del):

$$SER = \frac{C(Ins) + C(Subs) + C(Del)}{C(slotok_szama)}, \tag{1}$$

ahol $C(.)$ a számláló operátor, a slot-ok pedig a szavak utáni helyek a szövegben, ahova írásjelet helyezhetnek a modellek.

4.1. Magyar eredmények

Először összehasonlítjuk a MaxEnt szekvenciacímkéző baseline rendszer teljesítményét (lásd 3.1 alfejezet) az RNN-alapú írásjelező rendszerével (lásd 3.2. alfejezet), a magyar adathalmazon, két különböző konfigurációban. Az *on-line*

[1] https://github.com/tundik/HuPP

módban az aktuális szó előtti időpillanat írásjel predikciója jelenik meg, ez még épp akkora késleltetést eredményez, amely még valós idejű alkalmazáshoz elfogadható. A legjobb írásjelezési eredményének elérése érdekében az ún. *off-line módban* a jövőbeli kontextust is felhasználjuk.

Az eredményeket két táblázatban mutatjuk be: a kézi úton előállított feliratokon végzett írásjel-visszaállítás eredményei a 3. táblázatban, az automatikus (ASR) feliratokon végzett visszaállítás eredményei a 4. táblában találhatók. A MaxEnt modellek (i, j) jelölésében i a múltbeli, míg j a jövőbeli kontextust kontrolláló rádiusz paraméter értékét jelöli. Amint láthatjuk, a vesszők predikciója minden módszer és konfiguráció esetében kiemelkedik a többi írásjelé közül. Erre az lehet a legvalószínűbb magyarázat, hogy magyar nyelven a vesszők nagy aránya megbízhatóan becsülhető a slot-ot követő szó alapján (pl. 'hogy'). Ezzel ellentétben a mondatvégi pontok megbízható visszaállításához elengedhetetlennek tűnik az akusztikus jellemzők figyelembevétele is [13].

3. táblázat. Írásjel-visszaállítási eredmények magyar nyelvű, manuális feliratokon

Referencia-átirat	**Model**	**Vessző**			**Pont**			**Kérdőjel**			**Felkiáltójel**			**SER**
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mód	MaxEnt-(19,19)	72,5	59,6	65,5	52,1	40,0	45,2	**55,7**	21,8	31,3	31,1	31,5	31,3	63,5
	WE-BiLSTM	**72,9**	**71,2**	**72,0**	**59,1**	**56,1**	**57,6**	52,4	**38,7**	**44,5**	**51,3**	**36,1**	**42,4**	**50,1**
On-line mód	MaxEnt-(25,1)	71,8	58,1	64,2	47,5	35,7	40,8	50,4	16,2	24,5	29,3	**33,3**	31,2	66,9
	WE-LSTM	**72,7**	**69,5**	**71,1**	**56,2**	**48,3**	**52,0**	**60,4**	**31,1**	**41,1**	**61,1**	29,4	**39,7**	**53,6**

Ahogy a 3. táblázatban látható, a magyar nyelvű, manuálisan készült feliratok halmazán közel 20% relatív hibacsökkenés (SER) érhető el az RNN-alapú írásjelező modellel a baseline MaxEnt rendszerhez képest. A WE-BiLSTM és a WE-LSTM modellek különösen jól teljesítenek a pontok, kérdőjelek és felkiáltójelek helyreállításában, mivel a nagy kontextusokban rejlő információt sokkal hatékonyabban képesek kihasználni, mint a MaxEnt címkézők. A jövőbeni kontextus korlátozása az on-line konfigurációban sokkal kisebb mértékben befolyásolta az írásjelezés eredményességét, mint amire számítottunk. A jövőbeni szószekvencia-elemekből kinyert információk főként akkor hasznosak, ha a feladat megköveteli a felidézés maximalizálását, egyébként a pontosság tekintetében a WE-LSTM is alkalmas az írásjelek helyreállítására.

4. táblázat. Írásjel-visszaállítási eredmények magyar nyelvű, gépi feliratokon

ASR átirat	**Model**	**Vessző**			**Pont**			**Kérdőjel**			**Felkiáltójel**			**SER**
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mód	MaxEnt-(19,19)	**64,5**	55,8	59,9	41,1	31,2	35,6	**41,2**	8,8	14,4	48,8	17,1	25,4	79,2
	WE-BiLSTM	63,9	**67,7**	**65,7**	**50,5**	**49,0**	**49,8**	37,7	**24,1**	**29,4**	**60,9**	**24,0**	**34,4**	**70,1**
On-line mód	MaxEnt-(25,1)	**64,3**	54,9	59,2	38,9	29,4	33,5	36,0	7,1	11,9	47,1	20,6	28,6	81,3
	WE-LSTM	63,8	**65,1**	**64,4**	**47,8**	**42,0**	**44,7**	**48,5**	**20,5**	**28,9**	**61,8**	**21,7**	**32,1**	**73,1**

Amint azt a bevezetőben már vázoltuk, mind az ASR-hibák, mind a jövőbeli kontextus korlátozása azon tényezők közé tartoznak, amelyek nehezítik az írásjelek hatékony elhelyezését az élő TV-műsorok feliratait. Az eredményeink azt mutatják, hogy a két faktor közül a jövőbeni kontextus kevésbé fontos az írásjelek robusztus helyreállításához, ami ellentmond a várakozásainknak. Ezzel szemben az ASR-hibák jobban összefüggnek az írásjelek hibáival: a manuálisról a gépi úton készült feliratokra való áttérés során a SER értéke 15-20%-kal növekedett (lásd a 4. táblázatot). Habár ezzel párhuzamosan a MaxEnt és RNN modellek közötti különbség is csökkent az ASR feliratok feldolgozásakor, de az RNN még mindig jelentősen felülmúlta a baseline rendszert.

4.2. Műfaji analízis

A magyar nyelvű adatbázis teszthalmazának feliratait 6 műfaji kategóriába lehet osztani (lásd 1. táblázat). Ebben az alfejezetben az egyes műfajokon mérhető írásjel-helyreállítási pontosságokat hasonlítjuk össze, feltételezve, hogy az informálisabb, spontánabb műfajok írásjelezése nehezebb feladat, hiszen ezek esetében magasabb az ASR-hibák aránya (WER). Az egyes műfajokra vonatkozó írásjelek egy részét nem értékeltük (lásd az "N/A" feliratokat a 1. táblázatban), ha a felidézést vagy pontosságot nem lehetett meghatározni a tévesztési mátrix alapján.

Mivel az RNN-alapú rendszer minden műfajt tekintve meghaladta a MaxEnt rendszer teljesítményét, a könnyebb olvashatóság érdekében úgy döntöttünk, hogy csak a WE-BiLSTM és WE-LSTM rendszerek eredményeit adjuk meg a 5. és 6. táblázatban.

5. táblázat. Műfajonkénti írásjelezési eredmények magyar nyelvű referencia átiratokon

Referencia átirat	**Műfaj**	**Vessző**			**Pont**			**Kérdőjel**			**Felkiáltójel**			**SER**
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
RNN Off-line mód	Időjárás	61,2	54,3	57,5	46,7	46,7	46,7	N/A	N/A	N/A	**90,0**	45,0	60,0	69,3
	Híradó	**89,9**	**84,4**	**87,1**	**84,3**	**90,7**	**87,3**	**91,7**	**50,0**	**64,7**	83,9	**56,5**	**67,5**	**20,0**
	Sporthír	68,3	60,6	64,2	49,4	51,4	50,4	N/A	N/A	N/A	75,0	30,0	42,9	67,0
	Hírháttér	80,4	74,5	77,3	63,9	64,9	64,4	63,0	46,4	53,5	88,9	18,5	30,6	38,7
	Sportmagazin	61,2	61,1	61,1	43,9	49,3	46,5	55,2	37,5	44,7	38,5	9,4	15,2	73,1
	Magazin	67,6	67,6	67,6	45,1	46,3	45,7	50,5	29,7	37,5	50,0	5,6	10,1	58,6
RNN On-line mód	Időjárás	60,2	57,5	58,8	45,7	37,9	41,4	N/A	N/A	N/A	87,5	35,0	50,0	70,6
	Híradó	**88,4**	**83,1**	**85,7**	**86,6**	**81,3**	**83,9**	**75,0**	40,9	**52,9**	**100,0**	**67,4**	**80,5**	**24,1**
	Sporthírek	68,7	57,2	62,4	42,4	37,5	39,8	N/A	N/A	N/A	90,0	60,0	72,0	74,2
	Hírháttér	80,1	74,0	76,9	66,7	54,8	60,1	63,0	**45,6**	**52,9**	77,6	29,2	42,5	40,8
	Sportmagazin	60,8	59,7	60,3	42,3	34,8	38,2	53,3	38,3	44,5	20,0	7,5	11,0	77,3
	Magazin	67,6	65,1	66,3	43,5	32,8	37,4	57,3	27,2	36,9	36,4	11,3	17,2	61,5

Ha összevetjük az eredményeket a 1. táblázat statisztikáival, láthatjuk, hogy az írásjelek helyreállítása a híradó, hírháttér és magazin műfajok esetén sikerült a legjobban. Ez nem meglepő, hiszen ezekre a műfajokra rendelkezünk a

legtöbb tanítóanyaggal. Azonban a viszonylag nagy különbség a három jól modellezett műfaj között azt sugallja, hogy ezen kívül más tényezők is befolyásolják az eredményeket. Véleményünk szerint ezek közül kiemelkedik az adott szöveg tervezettsége. A nyelvi modellezéshez hasonlóan, minél formálisabb a beszédstílus a felvételeken, annál hatékonyabb az írásjelek elhelyezése is, pl. a híradók esetében. Nyilvánvaló, hogy a társalgási (pl. hírháttér) és az informális (magazin) beszédstílusok sajátosságai (melyekben a megakadásjelenségek és az alapvető grammatikai szabályokkal nonkonform kifejezésmód aránya magasabb) még inkább megnehezítik a predikciót, és a formálisabb stílusokhoz képest több írásjelezési hibát vezetnek be a rendszerbe.

6. táblázat. Műfajonkénti írásjelezési eredmények magyar ASR átiratokon

ASR átirat	Műfaj	Vessző			Pont			Kérdőjel			Felkiáltójel			SER	WER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1		
RNN Off-line mód	Időjárás	57,8	54,3	55,9	48,1	43,8	45,8	N/A	N/A	N/A	**100,0**	**60,0**	**75,0**	74,3	6,8
	Híradó	65,7	**82,2**	**73,0**	**63,5**	**82,7**	**71,8**	**52,9**	**40,9**	**46,2**	75,0	45,7	56,8	**57,4**	10,1
	Sporthírek	53,3	55,9	54,6	37,9	41,1	39,5	N/A	N/A	N/A	88,2	53,6	66,7	93,1	21,4
	Hírháttér	**70,4**	68,8	69,6	55,6	49,4	52,3	44,1	28,4	34,5	70,8	13,1	22,1	59,6	24,7
	Sportmagazin	52,8	58,0	55,3	37,5	41,1	39,2	42,9	18,8	26,1	N/A	N/A	N/A	93,2	30,3
	Magazin	59,6	59,9	59,7	34,6	29,1	31,9	13,9	19,4	16,2	16,7	1,4	2,6	82,3	38,7
RNN On-line mód	Időjárás	61,4	57,9	59,6	43,1	39,1	41,0	N/A	N/A	N/A	**88,9**	40,0	55,2	73,2	6,8
	Híradó	64,8	**80,6**	**71,8**	**62,4**	**73,5**	**67,5**	40,0	9,1	14,8	75,0	45,7	56,8	**61,8**	10,1
	Sporthírek	52,8	54,0	53,4	35,4	34,6	35,0	N/A	N/A	N/A	83,3	**53,6**	**65,2**	96,7	21,4
	Hírháttér	**70,2**	67,1	68,6	53,5	40,9	46,3	**46,4**	**25,6**	**33,0**	69,2	13,8	23,1	62,4	24,7
	Sportmagazin	53,1	55,4	54,2	35,6	30,9	33,1	41,4	22,7	29,3	16,7	5,7	8,5	94,0	30,3
	Magazin	58,5	59,9	59,2	36,0	22,3	27,6	46,3	12,0	19,1	42,9	4,2	7,7	83,2	38,7

Az időjárás-előrejelzés és a sportprogramok esetén a viszonylag magas SER érték jelzi annak fontosságát, hogy a tanítóhalmaznak elegendő mennyiségű témaspecifikus adatot kell tartalmaznia. Az ilyen alultanított témakörök esetén további tanítóanyagok bevonásával vagy adaptációs technikák alkalmazásával lehetne jobb eredményeket elérni.

Összehasonlítva a manuális és gépi feliratok írásjelezési hibáit, néhány érdekes következtetést vonhatunk le. A jól modellezett műfajok esetében a SER növekedése korrelál az ASR átirat szóhibaarányával (WER). Azonban a többi műfaj (időjárás-jelentés, sporthírek, sportmagazin) a SER és a WER között ilyen összefüggés nem fedezhető fel. A sporthírek viszonylag gyenge eredményeire külön kitérnénk. Míg az ASR átirata mérsékelt szóhibaarányt (24,7%) mutat, az írásjelezéshez kapcsolódó SER érték majdnem másfélszeresére nőtt (67% -ról 93-97%-ra). Feltételezzük, hogy ez a jelenség a sportprogramokban szereplő névelemek (Named Entity) megnövekedett mennyiségével függ össze, tekintve, hogy a legmagasabb OOV-arány (10%) itt figyelhető meg a teszthalmazban szereplő 6 műfaj közül.

4.3. Angol nyelvű eredmények

Ebben az alfejezetben néhány, a közelmúltban publikált írásjel-visszaállító megoldással hasonlítjuk össze a modellünket. Ebből a célból az IWSLT2011 adatbázist használjuk, amely TED előadások feliratait tartalmazza, és az angol nyelvű írásjelezési kísérleteknél egyfajta benchmark-nak tekinthető. A precízebb összehasonlíthatóság érdekében az alapértelmezett tanító, validációs és teszthalmazt használtuk, azonban a hiperparamétereket erre a feladatra újra optimizáltuk a validációs halmazon. Kiemeljük, hogy az IWSLT adatbázis nem tartalmaz mintákat felkiáltójelekhez.

Az angol írásjelezési eredményeket a 7. és 8. táblázatokban mutatjuk be. Amint látható, on-line módban a bemutatott egyirányú RNN modellünk (WE-LSTM) jelentősen felülmúlta a [11]-ben megjelenő ún. T-LSTM konfigurációt, amely az eddigi legjobb on-line eredményeket szolgáltatta ezen az adathalmazon legjobb tudomásunk szerint. Ha nem használunk előre tanított szóbeágyazási modellt (noWE-LSTM), eredményünk nagyon közel áll a T-LSTM konfigurációhoz.

7. táblázat. Írásjelezési eredmények angol nyelvű referencia átiratokon

Referencia átirat	Model	Vessző			Pont			Kérdőjel			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mód	MaxEnt-(6,6)	45,6	26,7	33,7	59,4	57,0	58,2	52,4	23,9	32,8	77,2
	WE-BiLSTM	55,5	45,1	49,8	65,9	**75,1**	70,2	57,1	52,2	54,5	59,8
	T-BRNN-pre [11]	**65,5**	47,1	54,8	73,3	72,5	72,9	**70,7**	**63,0**	**66,7**	**49,7**
	Corr-BiRNN [12]	60,9	**52,4**	**56,4**	**75,3**	70,8	**73,0**	70,7	56,9	63,0	50,8
On-line mód	MaxEnt-(10,1)	44,9	23,7	31,0	53,4	50,1	51,7	50,0	21,7	30,8	83,2
	noWE-LSTM	47,3	**42,7**	44,9	60,9	50,4	55,2	**68,2**	32,6	44,1	76,4
	WE-LSTM	**56,3**	40,3	**47,0**	**61,2**	**60,5**	**60,8**	55,5	**43,5**	48,8	**68,1**
	T-LSTM [17]	49,6	41,1	45,1	60,2	53,4	56,6	57,1	43,5	**49,4**	74,0

8. táblázat. Írásjelezési eredmények angol nyelvű ASR átiratokon

ASR Átirat	Model	Vessző			Pont			Kérdőjel			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mód	MaxEnt-(6,6)	40,6	23,9	30,1	56,2	53,5	54,8	31,6	17,1	22,2	84,0
	WE-BiLSTM	46,8	39,6	42,9	60,7	70,3	65,1	44,4	45,7	45,0	72,5
	T-BRNN-pre [11]	**59,6**	42,9	49,9	**70,7**	**72,0**	**71,4**	60,7	48,6	54,0	**57,0**
	Corr-BiRNN [12]	53,5	**52,5**	**53,0**	63,7	68,7	66,2	**66,7**	**50,0**	**57,1**	65,4
On-line mód	MaxEnt-(10,1)	42,6	23,9	30,7	53,2	48,9	51,0	33,3	17,1	23,0	87,0
	noWE-LSTM	40,2	**39,3**	39,7	56,2	46,6	51,0	**76,5**	38,2	**51,0**	86,5
	WE-LSTM	**48,8**	37,1	**42,2**	**57,6**	**57,3**	**57,4**	41,2	41,2	41,2	**78,3**
	T-LSTM [17]	41,8	37,8	39,7	56,4	49,3	52,6	55,6	**42,9**	48,4	83,7

Bár ebben a cikkben elsősorban az egyszerű, alacsony késleltetésű írásjelező modellek bemutatására fókuszáltunk, a WE-BiLSTM rendszert a jelenlegi state-of-the-art off-line megoldásokkal is összevetettük. A 7. és 8. táblázatokban látható, hogy mind a [11]-féle T-BRNN-pre modell, mind a [12]-féle Corr-BiRNN felülmúlta a WE-BiLSTM teljesítményét, főként a vesszők és a kérdőjelek beszúrását illetően. Az idézett írásjel-helyreállító rendszerek azonban sokkal összetettebb struktúrával rendelkeznek, és átlagosnak mondható kiépítettségű infrastruktúrán nem képesek valós időben működni, illetve bizonyos feladatokban elfogadhatatlan késleltetést visznek a rendszerbe. A WE-BiLSTM modellek esetén a pontok magas felidézési értékét a manuális és ASR feliratokban kedvező eredménynek tekintjük.

5. Összegzés

Cikkünkben bemutattunk egy alacsony késleltetésű, RNN-alapú, írásjel-visszaállító rendszert, amelynek a teljesítményét magyar és angol nyelvű adatbázisokon is kiértékeltük, valamint összehasonlítottuk egy Maximum Entrópia-alapú szekvenciacímkéző rendszerrel. Mindkét modellezési módszert kiértékeltük on-line módban, ahol a valós idejű működés lehetővé tétele érdekében csak a múltbéli szöveges jellemzők alapján hoztunk döntést; valamint off-line módban is, ahol mind a múltbeli, mind a jövőbeli jellemzőket figyelembe vettük. Az RNN-alapú megközelítés mindegyik tesztkonfigurációban jelentősen felülmúlta a MaxEnt baseline rendszer teljesítményét. Meglepő módon, azonban az on-line mód csak kis mértékben csökkentette az írásjelek helyreállításának pontosságát.

A magyar nyelvű felirat-adatbázisban a különböző televíziós műfajokon mérhető eredményeket összehasonlítva azt találtuk, hogy a szövegalapú modellekben az írásjelek helyreállításának pontossága (a nyelvi modellezéshez hasonlóan) a rendelkezésre álló tanítóadatok mennyiségétől és az adott feladat tervezettségétől függ. Megjegyezzük, hogy a magyar nyelvű beszédátiratok szövegalapú írásjel-helyreállításának témakörében nem ismerünk korábbi munkát.

Annak érdekében, hogy összehasonlítsuk modelljeinket a state-of-the-art RNN írásjelező rendszerekkel, angol nyelvű kísérleteket végeztünk az IWSLT adatbázison, mind on-line, mind az off-line üzemmódban. On-line írásjel-visszaállításban a mi WE-LSTM rendszerünk érte el a legjobb eredményt. Off-line módban a komplexebb megoldások természetesen felülmúlták a mi, elsősorban valós idejű működésre tervezett, egyszerű felépítésű megoldásunkat.

Rendszerünk jövőbeni továbbfejlesztésének egyik legfontosabb iránya, hogy egyesítjük a szöveges jellemzőket használó jelenlegi megoldásunkat a [13]-ban bemutatott prozódia alapú modellel. Az angol nyelvű modell további szöveges vagy akusztikai jellemzőkkel való kiterjesztése szintén ígéretes irány, miközben mindkét nyelven az alacsony késleltetést is szem előtt tartjuk.

Mindent összevetve munkánk egyik legnagyobb eredményének azt tekintjük, hogy sikerült létrehoznunk egy könnyű és gyors RNN modellt alacsony késleltetést megkövetelő írásjel-visszaállítási alkalmazások (pl. TV-műsorok feliratozása) támogatására. Ezen kívül fontos megemlíteni, hogy mindezt az agglutináló ma-

gyar nyelv esetére tettük meg, amely esetén sokkal kevésbé korlátozott a szórend, mint az angol nyelvben. A grammatikai funkciók sokkal kevésbé függnek a szórendtől, mint az utótagoktól (pl. esetvégződésektől), így az adatokban fellépő magasabb fokú változatosság is nehezebbé teszi a szekvencia modellezést.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatalnak, amely az FK-124413 projekt keretében a cikkben ismertetésre került kutatást támogatta. Köszönjük továbbá a Pro Progressio Alapítvány (Tarján Balázs), valamint az NVIDIA támogatását (GPU biztosítása az RNN tanításokhoz).

Hivatkozások

1. Batista, F., Moniz, H., Trancoso, I., Mamede, N.: Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. IEEE Trans. on Audio, Speech, and Language Processing **20**(2) (2012) 474–485
2. Paulik, M., Rao, S., Lane, I., Vogel, S., Schultz, T.: Sentence segmentation and punctuation recovery for spoken language translation. In: Proceedings of ICASSP, IEEE (2008) 5105–5108
3. Gravano, A., Jansche, M., Bacchiani, M.: Restoring punctuation and capitalization in transcribed speech. In: Proceedings of ICASSP, IEEE (2009) 4741–4744
4. Beeferman, D., Berger, A., Lafferty, J.: Cyberpunc: A lightweight punctuation annotation system for speech. In: Proceedings of ICASSP, IEEE (1998) 689–692
5. Lu, W., Ng, H.T.: Better punctuation prediction with dynamic conditional random fields. In: Proceedings of EMNLP, ACL (2010) 177–186
6. Huang, J., Zweig, G.: Maximum entropy model for punctuation annotation from speech. In: Proceedings of Interspeech. (2002) 917–920
7. Batista, F.: Recovering Capitalization and Punctuation Marks on Speech Transcriptions. PhD thesis, Instituto Superior Técnico (2011)
8. Cho, E., Niehues, J., Kilgour, K., Waibel, A.: Punctuation insertion for real-time spoken language translation. In: Proceedings of the Eleventh International Workshop on Spoken Language Translation. (2015)
9. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. on Signal Processing **45**(11) (1997) 2673–2681
10. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
11. Tilk, O., Alumäe, T.: Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: Proceedings of Interspeech. (2016) 3047–3051
12. Pahuja, V., Laha, A., Mirkin, S., Raykar, V., Kotlerman, L., Lev, G.: Joint learning of correlated sequence labelling tasks using bidirectional recurrent neural networks. arXiv preprint arXiv:1703.04650 (2017)
13. Moró, A., Szaszák, Gy.: A phonological phrase sequence modelling approach for resource efficient and robust real-time punctuation recovery. In: Proceedings of Interspeech. (2017)

14. Renals, S., Simpson, M., Bell, P., Barrett, J.: Just-in-time prepared captioning for live transmissions. In: Proceedings of IBC 2016. (2016)
15. Che, X., Wang, C., Yang, H., Meinel, C.: Punctuation prediction for unsegmented transcript based on word vector. In: Proceedings of LREC. (2016) 654–658
16. Tarján, B., Varga, Á., Tobler, Z., Szaszák, Gy., Fegyó, T., Bordás, Cs., Mihajlik, P.: Magyar nyelvű, élő közéleti- és hírműsorok gépi feliratozása. In: XII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2016, Szeged (2016) 89–99
17. Tilk, O., Alumäe, T.: LSTM for punctuation restoration in speech transcripts. In: Proceedings of Interspeech. (2015) 683–687
18. Ratnaparkhi, A., et al.: A maximum entropy model for part-of-speech tagging. In: Proceedings of EMNLP. (1996) 133–142
19. Recski, G., Varga, D.: A Hungarian NP chunker. The Odd Yearbook **8** (2009) 87–93
20. Makrai, M.: Filtering wiktionary triangles by linear mapping between distributed models. In: Proceedings of LREC. (2016) 2776–2770
21. Chollet, F.: Keras: Theano-based deep learning library. Code: https://github.com/fchollet. Documentation: http://keras.io (2015)
22. Ueffing, N., Bisani, M., Vozila, P.: Improved models for automatic punctuation prediction for spoken and written text. In: Proceedings of Interspeech. (2013) 3097–3101
23. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop. (1999) 249–252

Kísérletek az alapfrekvencia becslésére mély neuronhálós, ultrahang-alapú némabeszéd-interfészekben

Grósz Tamás[1,2], Tóth László[1], Gosztolya Gábor[2],
Csapó Tamás Gábor[3,5], Markó Alexandra[4,5]

[1]Szegedi Tudományegyetem, Informatikai Intézet
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[3]Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
[4]Eötvös Loránd Tudományegyetem, Fonetikai Tanszék
[5]MTA-ELTE Lendület Lingvális Artikuláció Kutatócsoport
{ tothl, groszt, ggabor } @ inf.u-szeged.hu
csapot @ tmit.bme.hu, marko.alexandra @ btk.elte.hu

Kivonat A legkorszerűbb némabeszéd-interfészek vokódert használva közvetlenül az artikulációs jellemzőkből generálnak beszédet. A legtöbb ilyen jellegű kutatásban csak a beszéd spektrális paramétereinek pontos becslésére koncentrálnak, a szintézishez szükséges alapfrekvencia-paraméterként konstans F0-értéket, az eredeti F0-görbét, vagy esetleg fehérzajos gerjesztést használnak. E megoldás mögött az a feltételezés húzódik meg, miszerint az F0-görbe nem becsülhető olyan artikulációs jellemzőkészletből, amely nem tartalmaz a hangszalagok rezgésére vonatkozó közvetlen mérési adatot. Jelen cikkünkben ilyen jellemzőkből, konkrétan ultrahang-felvételekből kíséreljük meg a beszédjel helyreállítását, kiemelt hangsúlyt helyezve a zöngésség, illetve az alapfrekvencia becslésére. Ellentétben a feltevéssel, miszerint az alapfrekvencia ilyen adatokból nem rekonstruálható, a becsült és az eredeti F0-görbe között 0,74-es korrelációs értéket mértünk. Ami még fontosabb, meghallgatásos kísérleteink alanyai nem tudtak különbséget tenni az eredeti és a becsült F0-görbével szintetizált mondatok között, azokat egyforma minőségűnek értékelték.
Kulcsszavak: némabeszéd-interfész, mély neuronháló, alapfrekvencia

1. Bevezetés

Az utóbbi évtizedben megnőtt az érdeklődés a beszédjel artikulációs jellemzőkből való helyreállítása iránt, ami az ún némabeszéd-interfészek (Silent Speech Interface, SSI) alapját képezi [1]. A probléma lényege, hogy valamilyen eszközzel rögzítjük az artikulációs szervek (pl. nyelv és/vagy ajkak) mozgását, majd ezekből az adatokból rekonstruáljuk a beszédjelet anélkül, hogy az alany valóban beszédjelet produkálna. A némabeszéd-interfészek hasznosak lehetnek a beszédképzésben sérültek (pl. gégeeltávolításon átesett betegek) számára, vagy olyan alkalmazásokban, ahol a beszédjel átvitele nem lehetséges, például extrém módon zajos

környezetben (lásd katonai alkalmazások). Az artikulációs adatok rögzítése történhet ultrahangos képalkotással (ultrasound tongue imaging, UTI) [2,3,4,5,6,7], elektromágneses artikulográffal (electromagnetic articulography, EMA) [8,9], állandó mágneses artikulográffal (permanent magnetic articulography, PMA) [10], elektromiográfiával (electromyography, EMG) [11], avagy a fentieket keverő multimodális megoldásokkal [12].

A jelenlegi legkorszerűbb SSI rendszerek a „közvetlen szintézis" alapelvét alkalmazzák, vagyis a beszédjelet közbeeső átalakítások nélkül (pl. beszédhangok felismerése) közvetlenül az artikulációs jellemzőkből állítják elő, vokóder használatával [3,4,5,9,10]. Az ilyen jellegű kísérleteket végző kutatók többsége a szintézishez szükséges spektrális jellemzők becslésére fókuszál (ilyen jellemzőkészlet pl. a Mel-Generalized Cepstrum, MGC). Ennek oka, hogy míg a spektrális burkológörbe nyilvánvaló módon a nyelv és az ajkak mozgásával korrelál, az alapfrekvencia (F0) paramétert a hangszalagok rezgése befolyásolja, ami viszont közvetlen módon nem függ a nyelv, az arc vagy az ajkak konfigurációjától [13]. Ennek ellenére vannak arra utaló kutatások, hogy a nyelv konfigurációja némiképp eltér zöngés és zöngétlen hangok esetén, például a hangszalag rezgése lelassul a mássalhangzók artikulációja során [14]. Egyéb tényezők mellett ezek a változások is korrelálnak az obstruens hangok artikulációs konfigurációjával, azaz a hangrés és az akadály közötti térfogattal [15]. Mindezen kutatási eredmények ellenére az SSI rendszerek fejlesztői az F0-görbe becslését legtöbbször reménytelen feladatként kezelik, ezért az egyszerűség kedvéért konstans F0-értéket, az eredeti F0-görbét, vagy esetleg fehérzaj-gerjesztést használnak a szintézis során.

Van azonban néhány szerző, aki próbálkozott a zöngésség, illetve az F0-görbe helyreállításával. Nakamura és tsai EMG-felvételekkel dolgoztak, és a feladatot két lépésre bontva igyekeztek megoldani. Egyrészt egy SVM-et használtak a zöngés/zöngétlen (voiced/unvoiced, V/U) jelszakaszok elkülönítésére, majd második lépésben egy GMM modellel becsülték a zöngés szakaszok konkrét F0-értékét. Ezzel a módszerrel 0,5 körüli korrelációs értéket értek el az eredeti és a becsült F0-görbe között, maga a V/U döntés pontossága pedig 84% volt [11]. Hueber és tsai ultrahangos adatok és ajakvideók kombinálásával tettek kísérletet arra, hogy a spektrális paraméterek mellett a V/U döntést is megbecsüljék [3]. A célra egy előrecsatolt (feed-forward) mély neuronhálót (deep neural network, DNN) használtak, és a V/U becslés pontosságára 82%-ot kaptak, ami nagyon hasonló Nakamura és tsai eredményéhez. Mivel a mért adatok nem tartalmaztak közvetlenül a hangszalagok működését reprezentáló jellemzőt, a relatíve magas pontosságot indirekt összefüggésekkel magyarázták, például azzal, hogy a stabil artikulációs konfigurációk általában magánhangzóknak, azaz zöngés jelszakaszoknak felelnek meg.

Két egészen friss publikáció EMA-felvételekből próbálta az F0-görbét becsülni. Liu és tsai különféle típusú mély hálók (DNN, RNN és LSTM) teljesítményét hasonlították össze a zöngésség becslésére. Azt kapták, hogy a kétlépéses becslés, azaz először a spektrális jellemzők becslése, majd a becsléssel kapott értékek eredeti jellemzőkkel való összefűzése minden esetben javít az eredményeken [16]. Zhao és tsai pedig úgy találták, hogy az EMA-adatok változási sebessége és

gyorsulása nagyon hasznosak az F0-görbe becslése során, és hogy az ún. LSTM neuronhálók jobban teljesítenek e téren, mint a standard DNN-ek. Objektív kiértékelési kritériumok szerint biztató F0-becslési eredményeket kaptak, meghallgatásos kiértékelést viszont nem végeztek [17].

Habár a fenti rövid összegzés azt mutatja, hogy néhány kísérlet ugyan történt az F0-görbe becslésére, konkrétan olyan kutatást nem találtunk, amely a legmodernebb mély hálós technológiát használná ultrahangos kiindulási adatokon. Csapó és tsai végeztek ugyan ilyen kísérleteket, de az F0 becslésével nem próbálkoztak, másokhoz hasonlóan az eredeti F0-görbét használták a szintézis során [6,7]. Az azonban nyilvánvaló, hogy az SSI-rendszerek későbbi gyakorlati használhatóságához az F0 becslését is meg kell oldani. Jelen cikkben ezzel kísérletezünk, Nakamura és tsai megközelítéséhez hasonlóan egy kétlépcsős modellel, amelyben az egyik gépi tanulási komponens a zöngésségi jegyet, míg a másik a zöngés jelszakaszok konkrét F0-értékét igyekszik megbecsülni. Abban viszont eltérünk a Nakamura-féle tanulmánytól, hogy mindkét részfeladatra mély neuronhálókat alkalmazunk [11]. További eltérés, hogy míg ők EMG-felvételekből indultak ki, esetünkben ultrahang-felvétel az input, amely direkt módon nem rögzíti a hangszalagok rezgését. Módszereink kiértékelését egy női beszélőtől származó felvételeken fogjuk végezni.

2. Kísérleti beállítások

A kísérletekhez használt felvételeket egy (42 éves) magyar anyanyelvű, beszédképzési problémával nem rendelkező nő segítségével rögzítettük, aki összesen 473 mondatot olvasott fel. Eközben a nyelv mozgását az Articulate Instruments Ltd. által gyártott „Micro” típusú ultrahang-berendezéssel rögzítettük 82 kép/másodperc sebességgel. Ezzel párhuzamosan a beszédjelet is felvettük egy Audio-Technica - ATR 3350 típusú kondenzátormikrofonnal. A továbbiakban ismertetett kísérletek inputját a nyers ultrahang-felvételek képezték, a képek mérete 64×119 pixel volt (további részletekért lásd [6,7]).

2.1. Előfeldolgozás és szintetizálás

A beszédjel elemzésére és szintetizálására a nyílt forrású SPTK eszköztár egyik vokóderét használtuk (http://sp-tk.sourceforge.net). A beszédjelet újramintavételeztük 11 050 Hz-en. Az F0-görbét a SWIPE algoritmussal nyertük ki [18]. A spektrális burkológörbét 12 MGC-LSP együtthatóval, valamint az energiaértékkel reprezentáltuk, ami összességében egy 13-dimenziós jellemzővektort eredményezett. A paramétereket az ultrahangképekkel szinkronban, 12 ms kereteltolással nyertük ki. A mély neuronhálók tanítása során az előbbi vektor képezte a megtanulandó célvektort. A szintézis során az eredeti paramétervektor helyett a DNN által az ultrahang-felvételekből becsült értékeket használtuk. Ehhez a vokóder az F0 paraméter segítségével impulzus-gerjesztőjelet produkált, majd ezt átengedte az MGC-LSP paraméterekből képzett ún. Mel-Generalized Log Spectral Approximation (MGLSA) szűrőn [19], így végezve el a beszédjel rekonstrukcióját.

1. ábra. Az F0 becslésére javasolt eljárás.

2.2. Az alapfrekvencia becslése mély neuronhálóval

Az alapfrekvencia becslése pusztán ultrahang-felvételek alapján nem egyszerű feladat. Nakamura és tsai eljárásához hasonlóan mi is kétlépcsős gépi tanulási megközelítést alkalmaztunk [11], ahol az egyik gépi tanuló a zöngésség jelenlétét, a másik pedig a zöngés keretek F0-értékét volt hivatva megbecsülni. Nakamura és társaival szemben mi mindkét feladatra mély neuronhálót alkalmaztunk. Mivel a V/U döntés kimenete bináris, ezt osztályozási feladatként kezeltük, a konkrét F0-érték becslését viszont regressziós feladatnak tekintettük, és ez utóbbi neuronhálót csak a zöngés időkeretekhez megadott F0-értékek becslésére tanítottuk. A kiértékelési lépés, azaz a szintézis során a két háló kimenetét kombináltuk, azaz a zöngésnek ítélt adatkeretekhez az F0-becslő háló kimenetét rendeltük, míg a zöngétlen pillanatokban egy megfelelő (a vokóder által definiált) konstans értéket adtunk vissza. (A becslés menetéről lásd az 1. ábrát.)

2. ábra. Egy példa a rendszerünk alapfrekvencia-becslési kimenetére, az eredeti F0-görbéhez képest.

Megjegyezzük, hogy az utóbbi években a képi felismerési feladatokban a konvolúciós neuronhálók használata vált dominánssá. Mi azonban egyszerű teljesen kapcsolt hálót használtunk az alábbi megfontolások miatt. Egyrészt a konvolúciós hálók olyan feladatokban bizonyulnak sokkal jobbnak a standard hálóknál, amikor a kép sok apró részletből, hierarchikusan épül föl. Esetünkben azonban a kép alacsony felbontású, és csupán pár építőelemből áll (a nyelv, az állkapocs, valamint a nyelvcsont árnyéka). Másrészt a konvolúciós hálók a hierarchikus felbontás során elveszítik az egyes komponensek pontos pozícióját. Általános esetben ez még előnyös is, de úgy éreztük, hogy esetünkben lényeges információt veszítenénk. A fenti indokok miatt egy egyszerű, 5 rejtett réteget tartalmazó előrecsatolt hálóval dolgoztunk, rétegenként 1000 ReLU neuronnal. A háló feladata a 12 LSP paraméter, az energia és az F0-érték becslése volt; előzetes kísérleteinkben úgy találtuk, hogy hatékonyabb ezeket együtt tanulni, mint külön-külön. Ezért a paraméterek együttes tanulására (regressziójára) egy 14 lineáris kimeneti neuront tartalmazó DNN-t használtunk. A zöngésség eldöntését tanuló neuronhálónk ugyanilyen felépítésű volt, de egy kétosztályos softmax kimeneti réteggel.

2.3. Az input reprezentálása

Csapó és tsai korábbi cikkükben számos kísérletet végeztek az optimális input-reprezentáció megtalálására. Ezért mi itt csak azokat a megoldásokat vizsgáltuk, amelyek ott a legjobbnak bizonyultak [6,7]. Mindkét neuronháló ugyanazon az input adaton tanult. Összehasonlítási alapként (a továbbiakban *alaprendszer*), a legegyszerűbb megoldás szolgált, amikor a DNN inputja egyetlen ultrahang-képkocka. Ezután az inputot kibővítettük, hogy 5 szomszédos képkockából álló adatvektort tartalmazzon (a továbbiakban *5-keret*). Ezzel az egyszerű megoldás-

	Development			Teszt		
	Zöng.	F_0		Zöng.	F_0	
Input	becslés	NMSE	Korr.	becslés	NMSE	Korr.
Alaprendszer (1 keret)	86,1%	0,562	0,719	85,7%	0,577	0,711
5 keret	88,2%	0,476	0,760	87,2%	0,516	0,742
17 keret + jellemzőkiv.	87,4%	0,506	0,747	86,9%	0,526	0,736

1. táblázat. A DNN zöngésségi és F0-becslési pontossága különféle input-jellemzőkészletek esetén

sal azonban jelentősen megnő a jellemzővektor mérete, ami gépi tanulási szempontból nem előnyös. Emiatt egy jellemzőkiválasztási módszert alkalmaztunk az egyes képek méretének 20%-ra való redukálására [6,7]. Ily módon 8-8 bal és jobb oldali szomszédos képkockát is figyelembe tudtunk venni az inputvektor méretének jelentős növekedése nélkül. Ezt a rendszert *17-keret+jk* néven fogjuk hivatkozni. További részletek a jellemzőkinyerésről Csapó és tsai cikkében találhatók [7].

A legjobbnak bizonyuló F0-becslő modell kiértékelésére szubjektív meghallgatásos teszteket is végeztünk, amelyekben ugyanazt a mondatot különböző F0-görbékkel is szintetizáltuk. Összehasonlítási alapként konstans F0-t alkalmaztunk, ahol a DNN csupán a V/U döntést becsülte. Ez a modell *F0-konstans* címkével szerepel az ábrákon. Az összehasonlítás másik végpontjaként a mondatot az eredeti F0-görbével szintetizáltuk (*F0-eredeti* az ábrákon). Végezetül, a becsült F0-görbét használó modell *F0-DNN* néven szerepel a továbbiakban.

3. Eredmények és diszkusszió

3.1. Objektív kiértékelés

A DNN-alapú F0-becslés pontosságának számszerű kiértékelésére a normalizált átlagos négyzetes hibát (Normalized Mean Square Error, NMSE) és a Pearson korrelációs együtthatót használtuk [7]. Az 1. táblázat összegzi a különféle, az előző fejezetben bemutatott inputreprezentálási módszerek esetén kapott eredményeket. Mindegyik megoldással elég jó V/U döntési pontosságot értünk el, a legjobb F0-becslő modell pedig 0,74 korrelációs értéket produkált. A 2. ábra egy példán érzékelteti az eredeti és a becsült F0-görbe eltérését. Az eredmények alapján az alapfrekvencia ultrahang-felvételekből való becslése egyáltalán nem reménytelen.

3.2. Szubjektív meghallgatásos tesztek

Az objektív hibamértékek sajnos nem sokat mondanak arról, hogy vajon a becslési hiba mennyire befolyásolja a szintetizált mondatok minőségét. Ezért a hangminőség szubjektív kiértékelésére online MUSHRA (MUlti-Stimulus test with

3. ábra. A különböző reprezentációs módszerek összehasonlítására végzett meghallgatásos tesztek eredménye.

Hidden Reference and Anchor) meghallgatásos teszteket végeztünk [20]. A tesztek célja az eredeti, a szintetizált és a DNN-nel becsült F0-görbével szintetizált hangminták minőségének összevetése volt. Az újraszintetizált mondatokat a vokóder segítségével, a vokódert az eredeti mondatokból kinyert F0 és MGC paraméterekkel futtatva kaptuk. A minőségi skála alsó pontjának belövését segítendő a tesztek mindig tartalmaztak egy rossz minőségű ún. "horgony" mondatot, ezt konstans F0-érték és torzított MGC-paraméterek használatával állítottuk elő. A tesztekben 10 olyan mondat szerepelt, amelyeket a gépi tanuló algoritmus nem látott. Összesen 24 személy vett részt a tesztelésben (13 nő és 11 férfi). Mindannyian magyar anyanyelvűek voltak, négyük rendelkezett beszédtechnológiai ismeretekkel. Életkoruk 18–74 év közé esett, az átlag 28 év volt. A teszt elvégzése átlagosan 18 percet igényelt.

Kétféle tesztet végeztünk. Az első célja a különféle input-reprezentálási módozatok összevetése volt, míg a második teszt az F0-becslés pontosságának a hangminőségre való hatását volt hivatva felmérni. Az 1. teszt az *Alaprendszert*, a *17-keret+jk*, és a *5-keret* elnevezésű reprezentálási stratégiákat hasonlította össze, míg a 2. teszt az *F0-konstans*, *F0-DNN*, és *F0-eredeti* nevű F0-becslési módozatokat vetette össze. Mind a 10 mondat, illetve az összes mondatvariáns (6 változat mondatonként) véletlenszerű (és minden tesztalany számára eltérő) sorrendben szerepelt a tesztekben. A tesztalanyok feladata az egyes mondatváltozatok minőségének pontozása volt egy 0-tól (nem természetes) 100-ig (természetes) terjedő skálán, ahol a 100 pontos minőség referenciájaként az eredeti mondatot adtuk meg.

A meghallgatásos tesztek eredménye A MUSHRA pontértékeket a 3. ábra mutatja az első, a 4. ábra a második teszt esetén (a 95%-os konfidenciaintervallumokat is feltüntetve). Mint látható, az eredeti mondatok közel 100%-ot értek

4. ábra. A különböző F0-predikciós módszerek összehasonlítására végzett meghallgatásos tesztek eredménye.

el a természetességi skálán. Az újraszintetizált mondatok ehhez képest csupán 60%-ot kaptak, míg a skála alsó pontját mutató „horgony” mondatok valóban messze a legalacsonyabb pontszámot kapták. A három DNN-alapú, eltérő jellemzőreprezentálást alkalmazó mondatokat a hallgatók 20–36% közé tették, ami azt mutatja, hogy a természetességük az újraszintetizált és a horgony mondatok között körülbelül félútra esik.

A meghallgatásos tesztben előállt rangsort Mann–Whitney–Wilcoxon-féle rangösszegpróbával is összevetettük, 95%-os konfidenciaszintet használva. Ennek alapján a DNN segítségével szintetizált mondatok szignifikánsan különböző minőségűnek bizonyultak az eredeti, a horgony és az újraszintetizált mondatok mindegyikéhez képest. Az első tesztben a *17-keret+jk* és az *5-keret* reprezentációs stratégiák szignifikánsan jobbak voltak, mint az *alaprendszer*, de kettőjük között szignifikáns különbség nem mutatkozott. A második tesztben az *F0-konstans* becslést használó modell szignifikánsan rosszabb eredményt ért el, mint az *F0-DNN* és az *F0-eredeti* rendszer, és ami a legfontosabb, az utóbbi kettő között az eltérés nem volt szignifikáns. Ez azt jelenti, hogy a tesztalanyok nem tudtak különbséget tenni a DNN-predikált és az eredeti F0-görbével szintetizált mondatok között, azokat egyforma minőségűnek ítélték.

4. Konklúzió

Cikkünkben kísérleteket végeztünk az F0-görbe becslésére ultrahang-felvételekből történő beszédjel-helyreállítás során. E célra két mély neuronhálót tanítottunk párhuzamosan, egyet a zöngésség, egyet pedig a konkrét F0-érték és a spektrális paraméterek becslésére. Objektív kiértékelési mértékeket használva azt találtuk, hogy a becsült és az eredeti alapfrekvencia-görbe korrelációs együtthatója 0,74. Szubjektív meghallgatási tesztjeink során a tesztalanyok az eredeti, illetve a

predikált F0-görbével szintetizált mondatokat közel egyformán természetesnek minősítették. Eredményeink alapján elmondhatjuk, hogy a közkeletű feltevéssel szemben az alapfrekvencia becslése nem reménytelen még akkor sem, ha a mérési adatok (esetünkben ultrahang-felvételek) a hangszalagok rezgését közvetlen módon nem reprezentálják. Mindamellett megjegyezzük, hogy kísérleteinkben egyetlen beszélő anyagával dolgoztunk, és ez a tény feltehetőleg nagymértékben hozzájárult a jó eredményekhez. Azt gondoljuk, hogy a beszélőfüggőség nem jelent hátrányt olyan értelemben, hogy a jövőbeni némabeszéd-interfészek minden bizonnyal személyfüggőek lesznek. Ennek ellenére tervezzük kísérleteinket megismételni több beszélőtől (férfiak és nők vegyesen) származó felvételekkel. További megjegyzésünk, hogy éles alkalmazásokban a rendszernek majd a hangszalagok rezgése nélkül is működnie kell, mi viszont egy egészséges tesztalannyal dolgoztunk, aki normál beszédet produkált. Ebből következően szükséges lesz a kísérletek valódi néma beszéddel való megismétlése is.

Köszönetnyilvánítás

A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta (FK 124584). Tóth László munkáját az MTA Bolyai János Kutatási Ösztöndíja támogatta. Grósz Tamást az Emberi Erőforrások Minisztériuma ÚNKP-17-3 kódszámú Új Nemzeti Kiválóság Programja támogatta. A cikk elkészítéséhez használt Titan-X grafikus kártyát az NVIDIA Corporation adományozta.

Hivatkozások

1. Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S.: Silent speech interfaces. Speech Communication **52**(4) (2010) 270–287
2. Denby, B., Stone, M.: Speech synthesis from real time ultrasound images of the tongue. In: ICASSP, Montreal, Kanada (2004) 685–688
3. Hueber, T., Benaroya, E.l., Denby, B., Chollet, G.: Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface. In: Interspeech, Florence, Olaszország (2011) 593–596
4. Hueber, T., Bailly, G., Denby, B.: Continuous articulatory-to-acoustic mapping using phone-based trajectory HMM for a silent speech interface. In: Interspeech, Portland, USA (2012) 723–726
5. Jaumard-Hakoun, A., Xu, K., Leboullenger, C., Roussel-Ragot, P., Denby, B.: An articulatory-based singing voice synthesis using tongue and lips imaging. In: Interspeech. (2016) 1467–1471
6. Csapó, T.G., Grósz, T., Tóth, L., Markó, A.: Beszédszintézis ultrahangos artikulációs felvételekből mély neuronhálók segítségével. In: MSZNY 2017, Szeged (2017) 181–192
7. Csapó, T.G., Grósz, T., Gosztolya, G., Tóth, L., Markó, A.: DNN-based ultrasound-to-speech conversion for a silent speech interface. In: Interspeech, Stockholm, Svédország (2017) 3672–3676
8. Wang, J., Samal, A., Green, J.: Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph. In: SPLAT, Baltimore, USA (2014) 38–45

9. Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., Yvert, B.: Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. PLOS Computational Biology **12**(11) (2016) e1005119
10. Gonzalez, J.A., Cheah, L.A., Green, P.D., Gilbert, J.M., Ell, S.R., Moore, R.K., Holdsworth, E.: Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary. In: Interspeech, Stockholm, Svédország (2017) 3986–3990
11. Nakamura, K., Janke, M., Wand, M., Schultz, T.: Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0. In: ICASSP, Prága, Csehország (2011) 573–576
12. Freitas, J., Ferreira, A.J., Figueiredo, M.A.T., Teixeira, A.J.S., Dias, M.S.: Enhancing multimodal silent speech interfaces with feature selection. In: Interspeech, Szingapúr (2014) 1169–1173
13. Jiang, J., Alwan, A., Bernstein, L.E., Keating, P., Auer Jr., E.: On the correlation between facial movements, tongue movements, and speech acoustics. EURASIP Journal on Applied Signal Processing (11) (2002) 174–1188
14. Bickley, C.A., Stevens, K.N.: Effects of a vocal tract constriction on the glottal source: experimental and modeling studies. Journal of Phonetics **14** (1986) 373–382
15. Westbury, J.R., Keating, P.A.: On the naturalness of stop consonant voicing. Journal of Linguistics **22** (1986) 145–166
16. Liu, Z.C., Ling, Z.H., Dai, L.R.: Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks. In: Interspeech. (2016) 1502–1506
17. Zhao, C., Wang, L., Dang, J., Yu, R.: Prediction of F0 based on articulatory features using DNN. In: ISSP, Tienjin, Kína (2017)
18. Camacho, A., Harris, J.G.: A sawtooth waveform inspired pitch estimator for speech and music. The Journal of the Acoustical Society of America **124**(3) (2008) 1638–1652
19. Imai, S., Sumita, K., Furuichi, C.: Mel log spectrum approximation (MLSA) filter for speech synthesis. Electronics and Communications in Japan (Part I: Communications) **66**(2) (1983) 10–18
20. : ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality (2001)

A diszfónia súlyosságának automatikus becslése a szakértői értékelések szubjektív jellegének figyelembevételével

Tulics Miklós Gábriel[1], Jászai Henrietta[1], Vicsi Klára[1]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
{tulics,vicsi}@tmit.bme.hu
{jaszaiheni}@gmail.com

Kivonat Ebben a munkában egy a diszfónia súlyosságát automatikusan becslő eljárás kerül bemutatásra. A k-közép algoritmus segítségével kimutattuk, hogy a különböző fonetikai osztályokba tartozó beszédhangokon mért akusztikai paraméterek alkalmasak a szakemberek négyfokozatú értékelésének (RBH szubjektív skála) modellezésére. A kapott felügyelet nélküli automatikus becslési eredmény meglepően közel áll a szubjektív megítéléshez. Ezen kívül megvizsgáltuk négy szakember értékelésének következetességét. A négy szakember egyike állította fel a diagnózist és értékelte a beteg beszédének minőségét a konzultációk során; a másik három szakember nem ismerte a pácienst, és csak a korábban rögzített hangfájlok alapján határozta meg a diszfónia súlyosságát. A diszfónia súlyosságának automatikus becslése során két regressziós modell került összehasonlításra: a négy szakember értékelésének átlaga alapján készített és a betegeket kezelő klinikus ítélete alapján létrehozott modell. A két modell alacsony RMSE és magas korrelációs értékeket ért el az automatikusan becsült súlyosságot és a perceptuális értékeléseket összevetve.
Kulcsszavak: beszédelemzés, folyamatos beszéd, kóros beszéd, következetesség vizsgálat, regresszióelemzés, klaszterelemzés, diagnosztika

1. Bevezetés

A fejlett országokban a dolgozók egynegyedénél elengedhetetlen a beszéd használata kommunikációs eszközként. A beszédhang-képzés bármilyen rendellenessége kihatással van a magánéletre, érintheti a szakmai előmenetelünket és a megélhetésünket is. Új, olcsó és hatékony módszerekre van szükség, amelyek segítik a háziorvosok, fül orr gégészek munkáját a betegségek feltárásában.

A kutatók célja egy olyan automatikus becslési rendszer létrehozása, amely felismeri a kóros hangot, és automatikusan meghatározza a diszfónia súlyosságát, ezáltal lehetőség nyílik a hang non-invazív és objektív diagnosztizálására, ezzel együtt a korai felismerésre. A diszfónia a hangképzés diszfunkciójára utal. A diszfóniás hang rendszerint rekedt, levegős, fátyolos, préselt vokális tulajdonságokkal jellemezhető [1]. A folyamatos beszéd vizsgálatának számos előnye van

a kitartott magánhangzók elemzésével szemben, mivel a folyamatos beszéd tartalmaz számos alapfrekvencia variációt, szüneteket és lehetőség van a beszédhangok különböző változatainak elemzésére. A kutatási eredmények várhatóan jobban alkalmazhatók a gyakorlatban, mivel a valós életben a folyamatos beszédet használjuk [2,3]. Az egészséges és kóros hangok automatikus osztályozásában az eredmények jelentősen javultak, amikor folyamatos beszédet kezdtek el használni a kitartott magánhangzók helyett [4].

A kutatók olyan akusztikai paraméterek kifejlesztésére összpontosítottak, amelyek hatékonyan reprezentálják a hangképző rendszer kóros állapotát. A legszélesebb körben használt akusztikai paraméterek közé tartoznak a következők: jitter, shimmer és Harmonics to Noise Ratio (HNR). Korábbi kutatások igazolják, hogy olyan akusztikai paraméterek, mint a jitter, shimmer, HNR és az MFC-együtthatók (Mel-frequency cepstral coefficients) első komponense (c_1) (a továbbiakban: mfcc01) az egészséges és kóros hangok automatikus osztályozásában hasznos paraméterek [4,5,6].

Korábban kimutattuk [7], hogy ezek az akusztikai paraméterek magas korrelációt mutatnak a diszfónia súlyosságával, ugyanúgy, mint a különböző fonetikai osztályokon mért Soft Phonation Index (SPI) és Empirical mode decomposition (EMD) alapú frekvenciasáv-arányok. Ezek a paraméterek hasznosak lehetnek a diszfónia különböző típusainak, például funkcionális diszfónia és recurrens paresis megkülönböztetéséhez.

A diszfónikus beszéd diagnózisát és kezelését a páciens hangminőségét vizsgáló orvos végzi. A beszéd értékelése természeténél fogva szubjektív. A diszfóniás hang súlyosságát általában egy orvos értékelése, vagy több tapasztalt értékelő ítéletének átlaga vagy mediánja alapján határozzák meg [8,9]. Ha több értékelő áll rendelkezésre, a szakemberek a diszfónia súlyosságának értékelését korábban rögzített hangminták meghallgatásával végzik. Az értékelések változhatnak az értékelő szakemberek között, ezért tanácsos konzisztenciaelemzést végezni.

Ebben a tanulmányban organikus és funkcionális diszfónia súlyosságának automatikus becslésére összpontosítunk. Egy diszfónia súlyosságát becslő módszert javaslunk, amely magyar anyanyelvű betegek és kontrollpopuláció felolvasott szövegein alapul. A diszfónia súlyosságát négy szakember határozta meg: a szakember, aki a pácienst kezelte és közvetlenül értékelte a páciens beszédének minőségét a konzultáció során, valamint további három szakember, aki nem ismerte a pácienseket, csak az előzetesen rögzített hangfájlokat hallgatta meg és azok alapján határozta meg a diszfónia súlyosságát. A diszfónia súlyosságát az RBH-skála szerint határozták meg a szakemberek, ahol az R az érdességet, a B a levegősséget, a H pedig a rekedtséget jelenti [10]. Korábbi eredményeink alapján [7] 18 különböző fonetikai osztályon (például nazálisokon, magas magánhangzókon, zöngés spiránsokon stb.) mért akusztikai paramétert választottunk ki és végeztünk osztályozási, valamint regressziós elemzéseket. Négyosztályú osztályozáshoz k-közép felügyelet nélküli osztályozót használtunk, annak érdekében, hogy megtudjuk a használt akusztikai paraméterek alkalmasak-e a szakemberek négyfokozatú értékelésének modellezésére. Megvizsgáltuk az RBH szubjektív jellegét, valamint a négy szakember minősítésének konzisztenciáját. A pácienseket kezelő szakem-

ber értékelését összehasonlítottuk azon szakemberek értékeléseivel, akik az előre rögzített hangfelvételek alapján határozták meg a diszfónia súlyosságát.

A Bevezetés után, a 2. fejezetben röviden ismertetjük a kísérletekben használt beszédadatbázist, a mért akusztikai paramétereket, valamint a kiértékelési módszereket. Eredményeinket a 3. fejezetben mutatjuk be, majd a 4. fejezetben ismertetjük az eredmények interpretálását és a jövőbeli terveinket.

2. Módszerek és beszédanyagok

2.1. Patológiás és egészséges beszédadatbázis

A páciensek hangmintáinak felvétele, az Országos Onkológiai Intézet Fül-orr-gégészeti osztályának járóbeteg ellátásai során történt. A vizsgálat során számos betegség fordult elő: funkcionális diszfónia, recurrens paresis (hangszalagbénulás), a hangképző szervrendszer különböző pontjain előforduló tumorok, gasztroesophageal reflux (GERD), krónikus gégegyulladás, bulbar paresis (agyideggyulladás), amiotrófiás laterálszklerózis (ALS), leukoplakia, spazmodikus diszfónia, stb. Összehasonlítás végett egészséges páciensekről is készültek felvételek. Ezek rögzítése a korábban említett esetekhez nem kapcsolódó kivizsgálások során kerültek felvételre.

2.2. Az RBH skála

A hang kórosságának súlyosságát négy szakember határozta meg: a szakorvos, aki kezelte a pácienst, felállította a diagnózist és meghatározta a hangjának súlyosságát a konzultációk során (Szakember 1), valamint további három szakember (Szakember 2, Szakember 3 és Szakember 4) aki nem került személyes kapcsolatba a páciensekkel, csupán a hangfelvételek visszahallgatása alapján határozták meg a diszfónia súlyosságát. Mind a négy szakembernek nagy tapasztalata van a hangképzési rendellenességekkel rendelkező páciensek kezelésében. A diszfónia súlyosságát az RBH skálával lehet meghatározni, ahol az R (roughness) az érdességet, a B (breathiness) a levegősséget, míg a H (hoarseness) az általános rekedtséget hivatott mutatni. A H értéke nem lehet kisebb, mint a másik két kategória maximuma. Például ha a B=3 és az R=2, akkor a H értéke 3, nem lehet 2, vagy 1. Az egészséges hangok kódja így R0B0H0 szerint alakul. Ptok és társai igazolták, hogy az RBH skála alkalmazása megfelelő a klinikai alkalmazásokra [11]. A tanulmányunk során az általános rekedtség (H) vizsgálata, és diagnosztizálása történt. 138 felvétel került felhasználásra. Minden felvétel esetén a szakemberek külön külön meghatározták a diszfónia súlyosságának fokát az RBH skála szerint, majd a H értékeket felvételenként átlagoltuk, ezt használtuk fel később célváltozóként. A szakemberek által meghatározott H értékek átlagának eloszlása az adatbázisban az 1. ábrán látható.

1. ábra: A H érték eloszlása az adatbázisban.

2.3. A felvételek rögzítési környezete és szövege

A felvételek (Monacor ECM-100) közeltéri mikrofonnal, alacsony zajszinttű külső hangkártyával (Creative Soundblaster Audigy 2 NX), jó minőségű A/D konverterrel (kódolás: PCM, mintavételezési frekvencia: 16 kHz, kvantálás: 16-bit) kerültek rögzítésre, csendes irodai környezetben (orvosi szobában). Minden páciens Aiszóposz meséjét, „Az északi szél és a nap” -ot olvasta fel. Ezen népmese gyakran használt a foniátriai kutatásokban, mivel a mese fonetikailag kiegyensúlyozott. Vagyis a szöveganyagát úgy szerkesztették meg, hogy az adott nyelvben előforduló minden beszédhang, valamint a leggyakoribb hangkapcsolatok szerepelnek benne. Számos nyelvre elkészült ez a szöveg, köztük a jelen esetben is használt magyarra. Az adatbázis fonéma szintű szegmentálása automatikus fonémaszegmentáló programmal történt, amit a Beszédakusztikai Laboratórium munkatársai fejlesztettek ki [12], majd szükség esetén kézileg történt ezek javítása.

2.4. Akusztikai paraméterek

A korábbi munkák során az alap paraméter szett a következő volt: jitter(ddp), shimmer(ddp), Harmonics-to-Noise Ratio (HNR) és mfcc01, melyeknek átlaga (mean), illetve szórása (std) a népmesében legtöbbet szereplő [E] (SAMPA) magánhangzóból kerültek kinyerésre [5] (továbbiakban Baseline paraméter szett). Ezt követően a paraméterek listája további két akusztikai mérőszámmal bővült, a Soft Phonation Index (SPI)-el és az Empirical mode decomposition (EMD) módszeren alapuló frekvencia sávok arányaival, melyeket különböző fonetikai osztályokba tartozó beszédhangokon vizsgáltuk meg.

Az SPI a beszéd jel alacsony frekvencia sávjának (70 - 1600 Hz) és a magas frekvencia (1600 - 4500 Hz) sávjának energiaaránya. Ha ez az arány magas, akkor az azt jelenti, hogy az energia az alacsony frekvencia sávra koncentrálódik, amely lágyabb hangot eredményez [13].

Fogalmilag az EMD (Empirical Mode Decomposition) a multikomponenses jelet szétbontja elemi jel komponensekké, amik aztán úgy ismertek, mint Intrinsic mód függvények (Intrinsic Mode Functions - IMF). Az IMF-ek frekvencia szerint vannak mátrixba rendezve. Az első pár IMF a magas frekvenciájú komponense a jelnek, míg a továbbiak a kisebb frekvenciájú komponenseket reprezentálják. Minden IMF esetén kiszámításra került az entrópiája (H). A frekvencia sávok entrópiáinak aránya pedig a következő módon került meghatározásra:

$$IMF_{entropy} = \frac{\sum_{d=1}^{2} H_d}{\sum_{d=2}^{D} H_d} \tag{1}$$

H_d a log transzformált IMF-ek Shannon entrópia értéke az egyes $d = 1, 2, \ldots D$ értékekre, ahol D az összes kinyert IMF száma. A Shannon entrópia diszkrét jelek esetén a következőképpen van definiálva:

$$H(p_i) = -K \sum_{i=1}^{n} p_i log p_i \tag{2}$$

Ahol K pozitív konstans. Az IMF paraméter meghatározása a [14] Matlab toolkittel történt.

Az SPI, valamint az IMF entrópia értékét nem csak az [E] magánhangzóra, hanem további fonetikai csoportokra is meghatároztuk. A vizsgálat során az SPI és az IMF entrópiáját a következő hat fonetikai osztályra számoltuk ki:

- az [E] magánhangzón
- nazális hangokon (Nasal), mint [m], [n] és [J]
- magas magánhangzókon (HighVowels), mint [E], [e:], [i], [2] és [y]
- mély magánhangzókon (LowVowels), mint [O], [A:], [o] és [u]
- spiránsokon (VoicedSpirants), mint [v], [z] és [Z]
- zárhangokon és affrikátákon (VoicedPlosives), mint [b], [d], [g], [dz], [dZ] és [d']

Az SPI paramétert az egész hangfájlra is kinyertük, a többi esetben felvételenként átlag és szórás is kiszámításra került. A SPI és az IMF akusztikai paramétereken kívül, a jitter, shimmer, HNR és az mfcc01 felvételenkénti átlag- és szórásértékeivel összesen 33 paraméter került vizsgálatra.

2.5. Döntési módszerek

Felügyelet nélküli klaszterelemzéssel megvizsgáltuk, hogy a különböző fonetikai osztályokon mért akusztikai paraméterek alkalmasak-e a szakemberek négyfokozatú (RBH szubjektív skála szerinti) értékelésének modellezésére. Ehhez k-közép klaszterező algoritmust használtunk. A k-közép az egyik legegyszerűbb algoritmus, amely felügyelet nélküli tanulási módszert alkalmaz az ismert klaszterezési

problémák megoldására. Ez az eljárás egyszerű és gyors megközelítése a problémának, könnyen implementálható és könnyen értelmezhető az eredménye.

Megvizsgáltuk a négy szakember értékeléseinek konzisztenciáját. A diszfónia súlyosságának becslésére lineáris regressziót és radiális bázisfüggvényű (RBF) SVR-t (Support Vector Regression, továbbiakban SVR) használtunk. A lineáris regresszió természetéből adódóan csak az egymástól függő és a független változók lineáris kapcsolatát vizsgálja; továbbá, azt is feltételezi, hogy a bemeneti változók és a célváltozó között lineáris kapcsolat van. Az RBF-kernellel rendelkező SVR-nek jó általánosító képessége van és robosztus a bemenő zajra tekintettel.

Az egyenletesebb eloszlás érdekében az adatbázist kibővítettük 10 extra egészséges felvétellel (H0) a klaszter- és a regresszióelemzés feladatokhoz. E populáció beszédét ugyanazzal a szöveges anyag felhasználásával rögzítettük. Az páciensek nem rendelkeztek más betegséggel, és nem voltak semmilyen orvosi kezelés alatt. A következetesség vizsgálat során az eredeti 138 mintát használtuk.

3. Eredmények

3.1. Felügyelet nélküli klaszterelemzés

Négy osztályos k-közép klaszter analízist végeztünk egy 18 paramétert tartalmazó bemenő jellemzővektorral, amely jó eredményekkel tudta osztályozni az egészséges és diszfóniás hangokat. A 18 paraméteres bemenő jellemzővektor egy Forward Feature Selection (FFS) algoritmus eredményeként került kiválasztásra. Az FFS egy iteratív algoritmus, amely kiválasztja a legjobb paramétert egy előre meghatározott költségfüggvény kielégítése alapján, minden lépésben egy újabb paramétert hozzáadva a paraméterek halmazába. Jelen esetben a paraméterek a legalacsonyabb RMSE érték szerint lettek kiválasztva. A kiválasztott paraméterek az 1. táblázat tartalmazza.

1. táblázat. 18 akusztikai paraméteres szett.

	Kiválasztott akusztikai paraméterek
18 paraméter szett	jitter$_{mean}$, shimmer$_{mean}$, hnr$_{mean}$, mfcc01$_{mean}$, jitter$_{std}$, shimmer$_{std}$, hnr$_{std}$, mfcc01$_{std}$, SPI->E$_{std}$, SPI->Nasal$_{mean}$, SPI->Nasal$_{std}$, SPI->LowVowels$_{std}$, SPI->VoicedSpirants$_{mean}$, SPI->VoicedSpirants$_{std}$, IMF->E$_{std}$, IMF->Nasal$_{mean}$, IMF->VoicedPlosives$_{mean}$, IMF->VoicedPlosives$_{std}$

Érdekes kérdés, hogy ezen kiválasztott paraméterek modellezni tudják-e az egyéni értékeléseket. Ha az akusztikai paraméterek szettje, és a felügyelet nélküli tanuló algoritmus rögzített, akkor összehasonlítható a négy szakember értékeléseinek klaszter modellje külön-külön. Az RBH skála szubjektív voltának vizsgálata céljából klaszter analízist végeztünk.

Az egyes szakemberek tévesztési mátrixait a 2., 3., 4. és 5. táblázat reprezentálja. Az osztályozások teljesítményét a 6. táblázat mutatja. A H döntések

2. táblázat. Tévesztési mátrix Szakember 1 esetén.

		Prediktált H				
		0	1	2	3	Összesen
Szakember 1 által megállapított H érték	0	**12**	13	9	2	36
	1	1	**33**	25	4	63
	2	2	5	**10**	6	23
	3	1	3	5	**17**	26
	Összesen	16	54	49	29	148

3. táblázat. Tévesztési mátrix Szakember 2 esetén.

		Prediktált H				
		0	1	2	3	Összesen
Szakember 2 által megállapított H érték	0	**11**	5	6	0	22
	1	3	**26**	24	2	55
	2	2	23	**16**	15	56
	3	0	0	3	**12**	15
	Összesen	16	54	49	29	148

4. táblázat. Tévesztési mátrix Szakember 3 esetén.

		Prediktált H				
		0	1	2	3	Összesen
Szakember 3 által megállapított H érték	0	**11**	2	3	0	16
	1	2	**20**	15	1	38
	2	2	25	**16**	9	52
	3	1	7	15	**19**	42
	Összesen	16	54	49	29	148

5. táblázat. Tévesztési mátrix Szakember 4 esetén.

		Prediktált H				
		0	1	2	3	Összesen
Szakember 4 által megállapított H érték	0	**12**	7	6	0	25
	1	2	**24**	18	6	50
	2	2	18	**17**	6	43
	3	0	5	8	**17**	30
	Összesen	16	54	49	29	148

6. táblázat. Teljesítménymetrikák minden szakember és címke esetén.

	Metrika	**H0**	**H1**	**H2**	**H3**
Szakember 1 által megállapított H érték	Precision	0,33	0,52	0,43	0,65
	Recall	0,75	0,61	0,20	0,59
	F-score	0,46	0,56	0,28	0,62
Szakember 2 által megállapított H érték	Precision	0,50	0,44	0,41	0,80
	Recall	0,69	0,44	0,47	0,41
	F-score	0,58	0,44	0,44	0,55
Szakember 3 által megállapított H érték	Precision	0,69	0,53	0,31	0,45
	F-score	0,69	0,43	0,32	0,54
Szakember 4 által megállapított H érték	Precision	0,48	0,48	0,40	0,57
	Recall	0,75	0,44	0,35	0,59
	F-score	0,59	0,46	0,37	0,58

átlagos pontossága az egyes szakemberek esetén sorrendben: 0,49; 0,44; 0,45; 0,47.

Ebből a vizsgálatból azt a következtetést vonhatjuk le, hogy a bemenő jellemzővektor alkalmas a diszfónia súlyosságának egyedi értékelésére. Összességében elmondható az egyes tévesztési mátrixok alapján, hogy amennyiben a klaszterezési eljárás nem tudta pontosan meghatározni a szakértő értékelését, akkor is főként a szomszédos klaszterekbe sikerült besorolnia az adott felvételt. Az egészséges (H0) és a patológiás (H1, H2 és H3) hangok elválasztása kielégítő. A tévesztési mátrixok megmutatják az egyes szakemberek minősítési stílusát, miszerint míg Szakember 1 a hangokat kevésbé ítélte súlyosnak, addig Szakember 4 több H3-as súlyosságú értékelést adott. Annak ellenére, hogy a Szakember 3 értékelései voltak a legkevésbé pontosak, mégis a H3 értékeléseinek pontossága meglepően magas, 0,80 volt. Az eredmények azt is mutatják, hogy mind a négy szakember értékelései esetén a H1 és H2-es esetek állnak legközelebb egymáshoz.

Ez jól mutatja, hogy jelen esetben egy folytonos skálát becslő eljárás, mint a regresszió, jobban tudná közelíteni az értékeléseket, mint a diszjunkt halmazokkal dolgozó klaszterezés, így pontosabb, kisebb hibával dolgozó rendszert eredményezve.

3.2. Következetesség vizsgálat

Ebben a részben az értékelések következetességének vizsgálata kerül bemutatásra, valamint, hogy megfigyelhető-e különbség a felvételek készítésénél ott lévő szakember és a hangfájlokat visszahallgató további három szakember értékelései között. A 7. táblázaton látható, hogy a szakemberek átlagos eltérése a H átlagától 0,4. Eszerint a Szakember 2 és 4 ad leginkább átlagos értékeléseket, így az ő értékelései változtatják meg legkevésbé az átlagos H értéket, míg Szakember 1 értékelései különböznek leginkább az átlagtól. Külön kiemelendő viszont, hogy egyedül Szakember 1 volt jelen a felvételek elkészítésénél.

A Cronbach alfa a leggyakrabban használt mérőszám a belső konzisztencia mérésére, kifejezésére. A 0,8 vagy afölötti Cronbach alfa értékek magas szintű belső konzisztenciát indikálnak. Jelen esetben a Chronbach alfa 0,891-nek adódott, ami jó belső konzisztenciát mutat. Ezen Chronbach értéket egyik szakember kihagyása sem növelné. Megjegyzendő, hogy Szakember 1 értékelésének törlésével csökkenne legkisebbet a belső konzisztencia, vagyis a többi szakember törlése esetén az jelentősebben csökkenne. Az eredményeket a 8. táblázat foglalja össze.

Az Intra Class Correlation Coefficient (ICC) is megállapításra került. Az ICC 0,75 feletti értéke kiváló, míg az alatti értéke jó következetességet mutat. Az értékelésekre jó konzisztencia adódott, míg a szakemberek értékeléseinek együttes (átlagos) konzisztenciája 0,89 lett, amely kiváló konzisztenciát mutat, 95% konfidencia intervallum esetén 0,857-tól 0,917-ig ($F(137;411) = 9{,}172$; $p < 0{,}001$).

3.3. Regresszióelemzés

A regresszió jelentős előnyt jelent a klaszterelemzéshez képest, mivel a becslés szinte folyamatosan követi a függvényt. Ez a tulajdonság jelentősen javíthatja

7. táblázat. A H átlagtól való átlagos eltérések az egyes szakembereknél.

	Összesen	Szakember 1	Szakember 2	Szakember 3	Szakember 4
H átlagtól való átlagos eltérések	0,4	0,5	0,3	0,4	0,3

8. táblázat. Megbízhatósági statisztika, ha valamelyik tag törölve van.

	Elem, ha törölt	
	item-rest korreláció	**Cronbach's α**
Szakember 1	0,714	0,881
Szakember 2	0,777	0,857
Szakember 3	0,782	0,852
Szakember 4	0,787	0,850

a modell minőségét. A regressziós módszerek jóságát, a négyzetes középérték hiba (RMSE), az R^2 (a determináció együtthatója, azaz a függő változó varianciájának aránya, amelyet a független változók magyarázhatnak) és a Pearson korreláció (a cél és a prediktált H értékek között) értékével vizsgáltuk. Ebben az elemzésben lineáris regressziót és RBF kernelfüggvényű SVR regressziót használtunk.

A legjobb modell elérése érdekében több paraméter szett is kipróbálásra került bemenő jellemzővektorként. Összehasonlításra került a kiindulási paraméter szett, a klaszteranalízisben használt 18 paraméteres szett, valamint az FFS algoritmus kimenete lineáris és RBF kernelfüggvényű regresszió esetén külön. Az RBF kernelt használó regressziónál a hiperparaméter-optimalizáció grid keresével történt.

9. táblázat. Az FFS algoritmus eredménye, a négy szakember értékelésének átlaga célváltozóként.

	Kiválasztott akusztikai paraméterek
FFS lineáris kernellel	mfcc01$_{mean}$, shimmer$_{mean}$, SPI->LowVowels$_{std}$, hnr$_{std}$, SPI->HighVowels$_{mean}$, IMF->Nasal$_{mean}$, SPI->VoicedPlosives$_{std}$, IMF->DeepVowels$_{std}$
FFS RBF kernellel	shimmer$_{mean}$, hnr$_{mean}$, mfcc01$_{mean}$, hnr$_{std}$, SPI->E$_{mean}$, SPI->E$_{std}$, SPI->Nasal$_{std}$, SPI->HighVowels$_{mean}$, SPI->LowVowels$_{mean}$, SPI->LowVowels$_{std}$, SPI->VoicedPlosives$_{mean}$, IMF->Nasal$_{mean}$, IMF->VoicedPlosives$_{mean}$, IMF->VoicedPlosives$_{std}$

A 11. táblázat mutatja a regresszióelemzés eredményét, amikor a négy szakember értékelésének átlaga volt a célváltozó. Az FFS algoritmus eredményeit, lineáris és RBF-kernel esetekben a 9. táblázat foglalja össze. A lineáris regressziót használó FFS adta a legnagyobb, 0,853-as korrelációt, viszont az RBF kernel használatával érhető el a H legkisebb RMSE értéke (0,454). A legkisebb korrelációt és a legnagyobb RMSE hibát adó jellemzővektor a Baseline paramétereket használó összeállítás volt. Ezek alapján megállapítható, hogy a különböző fonetikai osztályokba tartozó beszédhangokon mért akusztikai paraméterek (SPI és

IMF entropy energiahányadosok) növelik a hangképzési rendellenességek súlyosságát prediktáló modellek jóságát. Az FFS kimenetével képzett jellemzővektorok modelljei csak kis mértékben térnek el a 18 paraméteres jellemzővektoros modelltől, így ez is mutatja annak általános jóságát a probléma esetén.

A Szakember 1 predikcióira is modell épült, melynek eredményeit a 12. táblázat mutatja, mivel ő volt jelen a felvételek elkészítésénél, ő hallotta élőben a pácienst, és állította fel a diagnózist. Az FFS algoritmus eredményeit, lineáris és RBF-kernel esetekben a 10. táblázat foglalja össze.

A 12. táblázat azt mutatja, hogy a legjobb eredményt a lineáris regresszióval kapott paraméterkiválasztás adta, a legnagyobb 0,759-es korrelációval, és a legkisebb 0,687-es RMSE hiba értékkel. Annak ellenére, hogy a korreláció 0,1-el kisebb, és a hiba értéke 0,2-vel nagyobb az előzőhöz képest, fontos különbség van a két eset között: a súlyossági osztály felosztása. Amíg az előző esetben a négy szakember átlaga 0,25-ös felbontású skálát ad, addig itt az egyetlen szakember értékelése esetén 1-es a skála felbontása a rekedtség súlyosságára. Ennek ellenére azonos következtetés vonható le jelen eredmények kapcsán is, miszerint a különböző fonetikai osztályokon mért energiahányadosok (SPI and IMF$_{entropy}$) növelni tudják a felállítható modell jóságát.

10. táblázat. Az FFS algoritmus eredménye, Szakember 1 értékelése célváltozóként.

	Kiválasztott akusztikai paraméterek
FFS lineáris kernellel	hnr$_{mean}$, IMF->Nasal$_{mean}$, shimmer$_{mean}$, mfcc01$_{mean}$, SPI->E$_{mean}$, SPI->E$_{std}$, hnr$_{std}$, SPI->VoicedPlosives$_{std}$, SPI->LowVowels$_{std}$, IMF->VoicedSpirants$_{std}$
FFS RBF kernellel	shimmer$_{mean}$, hnr$_{mean}$, mfcc01$_{mean}$, shimmer$_{std}$, hnr$_{std}$ SPI->file, SPI->E$_{mean}$, SPI->Nasal$_{mean}$, SPI->Nasal$_{std}$, SPI->HighVowels$_{mean}$, SPI->LowVowels$_{mean}$, SPI->LowVowels$_{std}$, SPI->VoicedSpirants$_{mean}$, SPI->VoicedPlosives$_{mean}$, SPI->VoicedPlosives$_{std}$, IMF->E$_{mean}$, IMF->Nasal$_{mean}$, IMF->Nasal$_{std}$, IMF->HighVowels$_{std}$, IMF->LowVowels$_{mean}$, IMF->LowVowels$_{std}$, IMF->VoicedSpirants$_{std}$, IMF->VoicedPlosives$_{mean}$

11. táblázat. Regresszió elemzés eredménye, a négy szakember értékelés átlaga célváltozóként.

Akusztikai paraméter szett	**Regresszió típusa**	**R^2**	**Korreláció**	**H RMSE értéke**	**Hiperparaméterek**
Baseline paraméter szett	Lineáris	0,639	0,799	0,531	-
18 paraméter szett	Lineáris	0,700	0,837	0,502	-
FSS eredménye	Lineáris	**0,727**	**0,853**	0,462	-
Baseline paraméter szett	RBF kernel	0,660	0,812	0,500	C = 8, γ = 0,25
18 paraméter szett	RBF kernel	0,653	0,808	0,506	C = 2, γ = 0,125
FSS eredménye	RBF kernel	0,720	0,849	**0,454**	C = 4, γ = 0,25

12. táblázat. Regresszió elemzés eredménye, Szakember 1 értékelése célváltozóként.

Akusztikai paraméter szett	Regresszió típusa	R^2	Korreláció	H RMSE értéke	Hiperparaméterek
Baseline paraméter szett	Lineáris	0,476	0,690	0,759	-
18 paraméter szett	Lineáris	0,544	0,738	0,735	-
FSS eredménye	Lineáris	**0,577**	**0,759**	**0,687**	-
Baseline paraméter szett	RBF kernel	0,413	0,643	0,786	C = 4, γ = 0,0625
18 paraméter szett	RBF kernel	0,420	0,648	0,775	C = 2, γ = 0,125
FSS eredménye	RBF kernel	0,504	0,710	0,716	C = 4, γ = 0,25

A 2. ábra mutatja a diszfónia automatikusan prediktált súlyosságát az eredeti perceptuális referencia értékekhez képest, mind a két regressziós modell esetén: a négy szakember értékelésének átlaga alapján készített és a betegeket kezelő klinikus ítélete alapján létrehozott modell. Az ábra az FFS algoritmussal kapott lineáris regresszió modellt mutatja, mindkét esetben.

2. ábra: Automatikusan becsült diszfónia súlyosság a perceptuálisan megítélt H érték alapján.

A 2. ábra ismételten bemutatja a javasolt megközelítés becslési hatékonyságát, függetlenül a páciens patológiás hátterétől és a diszfóniájának súlyosságától. A két modell hasonló predikciókat ad, viszont Szakember 1 modellje nagyobb varianciával becsül, valamint ezen modell kisebb súlyossági fokúnak prediktálta a hangokat, mint a négyértékeléses modell. Látható, hogy mind a két modell jó predikciót ad a H1 súlyossági osztályra, súlyosabb esetekben pedig alulról becsli azokat.

4. Következtetés

Jelenlegi tanulmány a diszfónia súlyosságának automatikus becslésére irányult, a szakemberi értékelések szubjektív jellegének figyelembevételével. Ez a kutatás fo-

lyamatos beszéden alapult, mivel az jobban alkalmazható a gyakorlati munkára, mint a kitartott magánhangzók [2].

Az összes beszédfelvételt négy szakértő értékelte: a szakorvos, aki kezelte a pácienst, felállította a diagnózist és meghatározta a hangjának súlyosságát a konzultációk során, valamint további három szakember, aki nem került személyes kapcsolatba a páciensekkel, csupán a hangfelvételek visszahallgatása alapján határozták meg a diszfónia súlyosságát. A diszfónia súlyosságát az RBH skála szerint határozták meg, ebben a tanulmányban az általános rekedtséget (H) vizsgáltuk.

Az eredményeink azt mutatják, hogy a különböző fonetikai osztályokon mért energiahányadosok (SPI és $\mathrm{IMF}_{entropy}$) növelni tudják a felállítható modell jóságát, vagyis hasznosak a hangképzési rendellenességek súlyosságának predikciójában.

A k-közép, felügyelet nélküli tanulási módszer segítségével négy modellt hasonlítottunk össze, külön a négy szakember értékelése szerint. A kiválasztott akusztikai paraméterek alkalmasak a szakemberek négyfokozatú értékelésének modellezésére. A H döntés pontossága a négy szakember szerint 0,49; 0,44; 0,45; 0,47. A diszfónia súlyossági ítéletek között nagyfokú megbízhatóság volt tapasztalható a Cronbach Alpha és ICC-vel való belső konzisztenciájának mérésekor.

A diszfónia súlyosságának automatikus becslése során két regressziós modellt hasonlítottunk össze: a négy szakember értékelésének átlaga által készített modellt és a betegeket kezelő klinikus külön modelljét. A két modell alacsony RMSE és magas korrelációs értékeket ért el az automatikusan becsült súlyosság és perceptuális értékelések között.

A jövőbeni munka magában foglalja további szakemberek bevonását a diszfónia súlyosságának megítélésében, valamint az adatbázis bővítését. Egy nagyobb adatbázis lehetővé teheti különböző diszfónia típusok osztályozását. Úgy gondoljuk, hogy a tanulmányban bemutatott módszer általánosítható más nyelvekre is.

Hivatkozások

1. Hirschberg, J., Hacki, T., Mészáros, K.: Foniátria és társtudományok: A hangképzés, a beszéd és a nyelv, a hallás és a nyelés élettana, kórtana, diagnosztikája és terápiája (I. kötet). Budapest: Eötvös Kiadó. (2013)
2. Kim, J., Kumar, N., Tsiartas, A., Li, M., Narayanan, S.S.: Automatic intelligibility classification of sentence-level pathological speech. Computer speech & language **29**(1) (2015) 132–144
3. Liu, Y., Lee, T., Ching, P., Law, T.K., Lee, K.Y.: Acoustic assessment of disordered voice with continuous speech based on utterance-level asr posterior features. Proc. Interspeech 2017 (2017) 2680–2684
4. Vicsi, K., Imre, V., Mészáros, K.: Voice disorder detection on the basis of continuous speech. In: 5th European Conference of the International Federation for Medical and Biological Engineering, Springer (2011) 86–89
5. Kazinczi, F., Mészáros, K., Vicsi, K.: Automatic detection of voice disorders. In: International Conference on Statistical Language and Speech Processing, Springer (2015) 143–152

6. Grygiel, J., Strumiłło, P., Niebudek-Bogusz, E.: Application of mel cepstral representation of voice recordings for diagnosing vocal disorders. Delta **12** (2012) 2
7. Tulics, M.G., Vicsi, K.: Phonetic-class based correlation analysis for severity of dysphonia. In: Cognitive Infocommunications (CogInfoCom), 2017 8th IEEE Conference on, IEEE (2017) 21–26
8. Chien, Y.R., Borskỳ, M., Guðnason, J.: Objective severity assessment from disordered voice using estimated glottal airflow. Proc. Interspeech 2017 (2017) 304–308
9. Laaridh, I., Kheder, W.B., Fredouille, C., Meunier, C.: Automatic prediction of speech evaluation metrics for dysarthric speech. Proc. Interspeech 2017 (2017) 1834–1838
10. Schönweiler, R., Hess, M., Wübbelt, P., Ptok, M.: Novel approach to acoustical voice analysis using artificial neural networks. JARO-Journal of the Association for Research in Otolaryngology **1**(4) (2000) 270–282
11. Ptok, M., Schwemmle, C., Iven, C., Jessen, M., Nawka, T.: On the auditory evaluation of voice quality. HNO **54**(10) (2006) 793–802
12. Kiss, G., Sztaho, D., Vicsi, K.: Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features. In: Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on, IEEE (2013) 579–582
13. Roussel, N.C., Lobdell, M.: The clinical utility of the soft phonation index. Clinical linguistics & phonetics **20**(2-3) (2006) 181–186
14. Tsanas, A.: Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms. Models and analysis of vocal emissions for biomedical applications **2** (2013) 37–40

Az enyhe kognitív zavar és korai Alzheimer-kór automatikus azonosítása spontán beszédből akusztikus jellemzők segítségével

Gosztolya Gábor [1], Hoffmann Ildikó [2,3], Tóth László [4], Vincze Veronika [1,5], Pákáski Magdolna [6], Kálmán János [6]

[1]MTA-SZTE Mesterséges Intelligencia Kutatócsoport, Szeged
[2]Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék
[3]MTA Nyelvtudományi Intézet, Budapest
[4]Szegedi Tudományegyetem, Informatikai Intézet
[5]Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék
[6]Szegedi Tudományegyetem, Pszichiátriai Klinika
{ ggabor, tothl, vinczev } @ inf.u-szeged.hu

Kivonat Az enyhe kognitív zavar (EKZ) olyan tünetegyüttes, melynek fontos szerepe van néhány demenciatípus, például egy később lehetségesen kialakuló Alzheimer-kór korai megjósolásában, és így a kezelés minél korábbi elkezdésében. Korábbi kutatásainkban megmutattuk, hogy a korai vagy enyhe Alzheimer-kór (eAK) hatékonyan felismerhető a vizsgálati személy spontán beszéde alapján a megfelelő beszédjellemzők kinyerése révén. Ebben a cikkben a korábban bemutatott, beszédfelismerésen alapuló automatikus jellemzőkinyerést alkalmazzuk az eAK-ban szenvedő és az egészséges (kontroll) beszélők megkülönböztetése mellett az EKZ-val élők felismerésére is. A kísérleti eredmények alapján a kontroll csoport és az eAK beszélők megkülönböztetése történik a legnagyobb hatékonysággal, ugyanakkor az EKZ-val élő betegek felismerése is hatékony: a páronkénti pontosságértékek 74% és 84% közé esnek.
Kulcsszavak: beszédfelismerés, beszédfeldolgozás, demencia, enyhe kognitív zavar, korai Alzheimer-kór

1. Bevezetés

A demenciában szenvedő személyek száma az életkor kitolódása miatt az elkövetkező néhány évtizedben várhatóan megsokszorozódik [1]. Mindeközben a demencia korai felismerése nem egyszerű, és a diagnosztizált esetek aránya is alacsony. A demencia egyik leggyakoribb típusa az Alzheimer-kór (AK), melynek kezelése a betegség korai szakaszában a leghatékonyabb. A korai felismerés fontos a későbbi kezelés szempontjából, hiszen így lassítható a betegség előrehaladása és javítható a beteg életminősége.

A dementálódás folyamata már negyvenéves kor körül elkezdődhet az úgynevezett enyhe kognitív zavar (EKZ) tünetegyüttessel. Az enyhe kognitív zavar a

különböző mentális funkciók egy megelőző magasabb szintjéhez képest tapasztalható kezdődő, enyhe változás. Ez megjelenhet a rövidtávú emlékezet kapacitásának finom korlátozódásában, a szókeresés gyakoriságának enyhe növekedésében vagy figyelmi zavarokban. Az EKZ az egészséges és a demens állapot közötti skálán jelenik meg; előjele lehet egy később kialakuló demenciának. A demencia ugyanakkor folytonosan meglévő kognitív deficitet jelent a mentális funkciók legalább három területén, mely az emlékezetet, a nyelvet, a téri-vizuális készséget, az absztrakt gondolkodást, a számolást, az ítélőképességet, az érzelmi állapotot vagy a személyiség megváltozását érintheti [2,3,4].

A klinikai pszichiátria egyik legnehezebb feladata az EKZ diagnosztizálása. Habár a nemzetközi szakirodalom és a klinikai gyakorlat számos korai felismerést célzó diagnosztikai tesztet ismer — pl. Mini Mentál Teszt (MMT, [5]), Órarajzolási Teszt (Clock Drawing Test, CDT [6,7]), ADAS-Cog ([8]) —, ezek vagy túl időigényesek, vagy nem elég érzékenyek ahhoz, hogy felismerjék azokat a finom változásokat, melyek az EKZ-ban megjelennek, így eredményeik is gyakran egymásnak ellentmondóak. Léteznek ugyanakkor olyan pszicholingvisztikai és neurolingvisztikai tesztek, mint például a szótalálás vagy a történetfelidézés, melyek jelentősen hatékonyabbnak bizonyulnak az EKZ felismerésére, ugyanakkor ezen tesztek eredményei félre is diagnosztizálhatnak [9].

Újabb tanulmányok rámutatnak arra, hogy az AK szubklinikus szakaszában, azaz az EKZ-ban már tapasztalhatók nyelvi változások. Ezek a változások érinthetik a beszéd temporális jellemzőit [9,10,11,12,13,14,15], a mentális lexikon elemeinek lehívását [16,17,18,19,20,15] és a verbális fluenciát [21,9,18,14] is.

Jelen cikkünkben a beszéd temporális paramétereinek elemzésével foglalkozunk, melynek segítségével olyan beszéd közbeni kognitív folyamatokat vizsgálhatunk, mint a nyelv és a memória közti kapcsolatok. Végső célunk egy automatikus szoftveres alkalmazás kifejlesztése, mely alkalmas a potenciális páciensek előszűrésére, elősegítve az EKZ-ban és az enyhe AK-ban szenvedők korai diagnosztizálását. Erre egy, korábbi tanulmányainkban ([11,22,23,24]) már bemutatott automatikus kétlépéses eljárást alkalmazunk. Jelen cikkünkben az enyhe vagy korai AK (eAK) felismerése mellett célunk az EKZ detektálása is, valamint a két betegcsoport egymástól való elkülönítése.

2. A felhasznált adatbázis

A hangfelvételeket a Szegedi Tudományegyetem Pszichiátriai Klinikáján rögzítettük az EKZ és korai AK által leginkább veszélyeztetett korosztálytól. A kutatást a Szegedi Tudományegyetem Etikai Bizottsága hagyta jóvá, és azt a Helsinki Nyilatkozat betartásával végeztük. Három beszélőcsoporttól rögzítettünk felvételeket: EKZ-val és korai AK-val élőktől, valamint olyan kontroll beszélőktől, akik a felvétel időpontjában nem szenvedtek semmilyen mentális betegségben. A három csoportot korban, nemben és iskolázottságban illesztettük. A kontroll csoport alkohol vagy mentális funkciókat befolyásoló gyógyszerhatás alatt nem állt, sem látási, sem hallási deficittel nem rendelkezett. Kizártuk azokat, akik korábban fejsérülést szenvedtek, és azokat is, akik depressziós tüneteket mutattak.

	Beszélőcsoportok			Stat.
	Kontroll	**EKZ**	**eAK**	p
Életkor	70,72 ± 5,004	72,4 ± 3,594	73,96 ± 6,846	$p = 0{,}105$
Iskolázottság (év)	12,08 ± 2,326	10,84 ± 2,304	10,76 ± 2,818	$p = 0{,}118$
MMT érték	29,24 ± 0,523	27,16 ± 0,898	23,92 ± 2,488	$p < 0{,}001$
CDT érték	8,88 ± 2,007	6,44 ± 3,429	5,88 ± 3,244	$p = 0{,}001$
Adas-COG érték	8,575 ± 2,374	12,044 ± 3,205	18,675 ± 5,818	$p < 0{,}001$

1. táblázat. A vizsgálati csoportok demográfiai adatai.

Az EKZ és eAK személyek orvosilag diagnosztizáltak voltak, velük CT, MRI és általános kognitív állapotot felmérő tesztek készültek (MMT, Órarajzolási teszt, ADAS-Cog).

Mivel mind saját korábbi kutatásaink ([11,22,23,24]), mind más kutatócsoportok által végzett kutatások (például [17,9,13]) azt erősítették meg, hogy az EKZ és az eAK jobban befolyásolja a beszélő spontán beszédét, mint az előre eltervezettet, célunk spontán beszéd rögzítése volt. Emiatt jelen kísérletünkben a korábbi munkánkban bemutatott módon késztettük pácienseinket spontán beszédre, az alábbi forgatókönyv szerint. Miután megnéztek egy egyperces animációs filmet, tesztalanyainkat megkértük, hogy meséljék el a filmben látottakat (*azonnali felidézés*). Ezután a vizsgálati személyeket megkértük, hogy meséljék el a tegnapi napjukat (*spontán beszéd*). Végül egy második egyperces filmet vetítettünk le, majd a tesztszemélyeknek egyperces várakozás után kellett elmesélniük annak tartalmát (*késleltetett felidézés*). (A részletekért ld. Hoffmann és mtsai cikkét [11].)

EKZ és korai AK vizsgálati személyeink hangfelvételeit tartalmazó adatbázisunk folyamatosan gyarapszik; jelenleg több, mint 150 beszélő spontán beszédét tartalmazza. Ugyanakkor különböző okokból (pl. ellentmondásos diagnózis vagy rossz hangminőség) néhány személy felvételeit ki kellett zárnunk a további vizsgálatokból. Ezenfelül, mivel a három beszélői csoportot illeszteni kívántuk életkor, nem és iskolázottság szerint, további hangfelvételeket kellett kihagynunk jelen tanulmányunkból. Végül minden beszélői csoportból 25-25-25 személy anyagait használtuk, így kísérleteinket összesen 75 beszélő 225 hangfelvételén végeztük. Az 1. táblázat tartalmazza a kísérletben vizsgált alanyok néhány fontosabb adatát.

3. EKZ felismerésére használt beszédjellemzők

A spontán beszéd időviszonyainak elemzése különösen érzékeny neuropszichológiai módszernek bizonyult a kognitív folyamatok (mint például a beszéd) vizsgálatára [11]. A spontán beszéd temporális paramétereinek vizsgálata nagy pontossággal teszi mérhetővé a beszélő személy beszéd- és nyelvi képességeit [25,26].

Több tanulmány (pl. [21,27]) is megmutatta, hogy a demencia megléte növeli a hezitálás mennyiségét és időtartamát a spontán beszédben. Korábbi munká-

(1)	Artikulációs tempó (hezitációk nélkül számított másodpercenkénti beszédhang-szám).
(2)	Beszédtempó (hezitációkkal együtt számított másodpercenkénti beszédhang-szám).
(3)	A felvétel hossza, ezredmásodpercben mérve.
(4-5)	Kitöltött, illetve kitöltetlen szünetek összhossza.
(6-7)	Kitöltött, illetve kitöltetlen szünetek száma.
(8)	Hezitációs ráta (a kitöltött és kitöltetlen szünetek összhosszának és a felvétel hosszának aránya).

2. táblázat. Az alapvető akusztikus jellemzők, Hoffmann és mtsai [11] nyomán.

inkat (pl. [11]) követve, első lépésként a következő nyolc akusztikus jellemzőt határoztuk meg a manuális méréseinkhez, melyeket a Praat program segítségével végeztünk el: artikulációs tempó (1. jellemző), beszédtempó (2. jellemző), a jelidő hossza (3. jellemző), kitöltött, illetve kitöltetlen szünetek összhossza (4-5. jellemzők), kitöltött, illetve kitöltetlen szünetek száma (6-7. jellemzők), hezitációs ráta, azaz az összes szünet aránya a jelidőben (8. jellemző). Hezitáción a 30 ms hosszúságúnál nagyobb szünetet értjük [28]. A nyolc alapvető akusztikus paramétert és kiszámításuk módját a 2. táblázat foglalja össze.

3.1. A jellemzők automatikus kinyerése beszédfelismerési technikákkal

A fönti jellemzők manuális kiszámítása meglehetősen munkaigényes, ráadásul hozzáértő annotátorokat igényel. Ezért kívánatos a jellemzők kinyerésének automatizálása, melyet mi beszédfelismerési technikák használatával oldottunk meg. A jellemzők egy része ugyan egyszerű jelfeldolgozási eszközökkel is meghatározható (pl. beszéd/csend részek elkülönítése), azonban a beszédtempó és egyéb más, a beszédhangok hosszán alapuló jellemzők kinyerésére ezek alkalmatlanok. Ezen felül a kitöltött szüneteket is képtelenek megkülönböztetni az alany beszédétől.

A föntiek miatt egy beszédfelismerő rendszert tanítottunk be egy spontán beszédet tartalmazó adatbázisra. Vegyük észre, hogy a korábban felsorolt jellemzők kiszámításához általában nem szükséges az egyes beszédhangok *megkülönböztetése*, csupán azok *megszámolása*, illetve hosszának megmérése. Mivel a beszédhangok téves azonosítása általában nem okoz gondot, ezért egy egyszerűsített beszédfelismerő modellt használtunk, amely nagy pontossággal képes megtalálni és azonosítani a beszédjelben szereplő beszédhangokat [29]. A kitöltött szüneteket önálló „fonémának” tekintettük, így a beszédfelismerő rendszer kimenete minden felvételre egy időzített beszédhang-sorozat. Ennek alapján a korábban javasolt jellemzőket már automatikusan is ki tudjuk számítani. A beszédjellemzők alapján a beszélőket valamilyen gépi tanulási eljárás (pl. SVM, véletlen erdő) segítségével, automatikusan különítjük el (ld. 1. ábra).

1. ábra: Az alkalmazott jellemzőkinyerési és osztályozási megközelítés manuális (bal oldal) és automatikus (jobb oldal) jellemzőkinyerés esetén.

Korábbi tanulmányaink ([22,23,24]) kísérleti eredményei igazolták, hogy a fönti automatikus jellemzőkinyerés alkalmazható az EKZ-ban szenvedő beszélők nagypontosságú megkülönböztetésére az egészséges kontroll alanyoktól. Jelen munkánkban ugyanezt a jellemzőkészletet és jellemzőkinyerési megközelítést használjuk, viszont most mindhárom beszélőcsoportot (EKZ, korai AK és kontroll) el szeretnénk különíteni egymástól.

3.2. Kibővített jellemzőkészlet

A fönti (ld. 2. táblázat) jellemzők manuális számításokra lettek tervezve. Ugyanakkor a 3.1. fejezetben ismertetett automatikus beszédfelismerésen alapuló módszer a felvételhez első lépésben egy időzített fonémasorozatot határoz meg, majd

(1) Az adott fonéma előfordulásainak száma osztva a felvétel összfonémaszámával.
(2) Az adott fonéma összhossza osztva a felvétel teljes hosszával.
(3) Az adott fonéma előfordulásainak átlaghossza.
(4) Az adott fonéma előfordulásainak hosszainak szórása.

3. táblázat. A négy statisztikai fonémajellemző, Tóth és mtsai ([22]) nyomán.

abból számítja ki a felvételenként nyolc beszédjellemzőt. A folyamat legidőigényesebb lépése nyilvánvalóan a beszédfelismerő rendszer alkalmazása, így a beszédjellemzők tényleges meghatározása csak elhanyagolható arányban járul hozzá a teljes elemzés futási idejéhez. Ezen a ponton könnyen bővíthetjük a számolt beszédjellemzők listáját további jellemzőkkel, amennyiben azok az időzített fonémasorozatból számolhatóak. Ez a felismerés vezetett minket arra, hogy további, az EKZ (és remélhetőleg a korai AK) felismerését segítő jellemzőket adjunk a meglévő jellemzőhalmazhoz (*kiterjesztett* jellemzőkészlet).

Első lépésben megtartottuk a 2. táblázatban felsorolt jellemzőket, bár a (6)-os és (7)-es jellemzőket kicsit módosítottuk: ahelyett, hogy a kitöltött és kitöltetlen szünetek nyers darabszámára hagyatkoztunk volna, normalizált értékeket számoltunk úgy, hogy a szünetek számát elosztottuk az adott hangfelvétel fonémaszámával. A következő lépésben, mivel már rendelkezésünkre állt a kitöltött és kitöltetlen szünetek minden egyes előfordulásának hossza, a beszédjellemzők közé felvettük ezen hosszok átlagát és szórását is.

Észrevettük továbbá, hogy a beszédfelismerő rendszer gyakran keverte össze a kitöltött szüneteket bizonyos fonémákkal; például az „öööö” hezitáció a fonetikai szintű kimenetben többször egyszerű „ö” fonémaként jelent meg, míg a „hmm” hezitáció rövid vagy hosszú „m”-ként. Ez alapján azt feltételeztük, hogy ezen fonémák számának, illetve összhosszának növekedése a fonetikai szintű beszédfelismerő-kimenetben valószínűleg a kitöltött szünetek számának gyarapodását jelzi, melyeket az akusztikus modell rosszul kategorizált. Így akusztikus jellemzőkészletünket olyan jellemzőkkel bővítettük ki, melyek ezen fonémák adott hangfelvételen belüli gyakoriságát írják le: az „m”, „n” és „ö” fonémák előfordulásaiból négy-négy jellemzőt számoltunk (ld. 3. táblázat). Ezzel a kiterjesztéssel a három felvételre összesen 81 jellemzőt kaptunk, melyre kísérleteink során mint a *kibővített* jellemzőkészletre fogunk hivatkozni.

Az EKZ és korai AK kockázatai különböznek férfiak és nők esetében, és a korral is változnak [30]. Mivel mind a nem, mind az életkor rendelkezésünkre állt a kísérletben részt vevő összes alany esetében, ezeket a jellemzőket és az iskolázottságot (iskolában elvégzett évek számát) is hozzávettük a jellemzőkészlethez, így alap jellemzőkészletünk 27, míg kiterjesztett jellemzőhalmazunk végül 84 jellemzőből állt. Habár mind a nem, mind az életkor elég megbízhatóan meghatározható a beszédhangból (ld. pl. [31,32]), a tervezett alkalmazásban egyszerűen meg is kérhetjük az alanyt, hogy adja meg ezeket az adatokat.

4. Kísérletek és eredmények

Az automatikus beszédfelismerő rendszert a BEA Spontánbeszéd-adatbázis [33] kb. hétórányi beszédadatára tanítottuk be. Mivel fontos volt, hogy a kitöltött szüneteket képesek legyünk azonosítani, az elérhető annotációt módosítanunk kellett, hogy tartalmazzon bizonyos, a spontán beszédben előforduló elemeket (pl. kitöltött szüneteket, be- és kilégzéseket, nevetést, köhögést). Fonémaszintű felismerést végeztünk, a kitöltött szüneteket egy speciális fonémának tekintve. Akusztikus modellként egy standard előrecsatolt (feed-forward) neurális hálót alkalmaztunk három rejtett réteggel, rétegenként ezer ún. rectified aktivációs függvényt használó neuronnal. Nyelvi modellünk egy egyszerű fonémabigram volt, mely tartalmazta a fönt felsorolt nemverbális hangjelenségeket (kitöltött szünetek, köhögések stb.).

A beszédfelismerő rendszer kimenete egy időzített, fonetikai szintű kimenet volt, mely tartalmazta mind a kitöltött, mind a kitöltetlen szüneteket. Ebből a fonetikai szintű átiratból már mind az alapvető, mind a kibővített jellemzőkészletet könnyen ki lehetett számítani.

4.1. Beszélőosztályozás

Gépi tanulási szemszögből egy 75 példából (most: alanyból) álló adatbázis nagyon kicsinek számít. Azonban a diagnosztizált EKZ és korai AK alanyok száma erősen véges, spontán beszédük rögzítése fix protokoll mellett pedig nagyon munkaigényes. Az általunk ismert többi, hasonló területet vizsgáló tanulmány (pl. [13,14,34,9]) mind 100-nál kevesebb beszélőt vizsgált.

Mivel ilyen kisméretű adatbázissal dolgoztunk, az alanyokat nem osztottuk tovább külön tanító- és teszthalmazra, hanem ötszörös keresztvalidálást (cross-validation, CV) alkalmaztunk. Hogy minden esetben minden beszélőcsoportból ugyanannyi jusson mind az aktuális tanító-, mind az aktuális teszthalmazba: mindig 60 beszélő felvételein tanítottunk, melyből 20 EKZ-ban, 20 enyhe AK-ban szenvedett, míg 20 a kontrollcsoportba tartozott. Ezután a betanított gépi tanulási modellt kiértékeltük a fennmaradó 15 beszélőre. Ezt ötször megismételve az összes beszélőre kaptunk egy diagnózis hipotézist, melyeket egyetlen végső pontosságértékben összegeztünk.

Osztályozásra szupportvektor-gépet (Support-Vector Machines, SVM [35]) alkalmaztunk lineáris kernellel; a Weka ingyenesen elérhető programcsomagban [36] található *SMO* implementációt használtuk.

4.2. Kiértékelés

A kiértékelési metrika kiválasztása nem egyértelmű ezen a feladaton. Használhatjuk a hagyományos osztályozási pontosságot, hiszen jelen esetben az osztáleloszlás tökéletesen kiegyensúlyozott. Ugyanakkor amellett, hogy azt számoljuk ki, mennyire pontosan sikerült azonosítanunk a három beszélőcsoport tagjait, az adott feladatot detektálási feladatnak is tekinthetjük, melyben a kérdés az,

Jellemző-kinyerés	Jellemző-készlet	3-oszt.	2-osztályos				
		Pont.	Pont.	UAR	Prec.	Fedés	F_1
Manuális	Alapvető	61,3%	76,0%	74,0%	83,3%	80,0%	81,6%
Automatikus	Alapvető	61,3%	70,7%	67,0%	78,0%	78,0%	78,0%
	Kibővített	62,7%	74,7%	73,0%	83,0%	78,0%	80,4%

4. táblázat. A különböző akusztikus jellemzőkészletek használatával elért pontosságértékek háromosztályos modellt tanítva.

hogy az adott beszélőnek van-e *bármilyen* mentális betegsége. Ezt a gyakorlatban úgy a legegyszerűbb megoldani, hogy az EKZ és a korai AK csoportokat egyesítjük. Mivel ekkor az osztályok példaszámai kiegyensúlyozatlanná válnak (hiszen 25 kontroll és 50 valamilyen fajta demenciában szenvedő alany van), a (kétosztályos) osztályozási pontosság mellett az átlagos fedést (*Unweighted Average Recall*, *UAR*) is feltüntetjük. Emellett a szokásos információ-visszakeresési metrikákat is kiszámítottuk: pontosságot (*precision*, *Prec.*), fedést (*recall*, *Rec.*) és F-értéket (*F-measure* vagy F_1). Az SVM C meta-paraméterét úgy állítottuk be, hogy a legmagasabb (háromosztályos) osztályozási pontosságot kapjuk; azonos pontosságértékek esetén a legmagasabb F-értékkel rendelkező konfigurációt választottuk.

4.3. Eredmények

A 4. táblázat mutatja az osztályozási eredményeket a különböző akusztikus jellemzőkészletek használata mellett. Első pillantásra a háromosztályos pontosságértékek alacsonynak tűnhetnek, de tekintve, hogy három osztállyal dolgoztunk (EKZ, korai AK és kontroll) és az alanyok eloszlása is kiegyenlített volt, találgatással csak 33,3%-ot érhetünk el. Ennek figyelembe vételével a manuálisan számított jellemzők elfogadható pontossághoz vezettek: a 61,3%-os érték ezt jelentősen meghaladja. A feladatot kétosztályos detektálási feladatként értelmezve, az előálló pontosságértékek elég magasak, ami azt jelzi, hogy a téves osztályozások nagy része az EKZ és korai AK osztályok keveréséből adódhatott.

Ugyanezen (*alapvető*) jellemzők automatikus meghatározásával is 61,3%-os (háromosztályos) pontosságértéket kaptunk, bár a kétosztályos detektálási metrikák valamivel alacsonyabbnak adódtak. A kibővített jellemzőkészlet használatával viszont az összes pontosságérték nőtt (a fedés kivételével, mely változatlan maradt): bár a (háromosztályos) pontosság csak 1,4%-kal növekedett, az F-érték 2,4%-os (abszolút) növekedése azt jelzi, hogy ezzel a jellemzőkészlettel nagyobb pontosságú demenciadetektálásra van lehetőség. Ráadásul ezen jellemzőkészlet manuális kiszámítása irreálisan sok erőfeszítést igényelt volna, míg automatikusan csak elhanyagolhatóan nőtt a futási idő.

A 4. táblázat alapján a kibővített jellemzőkészlettel érhetőek el a legjobb eredmények. Ugyanakkor nem minden hiba feltétlenül azonos súlyú; érdemes lehet azt is megvizsgálni, hogy az egyes beszélőcsoportok milyen hatékonysággal

Jellemzőkészlet	**Pont.**	**UAR**	**Prec.**	**Fedés**	F_1
Kontroll vs. EKZ + eAK	74,7%	72,0%	81,6%	80,0%	80,8%
Kontroll vs. EKZ	76,0%	76,0%	84,2%	64,0%	72,7%
Kontroll vs. eAK	84,0%	84,0%	81,5%	88,0%	84,6%
EKZ vs. eAK	76,0%	76,0%	—	—	—

5. táblázat. A különböző kétosztályos gépi tanulási esetekben elért pontosságértékek, a kibővített akusztikai jellemzőkészlet használata mellett.

különíthetőek el egymástól. Habár a tévesztési mátrix elemzésével is vonhatunk le erről következtetéseket, szerintünk ennél jobb módszer az egyes osztálypárokra bináris osztályozó modelleket tanítani és azok teljesítményét vizsgálni. Így utolsó tesztsorozatunkban minden lehetséges osztálypárra kétosztályos SVM modelleket tanítottunk. Az első tesztesetben mind a 75 beszélő adatait felhasználtuk, de az EKZ-val és korai AK-val diagnosztizált pácienseket ugyanabba az osztályba (*demens alanyok*) soroltuk, míg a másik osztály a kontroll csoport volt. A fennmaradó három esetben csak két-két beszélőcsoport 25-25 alanyának felvételeit használtuk. Mivel a korábbi kísérletben a kibővített jellemzőkészlet vezetett a legjobb eredményekhez, most is ezeket a jellemzővektorokat használtuk.

Az 5. táblázat tartalmazza a bináris osztályozó modellek használatával kapott eredményeket. Mikor mind a 75 beszélő adatait használtuk, az eredmények hasonlónak adódtak, mint mikor háromosztályos SVM modelleket tanítottunk és a predikciókban a két demens osztály címkéit összevontuk: a megegyező (74, 7%) pontosságérték mellett valamivel alacsonyabb UAR-t és pontosságot (precision) kaptunk, míg a fedés és az F-érték kicsit nőtt. A további eredményeket megvizsgálva azt láthatjuk, hogy a kontroll csoportba tartozó alanyokat volt a legkönnyebb elkülöníteni az enyhe AK-s páciensektől, amely elég logikus, hiszen ez a két csoport található a vizsgált mentális betegségskála két végén. Ehhez képest annak eldöntése, hogy az alany EKZ-ben szenved-e vagy kontroll, illetve hogy EKZ-ben vagy enyhe AK-ban szenved-e, hasonlóan nehéznek bizonyult, bár a 76%-os pontosságértékek véleményünk szerint hatékony elkülönítést tükröznek. (Mikor az EKZ és enyhe AK csoportokba tartozó alanyokat próbáltuk megkülönböztetni egymástól, nem számítottunk ki minden pontosságértéket: mivel ebben az esetben nem volt kézenfekvő pozitív osztályunk, nem láttuk értelmét az információ-visszakeresési jellemzőknek.)

Összességében az eredmények azt tükrözik, hogy a kifejlesztett jellemzőkészlet és a javasolt kétlépéses eljárás sikeresen alkalmazható mind korai enyhe kognitív zavar, mind korai Alzheimer-kór felismerésére. Célunk, hogy az akusztikus jellemzőkre épülő elemzést a későbbiekben megfelelő nyelvészeti jellemzőkkel is kiegészítsük, hiszen a beszélő bizonytalanságát egyes kifejezések (pl. „azt hiszem”, „úgy gondolom”) jelenléte is jelzi (ld. pl. [37]). Ez azonban egyértelműen további kutatásokat igényel.

5. Összegzés

A különböző demenciatípusok, így például az enyhe kognitív zavar és az Alzheimer-kór korai felismerése kritikus a kezelés mielőbbi megkezdése, és így a páciens életminősége szempontjából. Korábbi tanulmányainkban már javasoltunk egy manuális eljárást, mely a beszéd temporális paramétereinek (beszédtempó, artikulációs tempó, hezitációs ráta stb.) vizsgálatával kísérli meg elkülöníteni az EKZ-ban szenvedő alanyokat az egészséges kontrolloktól, majd javasoltunk egy beszédfelismerésen alapuló eljárást az akusztikus jellemzők automatikus kiszámítására is. A diagnózis-hipotézist második lépésben, ezen jellemzők alapján valamely gépi tanulási eljárás szolgáltatta. Jelen dolgozatunkban azt vizsgáltuk, hogy ezek a technikák alkalmasak-e az EKZ mellett a korai AK felismerésére is. Eredményeink azt tükrözik, hogy az eljárás használatával hatékonyan (80% fölötti F-értékkel) meg lehet határozni, hogy az alany rendelkezik-e bármilyen demenciával. A három beszélőcsoport (EKZ, korai AK és kontroll) páronkénti elkülönítésére is $72-84\%$ közötti pontosságértékek adódtak; nem meglepő módon az enyhe AK és a kontroll csoportot volt a legkönnyebb elkülöníteni egymástól, mivel ez a két csoport található a vizsgált mentális betegségskála két végén.

Köszönetnyilvánítás

A kutatást az EFOP-3.6.1- 16-2016-00008 azonosítójú, EU társfinanszírozású projekt támogatta. Tóth László kutatásait az MTA Bolyai János ösztöndíja támogatta. Vincze Veronika kutatásait az Emberi Erőforrások Minisztériuma ÚNKP-17-4 kódszámú Új Nemzeti Kiválóság Programja támogatta.

Hivatkozások

1. Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.: The global prevalence of dementia: a systematic review and metaanalysis. Alzheimer's and Dementia **9**(1) (2013) 63–75
2. Cummings, L., Mega, S.: Neuropsychiatry and Behavioral Neuroscience. Oxford University Press (2003)
3. Reichman, W., Cummings, J.: Dementia. In Duthie, E., Katz, P., eds.: Practice of Geriatrics. W.B. Saunders, Philadelphia, PA (1999) 395–304
4. Turi, Zs., Németh, D., Hoffmann, I.: Nyelv és emlékezet. In Pléh, Cs., Lukács, Á., eds.: Pszicholingvisztika 1-2. Akadémiai Kiadó, Budapest (2014) 743–776
5. Folstein, M., Folstein, S., McHugh, P.: Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research **12**(3) (1975) 189–198
6. Freedman, M., Leach, L., Kaplan, E., Winocur, G., Shulman, K., Delis, D.: Clock Drawing: A Neuropsychological Analysis. New York: Oxford University Press (1994)
7. Kálmán, J., Maglóczky, E., Janka, Z.: Óra Rajzolási Teszt: gyors és egyszerű dementia szűrőmódszer. Psychiatria Hungarica **10**(3) (1995) 11–18

8. Rosen, W., Mohs, R., Davis, K.: A new rating scale for Alzheimer's disease. Journal of Psychiatric Research **141**(11) (1984) 1356–1364
9. Roark, B., Mitchell, M., Hosom, J.P., Hollingshead, K., Kaye, J.: Spoken language derived measures for detecting mild cognitive impairment. IEEE Transactions on Audio, Speech, and Language Processing **19**(7) (2011) 2081–2090
10. Forbes-McKay, K., Venneri, A.: Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. International Journal of Geriatric Psychiatry **2005**(26) (2005) 243–256
11. Hoffmann, I., Németh, D., Dye, C.D., Pákáski, M., Irinyi, T., Kálmán, J.: Temporal parameters of spontaneous speech in Alzheimer's disease. International Journal of Speech-Language Pathology **12**(1) (2010) 29–34
12. Meilán, J.J.G., Martínez-Sánchez, F., Carro, J., Sánchez, J.A., Pérez, E.: Acoustic markers associated with impairment in language processing in Alzheimer's disease. Spanish Journal of Psychology **15**(2) (2012) 487–494
13. Satt, A., Hoory, R., König, A., Aalten, P., Robert, P.H.: Speech-based automatic and robust detection of very early dementia. In: Interspeech, Szingapúr (2014) 2538–2542
14. Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L., Ogar, J.: Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In: CLPsych, Baltimore, USA (2014) 27–37
15. Laske, C., Sohrabi, H.R., Frost, S.M., de Ipiña, K.L., Garrard, P., Buscema, M., Dauwels, J., Soekadar, S.R., Mueller, S., Linnemann, C., Bridenbaugh, S.A., Kanagasingam, Y., Martins, R.N., O'Bryant, S.E.: Innovative diagnostic tools for early detection of Alzheimer's disease. Alzheimer's & Dementia **11**(5) (2015) 561–578
16. Garrard, P., Maloney, L.M., Hodges, J.R., Patterson, K.: The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain **128**(2) (2005) 250–260
17. Taler, V., Phillips, N.: Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. Journal of Clinical and Experimental Neuropsychology **30**(5) (2008) 501–556
18. Santos, V.D., Thomann, P.A., Wüstenberg, T., Seidl, U., Essig, M., Schröder, J.: Morphological cerebral correlates of CERAD test performance in mild cognitive impairment and Alzheimer's disease. Journal of Alzheimer's Disease **23**(3) (2011) 411–420
19. Cardoso, S., Silva, D., Maroco, J., de Mendonca, A., Guerreiro, M.: Non-literal language deficits in Mild Cognitive Impairment. Psychogeriatrics **14**(4) (2014) 222–228
20. Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., Gorno-Tempini, M.L.: Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. Cortex **55** (2014) 122–129
21. Barth, S., Schönknecht, P., Pantel, J., Schröder, J.: Mild cognitive impairment and Alzheimer's disease: an investigation of the CERAD-NP test battery. Fortschritte der Neurologie-Psychiatrie **73**(10) (2005) 568–576
22. Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Bíró, E., Zsura, F., Pákáski, M., Kálmán, J.: Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: Interspeech, Drezda, Németország (2015) 2694–2698
23. Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Bíró, E., Zsura, F., Pákáski, M., Kálmán, J.: Az enyhe kognitív zavar automatikus azonosítása. In: Neumann Kollokvium, Veszprém, Magyarország (2015) 112–117

24. Gosztolya, G., Tóth, L., Grósz, T., Vincze, V., Hoffmann, I., Szatlóczki, G., Pákáski, M., Kálmán, J.: Detecting Mild Cognitive Impairment from spontaneous speech by correlation-based phonetic feature selection. In: Interspeech, San Francisco, USA (2016) 107–111
25. Baum, S.R., Blumstein, S.E., Naeser, M.A., Palumbo, C.L.: Temporal dimensions of consonant and vowel production: An acoustic and CT scan analysis of aphasic speech. Brain and Language **39**(1) (1990) 33–56
26. Illes, J.: Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, Huntington's, and Parkinson's. Brain and language **37**(4) (1989) 628–642
27. Juncos-Rabadán, O., Pereiro, A.X., Facaly, D., Rodríguez, N.: Una revisión de la investigación sobre lenguaje en el deterioro cognitivo leve. Revista de Logopedia, Foniatría y Audiología **30**(2) (2010) 73–83
28. Gósy, M.: The paradox of speech planning and production. Magyar Nyelvőr **12**(1) (1998) 3–15
29. Tóth, L.: Phone recognition with hierarchical Convolutional Deep Maxout Networks. EURASIP Journal on Audio, Speech, and Music Processing **2015**(25) (2015) 1–13
30. Sachdev, P., Lipnicki, D., Crawford, J., Reppermund, S., Kochan, N., Trollor, J., Draper, B., Slavin, M., Kang, K., Lux, O., Brodaty, K.M.H.: Risk profiles for mild cognitive impairment vary by age and sex: the Sydney memory and ageing study. The American Journal of Geriatric Psychiatry **20**(10) (2012) 854–865
31. Kockmann, M., Burget, L., Cernocký, J.: Brno University of Technology system for Interspeech 2010 Paralinguistic Challenge. In: Interspeech, Makuhari, Chiba, Japán (2010) 2822–2825
32. Grzybowska, J., Kacprzak, S.: Speaker age classification and regression using i-vectors. In: Interspeech, San Francisco, USA (2016) 1402–1406
33. Gósy, M.: BEA: A multifunctional Hungarian spoken language database. The Phonetician **105**(106) (2012) 50–61
34. Lehr, M., Prud'hommeaux, E., Shafran, I., Roark, B.: Fully automated neuropsychological assessment for detecting Mild Cognitive Impairment. In: Interspeech, Portland, USA (2012) 1039–1042
35. Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support of a high-dimensional distribution. Neural Computation **13**(7) (2001) 1443–1471
36. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter **11**(1) (2009) 10–18
37. Vincze, V., Hoffmann, I., Szatlóczki, G., Bíró, E., Gosztolya, G., Tóth, L., Pákáski, M., Kálmán, J.: Az enyhe kognitív zavar automatikus azonosítása beszédátiratok alapján. In: MSZNY, Szeged (2015) 249–256

Kriminalisztikai alapú beszélőiprofil-alkotás

Beke András

Spicy Analytics Kft.
1026 Budapest, Harangvirág utca 5.
{beke.andras}@gmail.com

Kivonat: A beszélő hangja alapján az ismert személyek felismerése mellett képesek vagyunk az ismeretlen személyekről profilt készíteni, vagyis olyan általános információkat becsülni, mint például a nem, az életkor, a testalkat vagy a beszélő hangulata. Korábbi kutatások igazolták, hogy erős összefüggés van a toldalékcső hossza és a beszélő személy fizikai állapota, mint az életkor, a nem, a testmagasság stb. között. Ezen összefüggés alapján feltételezzük, hogy az emberi beszéd akusztikai jellemzői kódolják az adott beszélő testi fizikai felépítésére utaló jegyeket. A jelen kutatásban ezen összefüggés érvényességét vizsgáljuk tanuló algoritmusok segítségével. A kutatásban elemezzük, hogy a beszédből milyen eredményességgel lehet automatikusan becsülni a beszélő nemét, életkorát, testsúlyát, illetve testtömegét. A fizikai tulajdonságok becsléséhez a beszédből kinyert akusztikai jellemzőket használunk: prozódiai alapú, beszédminőség-alapú, spektrális alapú. Az eredmények azt mutatják, hogy a nem, a testtömeg és a testsúly becslése nagy pontosságú, míg az életkor becslése kevésbé.

1 Bevezetés

Az elmúlt években láthatóan megnövekedett az érdeklődés a hangalapú biometria iránt, ami ma már szintén kiegészítő eleme a személyazonosításnak, beléptető, illetve felügyeleti eszközöknek stb. Ugyanakkor fontos megjegyezni, hogy a hangalapú biometria korántsem rendelkezik olyan megbízhatósági szinttel, mint például a DNS, az ujjlenyomat vagy íriszminta. Ezzel szemben a hangalapú biometria alkalmazásának költsége jóval alacsonyabb és gyorsabb, mint a fent említetteké, így mind a gyakorlat, mind a tudományos szakma egyre nagyobb figyelmet fordít ennek fejlesztésére. A kurrens hangalapú biometriai eszközök közé tartozik a beszélőazonosítás, a beszélőhitelesítés. A beszélőazonosítás során egy *n*-elemű halmazból (általában a gyanúsítottak köréből) kell kiválasztani a legvalószínűbb beszélőt (vagyis a kérdéses mintát hozzárendelni a gyanúsítottak valamelyikéhez, vagy kizárni azt a csoportból). A beszélőhitelesítéskor azt a hipotézist kell ellenőrizni, hogy az adott beszélő az a személy-e, akinek álltja magát. A közös mindkettőben az, hogy az adott/vizsgált beszélői modell ismert személytől származik. Ugyanakkor a kriminalisztikai gyakorlatban sokszor olyan esetek is felmerülnek, amikor a beszélő személy kiléte nem ismert, illetve a hatóság nem rendelkezik gyanúsítotti körrel. A hangalapú biometriai eljárások közül a beszélőiprofil-alkotás foglalkozik ennek a kérdések a

vizsgálatával. A beszélőiprofil-alkotás a kriminalisztika egy olyan ága, amely deduktív érveléseket használ, hogy logikus következtetéseket vonjon le az elkövetőről annak beszédbeli és/vagy nyelvi ismérvei alapján. Ez a tudomány arra a megfigyelésre épül, hogy a beszélő hangja alapján az ismert személyek felismerése mellett képesek vagyunk az ismeretlen személyekről profilt készíteni, vagyis olyan általános információkat becsülni, mint a nem [14], az életkor [7,20], a testalkat [2,8,24] vagy a beszélő hangulata [22] stb. Korábbi kutatások igazolták, hogy erős összefüggés van a toldalékcső hossza és a beszélő személy fizikai állapota, mint az életkor, a nem, a testmagasság stb. között [4,5]. Ezen tényre alapulva számos olyan kutatás született, amely a beszédből automatikusan kívánta becsülni a beszélő fizikai tulajdonságait annak beszédbeli jellemzői alapján.

A nemzetközi szakirodalomban igen sok kutatás történt a beszélő nemének és életkorának automatikus becslésére annak beszéde alapján. Mindemellett van néhány olyan munka, amely a beszélő testsúlyának, illetve testmagasságának becsélét vizsgálja a beszélő hangja alapján. A hazai beszélői profilalkotás előzményeinek tekinthető: [7,8]. Ezen kutatások azonban a jelenség percepciós oldalát vizsgálták, vagyis hogy a humán hallgató milyen eredményességgel képes a beszélő fizikai paraméterinek becslésére. A következőkben bemutatjuk az egyes beszélői fizikai paraméterek becslésére történt kísérleteket.

1.1 A beszélő nemének automatikus osztályozása

A beszéd elsődleges információt hordozhat a beszélő nemére vonatkozóan. Ennek hátterében az áll, hogy a beszéd létrehozása biológiailag meghatározott. A nők és a férfiak fiziológiai jellemzőiket – többek között a beszédszervek méretét – tekintve eltérőek. A hangszalagok, valamint az artikulációs csatorna hossza kisebb a nők esetében (a hangszalagok átlagos hossza: 1,5 és 2 cm közötti, az artikulációs csatorna átlagos hossza: 14 cm; térfogata 130 cm^3), mint a férfiaké (a hangszalag átlagos hossza: 1,9 és 2,9 cm közötti, az artikulációs csatorna átlagos hossza: 17 cm, térfogata 170 cm^3). Az első és legdominánsabb akusztikai következménye ennek, amely hallás alapján is jól feldolgozható, hogy a nők hangmagassága magasabb, mint a férfiaké. Jóllehet a nemek azonosítása beszédhang alapján (akár percepciósan, akár gépi úton) egyszerű feladatnak tűnhet, mégis sok esetben téves eredményt adhat. Nehezíti a feladatot az is, hogy az életkor előrehaladásával a nők hangja mélyül, a férfiaké pedig magasabbá válik. A nemek felismerését nehezítik továbbá az alkoholfogyasztás, a dohányzás, illetve a betegségek is. A fiziológiai különbözőségeken kívül a nők és a férfiak nyelvhasználata is eltérő lehet, amely a szóhasználatban, a létrehozott szöveg nyelvtani összetettségében, a pragmatikai tényezőkben, avagy a nonverbális kommunikációban mutatkozhat meg [12]. A nők például gyakrabban használják a visszakérdezést (*ugye?*, *igaz?*, *hát nem?* stb.) a szolidaritás és az udvariasság eszközeként, mint a férfiak.

1.2 A beszélő életkorának becslése

Az életkor előrehaladtával a beszédszervek változáson mennek keresztül (például a tüdőkapacitás, a gégenagyság, a fogazat és a nyelv izomzata is változik), amelynek következtében jelentősen módosulhat a beszéd hangzása. Ez a változás igen gyakran a beszélő hangmagasságát, artikulációs és beszédtempóját, hangerejét stb. érinti. A nemhez hasonlóan, az életkor is befolyásoló tényező lehet a szóhasználat, a létrehozott szöveg nyelvtani összetettségét, pragmatikai tényezőit, avagy a nonverbális kommunikációt illetően. A beszédkutatások három fő életkori szakaszra irányulnak: gyermekkor (18 éves korig), felnőttkor (18–65 éves kor között), időskor (65 éves kortól). A kutatási eredmények szerint azonban a hallgató az életkor vonatkozásában ennél jóval pontosabb becslést is képes adni, közel 70%-os pontossággal tudja a beszélő életkorát megbecsülni [24]. Gyermekkorban a beszéd a felnőttkori beszédtől az alábbi tényezőkben tér el: az alaphangmagasság magasabb (gyermekkorban a legmagasabb: ekkor a fiúk és lányok között nincs, vagy alig van különbség), a beszéd- és artikulációs tempó lassabb, a beszédhangok kiejtése még nem kiforrott, a beszéd tematikája és grammatikai kidolgozottsága eltérő. Időskorban a beszéd a felnőtt beszédétől az alábbi tényezőkben tér el: a beszéd és artikulációs tempó lassabb, a beszédtervezési nehézségek jellemzőek, a szünettartás gyakoribb, az artikuláció pontatlanabb, az egyes szavak kiejtése hosszabb időtartamú, a hiba típusú megakadások gyakoribbak [3].

1.3 A beszélő testmagasságának becslése

Az elmúlt években szintén több vizsgálatot végeztek a testmagasság becslésére a beszéd alapján. Ezek többsége a humán percepciót tesztelte, vagyis azt, hogy a kutatásban részt vevő hallgatók milyen eredményességgel tudják megítélni a beszélő testmagasságát a hallott beszédminta alapján. Ezek eredményei azonban nem mutatnak egy irányba. Lass [13,14] munkái alapján a hallgatók egyértelműen meg tudják becsülni a beszélők testmagasságát, míg a későbbi munkák ezt nem erősítették meg [10]. Az újabb kutatások szerint [24] csak a férfiak esetében találtak erős korrelációt a testmagasság és a becsült értékek között, míg más vizsgálatban egyáltalán nem tudtak ilyen jellegű összefüggést kimutatni. Számos akusztikai vonatkozású kutatás vizsgálta a toldalékcső hossza és a testmagasság közötti korrelációt. Fitsh és Giedd [5] erős korrelációt mutatott ki a toldalékcső és a testmagasság között röntgen- és mágnesesrezonancia-alapú vizsgálatok alapján. A testmagasság előrejelzésére a számos akusztikai jellemző közül vizsgálták már a szubglottális rezonanciát [1], az alaphangmagasságot [6], a formánsokat [9,21], az MFCC-t és a lineáris predikciós együtthatókat (LPC) [4,18]. Emellett vannak olyan kutatások, amelyek mindezen jellemzőket használják a testmagasság becsléséhez. Mporas és munkatársa [16] például az openSMILE akusztikai jellemzőkinyerő algoritmust használták, amely több mint 6000 akusztikai paramétert tartalmaz. Tanulmányukban vizsgálták a lineáris és nem-lineáris regressziós modellek hatékonyságát is. Az eredményeik azt mutatták, hogy 0,053 méter volt az átlagos

eltérés a valós testmagasság és prediktált testmagasság között, ha bagging regressziós modellt használtak.

1.4 A beszélő testtömegének becslése

A testtömeg és a beszéd közötti korreláció mögött szintén a vokális traktus és az akusztikai jel közötti korreláció adja [13]. De Sousa és munkatársai [23] kimutatták, hogy a testtömegindex megjelenik az akusztikumban is. Vizsgálatukban 84 nő testtömegindexét (19 alultáplált; 23 normál, 20 túlsúlyos and 22 elhízott) és akusztikai paramétereit hasonlították össze (alaphangmagasság, kitartott magánhangzó [a] és fonációs időtartam az [a], az [i] és az [u] hangnak. A legnagyobb eltérést az alaphangmagasságban találták az egyes csoportok között. Poorjam és munkatársai [19] az i-vektoron (Gaussok középértékének összefűzése) nem-negatív mátrixfaktorizálással (Gaussok súlyain alkalmazva) olyan rendszer hoztak létre, amely a NIST telefonos beszédadatbázisán a férfiak estében 0,56-os, míg a nők esetében 0,49-es korrelációt mutat az eredeti és a becsült testtömeg között.

2 Adatbázis: leíró statisztikák a vizsgált paraméterek mentén

A jelen kutatásunkat a BEA (BEszéltnyelvi Adatbázis [17]) adatbázisból véletlenszerűen kiválasztott 134 adatközlő hanganyagán végeztük. A BEA protokolljából csak a mondatfelolvasást választottuk ki: 25 mondat minden beszélőtől, így 3350 mondat volt a teljes adatbázisunk. A kiválasztott korpuszban a nemek eloszlása nem volt egyenlő: 54 férfi, 80 női beszélőt tartalmazott. A korpusz leíró statisztikáját az 1. táblázat mutatja.

1. Táblázat: A korpusz leíró statisztikája

Faktorok	Férfi		Női	
	Átlag	Átlagos eltérés	Átlag	Átlagos eltérés
Életkor (év)	38	18	42	21
Testmagasság (cm)	178	10	165	5
Testsúly (kg)	79	15	64	13

Az életkori eloszlás esetében jól látható, hogy a résztvevők többségének életkora 20–30 év között mozog (Fig. 1.). A többi életkori cella közel azonos számosságú. A testmagasság közel normál eloszlású, egy kiugróan alacsony testmagasságú adatközlővel (Fig. 1.). A testtömeg balra ferde, de normál eloszlású (Fig.1.).

Fig. 1. Az életkor, a testmagasság és a testtömeg eloszlása a korpuszban.

Az egyes testjellemzők közötti korrelációelemzés alapján azt lehet látni, hogy nem meglepő módon a testtömeg és a testmagasság között van közepesen erős korreláció (Fig. 2.). Az életkor és a többi jellemző között viszont nincs ilyen értelemben korreláció.

Fig. 2. Korreláció az életkor, a testmagasság és a testtömeg között.

A nemen belül a nők esetében jól látható, hogy a 25 évesek felülreprezentáltak a korpuszban, míg a férfiak esetében ez kevésbé tapasztalható (Fig.3.). Ugyanakkor a nők esetében több olyan adatközlő is van, aki az idősebb korosztályba tartozik.

Fig. 3. Az életkori eloszlás a nem tekintetében.

Jelentős különbséget figyelhetünk meg a nők és a férfiak között a testmagasság és a testtömeg esetében is. A férfiak átlagosan 13 cm-el magasabbak, mint a nők (Fig.4.).

Fig. 4. A testmagasság (balra) és testtömeg (jobbra) eloszlása a nem tekintetében.

A férfiak testtömegüket illetően átlagosan 15 kg-mal súlyosabbak, mint a nők (Fig.4.).

Megvizsgáltuk, hogy a nők és férfiak esetében hogyan korrelálnak az egyes testjellemzők (Fig.5.)

Fig. 5. Korreláció az életkor, a testmagasság és a testtömeg között a férfiaknál (balra) és a nőknél (jobbra).

Látható, hogy a testtömeg és a testmagasság közötti közepes korreláció csak a férfiak esetében igazolható, a nőknél nem. Ugyanakkor a nőknél megfigyelhető egy gyenge korreláció az életkor és a testtömeg között.

Összességében megfigyelhető, hogy az egyes testi jellemzők nem kiegyenlítettek a korpuszban, illetve igen erősen függnek a nemtől, vagyis hogy férfiról vagy nőről van-e szó.

3. Módszer

3.1 Akusztikai jellemzők

Az egyes testi jellemzők osztályozásához, illetve becsléséhez olyan általános jellemzőkészletet kívántunk összeállítani, amely mind a nem, az életkor, a testmagasság és a testtömeg esetében használható. A teljes jellemzőkészletet azonban minden egyes testjellemző esetében specifikáltuk dimenziócsökkentő vagy jellemzőválogató algoritmussal. Az akusztikai jellemzőkinyerést az R programban valósítottuk meg (’seewave’, ’sound’, ’signal’,’tuneR’), valamint az ezekben nem található függvényeket (HNR, jitter, shimmer) implementáltuk R nyelvre.

3.1.1 Időtartományban mért jellemzők

Tipikusan olyan jellemzők tartoznak ide, mint alaphangmagasságból vagy a hangszínezetből (voice quality) származtatott jellemzők. A jelen kutatásban az alaphangmagasságból számított alapstatisztikai mutatókat vettük ide: átlag, szórás, minimum, maximum, ferdeség, csúcsosság. Az alaphangmagasság az egyik legtöbbet idézett akusztikai jellemző, amely összefüggésbe hozható mind a nemmel (a férfiak

f0-értéke általában alacsonyabb, mint a nőké), a testmagassággal (az alacsonyabb embereknek általában magasabb az f0-értékük, mint a magasabb embereknek [2], illetve a testtömeggel (általában az alultáplátak f0 értéke magasabb, és akinek magas a testtömegük, azok f0 értéke alacsonyabb a normál testtömegűekhez képest [23]. Az f0 mellett olyan akusztikai paramétereket is kerestünk, amelyek utalhatnak a hangszínezetre. Ezek közül kimértük az időbeli rezgésentrópia szintjét, amely a jel rendezettségét kívánja reprezentálni, a nullátmenetek számát (ZRC), az amplitudó Hilbert-burkolóját, a jittert, a shimmert, a HNR-t, a teljes görbületet (amely az időbeli rezgés második deriváltjának teljes összege), az amplitudó négyzetes középértékét (RNS), valamint a jel energiáját (annak első két deriváltját, illetve ezekből számított átlagát és szórását).

3.1.2 Frekvenciatartományban mért jellemzők

Minden egyes mondatra kiszámoltuk a 12 koefficienst tartalmazó MFCC-t (0–8 kHz között), illetve az energiát, majd ezeknek első két deriváltját, majd ebből számítottunk alapstatisztikai paramétereket (átlag, szórás, ferdeség, csúcsosság: 13x2x4). A MFCC mellett ugyanilyen módon kiszámoltuk a 12 együtthatót tartalmazó PLP-t is (0–4 kHz között). Továbbá kimértük a frekvencia modulációt is 50–8000 Hz között, amelyből szintén leíró paramétereket származtattunk (átlag, szórás, ferdeség, csúcsosság). Ezek mellett vettük a 0–4 kHz közötti spektrum alapvető statisztikai paramétereit (átlag, szórás, medián, kvartilisek, centroid, ferdeség, csúcsosság, entrópia, flatness, módusz). Összesen 344 dimenziós sorvektorral jellemeztünk minden egyes mondatot.

2.3 Modellek: klasszifikációs és regressziós modellek

A modellek tanításához és teszteléséhez a teljes adatbázist random módon két részre osztottuk, tanító és teszt adatbázisra (70-30%-ban). A tanító halmazba 1940 mondat került, a teszthalmazba pedig 1410 mondat. Az egyes tanuló algoritmusok hyper-paramétereit a tanító adatbázison állítottuk be 5-szörös keresztkiértékeléssel. A modellek tanításához az R szoftvert használtuk (’caret’, ’h2o’).

2.3.1 Klasszifikáció

A jellemzőkészlet nem-specifikus reprezentáláshoz főkomponens-elemzést (PCA) használtunk. A PCA alapján összesen 74 főkomponenst tartottunk meg, ezzel végeztük az osztályozást. A nem osztályozásához bináris logisztikus regressziót, random forest algoritmust, szupport vektor gépet, mély neurális hálózatot és gradient boosting machine-t használtunk. Ezen felül készítettünk egy, a modelleket összevonó ún. ensemble modellt is, amelyben az összevonást végző algoritmus a GLM. Az ensemble esetén kinyertük a tanuló és tesztelő adatokra kapott score-okat az öt egyedi tanuló algoritmussal, vagyis az új adatdimenzió egy 5-dimenziós sorvektor minden egyes esetre. Ezekre az adatokra tanítottunk egy új GLM-et, amely így az egyes tanuló algoritmusok előnyeit emeli ki a nem osztályozási feladatában.

2.3.2 Regresszió

Az életkor, a testtömeg és a testmagasság becsléséhez a teljes jellemzőkészletet használtuk. Az egyes testi jellemzők becsléséhez általános lineáris modellt, random forest-et, gradient boosting machine-et és mély neurális hálózatot használtunk. Ezen felül kísérletet végeztük a modellek összevonásával, úgynevezett ensemble modellel. Az első ilyen modell esetén mind a négy tanuló algoritmus kimeneti score-ját átlagoltuk. A második ensemble modell esetén a négy kimentre egy újabb általános lineáris modellt tanítottunk.

3 Eredmények

3.1 Az nem osztályozásának eredménye

Az általunk elvégzett klasszifikáció a nem tekintetében igen jó eredményt adott. Ez megegyezik a korábbi kutatások eredményeivel. Mindegyik algoritmussal közel 98-99%-os pontosságot értünk el (Táblázat 2.). Az eredmények azt mutatják, hogy a legjobb osztályozási metrikákat az SVM, a két rejtett réteget tartalmazó DNN, illetve az Ensemble modell adta.

2. Táblázat: A nem osztályozásának eredménye

	Acc	Kappa	AUC
LogisticRegression	0,953	0,902	0,967
RandomForest	0,894	0,772	0,983
SupportVectorMachine	0,966	0,928	0,996
GradientBoosting	0,939	0,872	0,984
DNN [100]-Rectifier	0,953	0,902	0,992
DNN [100-100]-TahnDropout 50%	0,961	0,919	0,993
Ensemble	0,965	0,926	0,996

3.2 Az életkor becslésének eredménye

Az eredményekből az látszik, hogy a feladat lineáris megközelítése a GLM-mel sikertelen volt, ugyanakkor a többi modell eredménye jónak mondható. A becsült életkor közel 10-12 évet téved átlagosan. Az önálló modellek közül a DNN (2 rejtett réteget 200-200 unit, Tanh-függvény) adta a legjobb eredményt. A legjobb eredményt összességében a GLM-mel összevont ensemble modell adta, ebben az esetben volt a legkisebb a hiba: 10 éven belüli (Táblázat 3.).

3. Táblázat: Az életkor becslésének eredménye

	RMSE	Rsquare	MAE
GLM	15,8	0,426	12,6
RandomForest	13,2	0,665	9,8
GBM	12,7	0,634	9,3
DNN	10,2	0,763	7,2
Ensemble_Avg	12,3	0,720	9,3
Ensemble_GLM	**9,8**	**0,770**	**6,8**

A regresszió szerint a nők esetén illeszkednek jobban a valós értékekre a prediktált értékek (0,782), az illeszkedés férfiak estén kisebb (0,768). Ugyanakkor a várt értéktől való eltérés az RMSE és a MAE alapján a férfiak estében kisebb (RMSE: 8,96; MAE: 6,3), mint a nők esetében (RMSE: 10,3; MAE: 7,1) (Fig. 6.).

Fig. 6. A valós életkor és a becsült életkor közötti összefüggés a férfiak és nők esetében

3.4 A testmagasság becslésének becslése

A testmagasság becslése az eredmények szerint jobban megközelíthető az általunk készített módszerrel. A GLM-et leszámítva 0,8 feletti korrelációs értéket kaptunk a valós testmagasság, illetve a becsült testmagasság között, emellett az áltagos hiba 3–7 cm között mozgott. A legjobb eredményt itt a DNN-nel, illetve a GLM-mel egyesített modellel kaptuk (Táblázat 4.).

A regresszió szerint a férfiak esetén jobban illeszkednek a prediktált értékek a valós értékekre (0,788), mint a nők estén (0,750). Ugyanakkor a várt értéktől való eltérés az RMSE és a MAE alapján a nők estében kisebb (RMSE: 2,5; MAE: 2,0), mint a férfiak esetében (RMSE: 9,2; MAE: 3) (Fig. 7.).

4. Táblázat: A testmagasság becslésének eredménye

	RMSE	Rsquare	MAE
GLM	6,6	0,500	5,2
RandomForest	4,2	0,814	3,1
GBM	4,3	0,785	3,2
DNN	3,8	0,824	2,9
Ensemble_Avg	4,2	0,812	3,2
Ensemble_GLM	**3,2**	**0,873**	**2,4**

Fig. 7. A valós testmagasság és a becsült életkor közötti összefüggés a férfiak és nők esetében

3.5 A testtömeg becslésének eredménye

A testtömeg becslése szintén nehezebb feladatnak bizonyult. Az eredmények szerint a becsült testmagasság közel 4–6 kg közötti átlagos hibával becsülhető. Ugyanúgy, ahogy az előző regressziós feladatokban, a legjobb eredményt ebben az esetben is a DNN, illetve a GLM-mel egyesített modellel adta (Táblázat 5.).

A regresszió szerint a nők esetén jobban illeszkednek a valós értékekre a prediktált értékek (0,783), mint a férfiak estén (0,754). A becslés pontosság a két esetben közel ugyanolyan mértékű volt (nők: RMSE: 6,7; MAE: 4,8; férfiak: RMSE: 6,5; MAE: 4,8) (Fig. 8.).

4 Következtetések

A jelen kutatásban azt vizsgáltuk, hogy milyen eredményességgel lehet a beszédből becslést adni a beszélő fizikai adottságaira a nem, az életkor a testmagasság, illetve a testtömeg tekintetében. Az eredményeink szerint az olvasott mondatokon a nem

meghatározása igen jó minőségben megvalósítható. Az életkor becslése esetében a feladat jóval komplexebb, hiszen a beszédben észlelhető életkor, illetve a valós életkor között (vö. [15]) sok esetben ugrás észlelhető. Ez a különbség tovább nő az idő előrehaladtával, vagyis a becslés pontossága a fiatalok esetében jobb, mint az idősek esetében (vö. [11]) (lásd Fig. 6). A testmagasság esetében a modell pontossága igen jónak mondható, vagyis az általunk kinyert akusztikai paraméterek hordozzák azon jegyeket, amelyek alkalmasak ezen fizikai paraméter becslésére. Ugyanez elmondható a testtömeg esetén is, jóllehet ennek automatikus becslése valamivel pontatlanabbnak bizonyult.

5. Táblázat: A testtömeg becslésének eredménye

	RMSE	Rsquare	MAE
GLM	12,9	0,359	9,9
RandomForest	8,9	0,706	6,6
GBM	8,6	0,709	6,4
DNN	7,9	0,754	6,0
Ensemble_Avg	8,9	0,750	6,8
Ensemble_GLM	**6,6**	**0,817**	**4,8**

Fig. 8. A valós testtömeg és a becsült életkor közötti összefüggés a férfiak és nők esetében

Ebben a kutatásban egy alap, ún. baseline rendszert hoztunk létre, amely felolvasott mondatokból kinyert akusztikai jellemzők alapján tesz becslést automatikusan a beszélő fizikai paramétereire. Ezen baseline rendszert kívánjuk a jövőben tovább fejleszteni, kiterjeszteni más beszédtípusra, más korpuszra más/komplexebb modellek bevonásával, továbbá az akusztikai jellemzők mellett szöveges jellemzők felhasználásával.

Köszönetnyilvánítás

A jelen kutatás a Bolyai János Kutatási Ösztöndíj támogatásával készült.

Bibliográfia

1. Arsikere, H., Leung, G., Lulich, S., and Alwan, A. (2012). “Automatic height estimation using the second subglottal resonance,” in IEEE International Conference on Acoustics, Speech and Signal Processing 2012 (ICASSP) (2012) 3989–3992.
2. Barsties, B., Verfaillie, R., Dicks, P., & Maryn, Y. „Is the speaking fundamental frequency in females related to body height?”. Logopedics Phoniatrics Vocology, 41(1), (2016) 27–32.
3. Bóna, J. "A spontán beszéd sajátosságai az időskorban." ELTE Eötvös Kiadó, Bp. (2013).
4. Dusan, S. “Estimation of speakers height and vocal tract length from speech signal,” in INTERSPEECH (ISCA) (2005) 1989–1992.
5. Fitch, W. T., and Giedd, J., “Morphology and development of the human vocal tract: A study using magnetic resonance imaging”, J. Acoust. Soc. Amer. (1999)106(3): 1511–1522.
6. Ganchev, T., Mporas, I., and Fakotakis, N. “Audio features selection for automatic height estimation from speech,” in Lecture Notes in Computer Science. Artificial Intelligence: Theories, Models and Applications, edited by S. Konstantopoulos, S. Perantonis, V. Karkaletsis, C. Spyropoulos, and G. Vouros (Springer, Berlin) (2010) 6040, 81–90.
7. Gocsál Á. Életkorbecslés a beszélő hangja alapján. Beszédkutatás (1998), 122–134.
8. Gósy M. A testalkat és az életkor becslése a beszéd alapján. Magyar Nyelvőr; (2001) 25/4: 478–487.
9. Greisbach, R. (1999). “Estimation of speaker height from formant frequencies,” Forensic Ling. 6, (1999) 265–277.
10. Gunter, C. D., & Manning, W. H. (1982). Listener estimations of speaker height and weight in unfiltered and filtered conditions. Journal of Phonetics.
11. Huckvale, M. and Aimee W. "A Comparison of Human and Machine Estimation of Speaker Age." International Conference on Statistical Language and Speech Processing. Springer, Cham, 2015.
12. Huszár Á. Bevezetés a gendernyelvészetbe: Miben különbözik és miben egyezik a férfiak és a nők nyelvhasználata és kommunikációja? Budapest: Tinta Könyvkiadó (2009) 123.
13. Lass N. J. and Davis M. “An investigation on speaker height and weight identification,” Journal of the Acoustical Society of America, vol. 60 (1976) 700–703.
14. Lass N. J, Hughes KR, Bowyer MD, Waters LT, Bourne VT: Speaker sex identification from voiced, whispered, and filtered isolated vowels. The Journal of the Acoustical Society of America (1976) 59:675–678.
15. Moyse, E. "Age estimation from faces and voices: a review." Psychologica Belgica 54.3 (2014).
16. Mporas, I., and Todor G.. "Estimation of unknown speaker’s height from speech." International Journal of Speech Technology 12.4 (2009) 149–160.
17. Neuberger, T., Gyarmathy, D., Gráczi , T. E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative hungarian language. In: TSD2014x. (2014) 424–431
18. Pellom, B. L., and Hansen, J. H. L. “Voice analysis in adverse conditions: The Centennial Olympic Park Bombing 911 call,” in Proceedings of the 40th Midwest Symposium on Circuits and Systems 2. (1997), 873–876.

19. Poorjam, A. H., Mohamad H. B., and Hugo Van H. "Speaker weight estimation from speech signals using a fusion of the i-vector and NFA frameworks." Artificial Intelligence and Signal Processing (AISP), 2015 International Symposium on. IEEE (2015).
20. Ptacek P. H, Sander E. K: Age recognition from voice. Journal of Speech and Hearing Research (1966) 9/2:273–277.
21. Rendall, D., Kollias, S., Ney, C., and Lloyd, P. “Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: The role of vocalizer body size and voice-acoustic allometry,” J. Acoust. Soc. Am. 117, (2005) 944–955.
22. Scherer K. R, Banse R, Wallbott H: Emotional inferences from vocal expression correlate across languages and cultures. Journal of Cross-Cultural Psychology (2001) 32/1:76–92.
23. Souza de, Lourdes Bernadete R., and Marquiony M. dos S. "Body mass index and acoustic voice parameters: is there a relationship." Brazilian Journal of Otorhinolaryngology (2017).
24. Van Dommelen, W. A., Moxness, B. H: Acoustic parameters in speaker height and weight identification: sex-specific behaviour. Language and Speech (1995) 38/3:267–287.

V. Többnyelvűség

Etudes in Chinese-Hungarian Corpus-Based Lexical Acquisition

Gábor Ugray

zydeodict@gmail.com

Abstract: The paper reports on a series of experiments to extract matching lexical items from a 6.1 million segment corpus of movie subtitles in Mandarin Chinese and Hungarian, with the aim of expanding an existing bilingual dictionary. The challenges of data cleansing and tokenization are outlined, and the outcome of word alignment, vector space embeddings, neural machine translation and two standard statistical approaches is presented. A bilingual concordance tool for end users, based on word alignments, is introduced. A quantitative and qualitative evaluation of the results finds that the new methods drastically outperform simple collocation extraction, but also shows that human judgement is indispensable before including vocabulary in a published dictionary.

1 Introduction

The last few years have brought two developments with promising consequences for digital lexicography. The first is the emergence of large bilingual corpora, even for an uncommon language pair such as Chinese-Hungarian. The second is neural-network-based machine learning driven by affordable GPUs. In this paper I report on a series of experiments to harness these developments for the expansion of CHDICT[1] [16], an open-source Chinese-Hungarian dictionary initially conceived as a translation of CC-CEDICT[2].

My work builds on OpenSubtitles2016 [9], a corpus of movie subtitles with 6.1 million Chinese-Hungarian segment pairs. I investigate the usefulness and limitations of word alignment, vector space models (VSM), neural machine translation (NMT), and statistical collocation extraction, to acquire lexical information from the corpus. All of these approaches have proven to be valuable sources of lexicographical insight, with VSM and NMT grossly outperforming simpler statistical methods. Furthermore, word alignment enables a bilingual concordance tool that is itself valuable for a broad audience.

[1] https://chdict.zydeo.net/en/

[2] https://www.mdbg.net/chinese/dictionary; https://cc-cedict.org/wiki/

2 Pre-processing

2.1 Data hygiene

The corpus that this work builds on contains 6.1 million aligned movie subtitles. I subsequently refer to the corpus's units as segments. These are usually, but not always, full sentences, which the corpus's authors aligned chiefly on the basis of timestamps. That method inevitably introduces some noise into the data through misaligned pairs, which is complemented by dirty data in the form of encoding errors, content in the wrong language, and for Chinese, in the wrong script (traditional instead of simplified).

I applied a mix of strategies to fix and prune the data. This included (a) converting to simplified if the segment included traditional-only characters, using OpenCC[3]; (b) fixing Hungarian *õ* and *û*; (c) discarding pairs where the text contained *{[]}@*, indicating escape sequences; (d) discarding pairs where the ratio of source and target length was beyond a threshold, allowing for greater variance in shorter segments; (e) discarding segments where the proportion of punctuation characters exceeded a threshold; (f) discarding pairs where too many Hungarian words were left unanalyzed by the emMorph morphological analyzer [11]; (g) discarding pairs where the Chinese contains characters that are neither in the Latin nor the Chinese script; and (h) removing duplicates.

After this preparation, the remaining corpus used throughout the exercises contains 2.9 million segment pairs.

2.2 Chinese word segmentation[4]

A key challenge for any Chinese NLP task is the lack of word delimiters in written Mandarin. There is no single universally accepted word segmentation method, and as we will see, the optimal approach depends on the task at hand – including, even, treating each character as a separate token.

I am aware of two available segmenters, ICTCLAS[5] [18] and Jieba.[6] The first was used by Brysbaert et al. to obtain their corpus-based word frequencies for SUBTLEX-CH [2]. Unfortunately I was unable to compile and execute this tool, but in some places I rely on its output indirectly through the published SUBTLEX-CH word frequencies. Both ICTCLAS and Jieba are hybrid tools combining a dictionary and Hidden Markov Models.

Additionally, I experimented with an algorithm inspired by Gensim's [12] *Phraser*. Starting with individual characters, it iteratively merges adjoining units that co-occur more frequently than predicted by chance. My purpose was to prevent a perceived

[3] https://github.com/BYVoid/OpenCC by BYVoid
[4] Word segmentation is not to be confused with segments, the corpus's sentence-level units.
[5] https://github.com/NLPIR-team/NLPIR/tree/master/NLPIR%20SDK/NLPIR-ICTCLAS
[6] https://github.com/fxsjy/jieba by Sun Junyi (2013)

over-eagerness of Jieba in joining measure words to determiners; complements to verbs; and compound nouns. Somewhat predictably, this home-grown approach underperformed Jieba, and I abandoned the experiment. I refer to this method as "Exp-merge" later in the text.

Since the goal is to acquire lexical units that humans expect to find in a dictionary, an established large dictionary's headword list is a good benchmark. For a sense of how a segmenter's "idea" of words corresponds to the dictionary's judgement, see *Figure 1* below.

To obtain this chart, I first created three word lists, ranked by frequency as measured on a corpus segmented by the three tools. The SUBTLEX-CH frequencies are those published by Brysbaert et al. from a 33 million word corpus. The others are my own calculations on the pruned bilingual corpus. The figure is a histogram, showing values for 100 points on the X axis, each representing a 1000-word bucket in the ranked list of the 100,000 most frequent words. The Y values indicate the dictionary coverage of each bucket: how many of those 1000 words are found in CC-CEDICT.

Figure 1: Lexical coverage of CC-CEDICT's 107k headwords, depending on the choice of tokenizer.

For comparison, Exp-merge produced 66k distinct words; Jieba's output contains 238k distinct words; SUBTLEX-CH's list has 100k words. Conversely, of CC-CEDICT's 107k simplified headwords, only 27.2k, 54.6k and 42.6k are attested in the Exp-Merge, Jieba and SUBTLEX-CH frequency lists, respectively. The diagram and these figures show that all segmenters disagree greatly both among themselves and with CC-CEDICT about the "definition" of words in Mandarin Chinese.

2.3 Hungarian stemming and tokenization

The case for stemming Hungarian to avoid the data scarcity problem is evident. I used the HFST-based [8] emMorph [11] analyzer and processed its output with my own C#

port of the *Stemmer* class and its dependencies from the GATE wrapper[7] [13]. I used a naïve disambiguation rule, always opting for the shortest stem. If the shortest stem was a particle+verb or a compound, I subsequently treated those as separate tokens.

Unfortunately HFST proved to be prohibitively slow in practice; I had to analyze the corpus's 772k surface forms from the command line and use an in-memory dictionary. Stemming produced 63.8k distinct stems, of which 20,237 occur at least 3 times. 162k surface forms, of 20.9% of the total, were left unanalyzed. These are overwhelmingly typos or non-Hungarian proper names.

For certain tasks I used byte-pair encoding (BPE) [14] as an alternative tokenization method. BPE gained popularity in recent years in NMT systems because it addresses the closed-vocabulary problem, albeit at the cost of being an arbitrary and not linguistically motivated approach.

3 Word-aligned bilingual concordance

Word alignment of the training corpus has been a staple of statistical MT from the outset [1], and it gained relevance again as an aid to the attention mechanism in NMT [3]. The alignment method itself has been improved significantly as recently as 2013 by the authors of *fast_align* [5]. The approach has great appeal because it can combine corpus-wide co-occurrence probabilities with local sequence information in individual segment pairs.

The temptation to build a Linguee-like[8] tool for searching the cleansed bilingual corpus, word-aligned on a segment level, was irresistible. My initial aim was to create a research tool for dictionary authoring, but as we will see in *Results*, the outcome has immediate value for end users too.

3.1 Training

I executed *fast_align* after tokenizing both Chinese and Hungarian in several different ways. *Table 1* shows *fast_align*'s reported perplexity values for these combinations.

For Hungarian tokenization, *bpe20k*, *bpe30k* and *bpe40* refer to BPE with 20k, 30k and 40k merges, respectively. *surf-lo* stands for no stemming, only lower-case normalization. *stem-lo* stands for lower-cased stems. For Chinese, Jieba outperforms the experimental word segmenter. Interestingly, increasing BPE's output vocabulary leads to worse outcomes.

BPE with Jieba allows for a few insightful searches, where a Chinese preposition or verbal complement is (correctly) mapped to a Hungarian suffix. However, *stem-lo+Jieba* grossly outperforms all other combinations, and was chosen for the final tool.

---

[7] https://github.com/dlt-rilmta/hunlp-GATE/tree/master/Lang_Hungarian/resources/hfst/hfst-wrapper

[8] https://www.linguee.de/

Chinese word segmenter	Hungarian tokenizer	Perplexity
Jieba	bpe20k	284.48
Jieba	bpe30k	295.85
Jieba	bpe40k	300.16
Jieba	surf-lo	216.47
Jieba	stem-lo	86.97
Exp-merge	bpe20k	355.75
Exp-merge	stem-lo	112.44

Table 1: *fast_align*'s reported perplexity values after training for 5 iterations, depending on the choice of tokenizer.

3.2 Presentation

I integrated a custom-developed tool for searching the word-aligned bilingual corpus within the CHDICT website[9]. For an illustration, see *Figure 2*, with a few results for 汽车 qìchē. The tool allows searching for either Chinese or Hungarian text, and presents matching segment pairs from the corpus.

Figure 2: A few sample search results from the word-aligned bilingual concordance tool's output.

In the results, the search term and its matches in the opposite language are highlighted in each segment. I used a slightly modified version of *fast_align* that outputs confidence values, which are indicated by the strength of the highlight. If the Chinese search term happens to be a substring of a Jieba token, the full token is also shown with a lighter highlight to clarify what the alignment truly means.

For Hungarian searches the tool uses two separate indexes. One matches the query's exact surface form; the other matches stems. Because of HFST's performance issues, queries are "stemmed" through an auxiliary table mapping the corpus's 772k surface forms to their chosen stems. The tool uses Sphinx[10] for quick and memory-efficient indexing and retrieval.

[9] https://chdict.zydeo.net/en/corpus
[10] http://sphinxsearch.com/

4 Bilingual word embeddings

Vector Space Models (VSMs) [15] embody the idea that a word's paradigmatic and semantic properties can be captured by quantifying what other words they tend to co-occur with. TF-IDF has been widely used in document retrieval since the 1970s paper that Gerard Salton never wrote [4]. More recently, neural networks have been used to learn word embeddings [10], replacing the closed formulas based on term counts.

I attempted to extract translations by embedding words from two languages in a single vector space. The approach is similar to Vulić et al. [17], but while that work relies on sampling non-sentence-aligned document pairs, my corpus allowed creating bags of words directly from Chinese+Hungarian segment pairs.

4.1 Extraction

The standard way to build a term-context matrix is to observe a small window up to about a dozen words in monolingual text. For my experiment I created, instead, a single bag of words from each Chinese and Hungarian segment pair. For clarity I prefixed Hungarian tokens with *hu_* and Chinese ones with *zh_*, although the languages can easily be distinguished by script.

To create the word embeddings, I used Gensim's Word2Vec model in skip-gram mode, with a window beyond the largest combined segment length. Gensim is an efficient re-implementation of Mikolov et al.'s neural word2vec model [10]. I used 200 dimensions, a value lower than the 300-500 that is standard in neural MT systems.

To define what a "word" is, I used Jieba for Chinese and lower-cased word stems for Hungarian. In this task, I discarded BPE because it is not linguistically informed.

Once the word vectors were learned, I applied a brute-force quadratic search to find the 40 nearest (by cosine similarity) Hungarian words with a frequency of 3 or higher, for each of the 54.6k CC-CEDICT headwords that are attested in the Jieba-segmented data.

4.2 Filtering

The raw output was, predictably, extremely noisy. Scores of 0.8 or higher are very reliable, but only 2,673 Chinese words have such a close Hungarian neighbor. On the other hand, spurious Hungarian matches tend to recur often in the top 40 list of several Chinese words. These proliferous matches are invariably noise: *nore* tops the list, showing up with 7,060 Chinese headwords, followed by *tada*, *csatlakozatok*, *lndítsuk* and *áilítólag*.

This enabled a filtering approach that also keeps potentially useful matches with a lower score. After ignoring Hungarian words that occur in the top 40 list of at least 100 different Chinese words, I was left with shorter non-empty lists for 34k CC-CEDICT headwords.

4.3 Outcome

Vector similarity delivers on its promise, returning a collage of words related in various ways. The list is always mixed: apart from the remaining noise, it contains semantic equivalents; complementary parts of frequent collocations; or simply vaguely related concepts. For illustration, here is the list of 遗物 yíwù, which CC-CEDICT glosses as *remnant:*

0.58 holmi • 0.53 ereklye • 0.50 mamaji • 0.50 hamvaszt • 0.49 felipe • 0.49 gyűjtemény • 0.48 régiség • 0.48 taiáitam • 0.48 drágakő • 0.48 hagyaték • 0.47 davenport • 0.47 yukio • 0.46 szuvenír • 0.46 ékszer • 0.46 mohammad • 0.46 amun-ra • 0.45 josemaría • 0.45 itthagyott • 0.44 coggins • 0.44 anyakönyvi • 0.44 bizsu • 0.43 hamu • 0.43 irat • 0.43 régész • 0.43 tárgy

This impressionistic collage helps disambiguate *remnant* into the eventual Hungarian glosses: *maradvány; ereklye; tárgyi emlék; hagyaték*. The vector space is a weak and noisy source of "translations" as such, but it has proven very valuable as a lexicographical tool to chart a headword's associations, connotations and register.

5 Neural MT

Google MT was already one of several sources for the compilation of CHDICT's original 11k entries [16]. Direct Chinese-Hungarian translations were rarely useful, with strong hints that Google uses English as a pivot language. I now investigated if custom NMT models trained from a Chinese-Hungarian corpus would yield useful headword translations.

I used OpenNMT [7] to train several models on data tokenized in different ways. All models have word embeddings of 500 dimensions and a 500-node 2-layer RNN. They were trained for 13 epochs with SGD, an initial learning rate of 1, annealed at a factor of 0.7 starting at epoch 9. Each model took approximately 12 hours to train on an NVIDIA GTX 1080 GPU, with mini-batches of 256. *Table 2* shows the perplexity and BLEU score results.

Seg-ZH	Tok-HU	Perplexity	BLEU
chars	**bpe10k**	**14.10**	**9.58**
chars	bpe20k	16.27	8.85
chars	bpe40k	18.35	8.80
chars	**stem-lo**	**22.46**	**10.41**
exp-merge	bpe20k	16.13	9.99
Jieba	**bpe20k**	**16.29**	**10.03**

Table 2: Final perplexity values and BLUE scores reported by OpenNMT, depending on the choice of segmenter/tokenizer.

If the goal were machine translation of full sentences, these results would be underwhelming. My aim, however, was only to extract helpful Hungarian hints for individual Chinese words. To this end I translated CC-CEDICT's 107k simplified headwords with the three models highlighted above, using a beam size of 10 and keeping the 20 best results.

5.1 Outcome

The output frequently shows anomalies that are well known to NMT practitioners. One example is the "I don't know problem"[11] also observed in neural chatbots, where the system defaults to a generally likely target segment. From the film subtitles corpus, this produces output like *igen; mi; igen uram; nem; szia;* etc. The other salient anomaly occurs with BPE-segmented Hungarian, where the system gets stuck in repetition loops: *ho hoho; hohohoho; hohohohoho;* etc.

Because the prediction score produced by inference is not a good indicator of quality, I used the same filtering approach as with vector similarities, discarding target strings that recur for many inputs. In fact I applied a stricter filter and discarded all of an engine's results for a given input if the first (best-scored) translation was on the proliferous noise list.

For illustration, this is the filtered output of the 3 selected engines for the previously mentioned headword, 遗物 yíwù:

MT char-char
örökre • marad • maradvány • maradt • maradj • maradsz • örökség • maradványokat • hagyja • holmija • egy maradvány • maradványok • öröksége
MT char-stem
tárgy • örökség • ereklye • hagyaték • maradvány • egy ereklye • kincs • rom • zsákmány • holmi • egy tárgy • egy vagyon • búcsú • tulajdon • sajnál • vagyon
MT jie-char
tárgy • cucc • tárgyról • holmi • tárgya • egy tárgy • tárgyak • dolgokat • tárgyakat • dolgok • tárgyat • holmik • dolgot • ez a tárgy

This is a drastic improvement over Google MT, which translates 遗物 as *emlékei*. In the cases where a Chinese headword has direct equivalents in Hungarian, the most frequent ones usually show up among the translations of multiple engines.

A major advantage over the other approaches is that MT is occasionally capable of producing translations consisting of multiple tokens, such as compound words or short expressions.

[11] forum.opennmt.net/t/english-chatbot-advice/32/5

6 Collocation classics

Collocation extraction using established statistical formulas has little novelty value, but I also included it in my effort as a source of easily obtainable candidates. I gleaned Chinese-Hungarian token pairs with two scoring methods, log-likelihood and mutual information. In this case I used Jieba and lower-cased stems for tokenization, these being the only linguistically motivated ones.

Just as before, I relied on a frequency threshold and excessive proliferation for filtering, instead of an arbitrary score threshold. After filtering, LL and MI produced non-empty candidate lists for 16.7k and 31.8k CC-CEDICT headwords, respectively.

7 Results

Two factors make a quantitative evaluation of the investigated methods complicated. First, the very aim of these exercises is to aid in the expansion of a pioneering bilingual dictionary, which means that there is no *a priori* ground truth available.

To work around this fact, I selected a batch of entries from CC-CEDICT, picking headwords that had filtered candidates from VSM; from at least 2 MT engines; and at least 1 collocation method. There were 9k headwords matching these criteria, of which I randomly sampled 400. I then proceeded to manually compile their Chinese-Hungarian entries, consulting CC-CEDICT's English glosses as well as the candidates from the new corpus-based extraction methods, the concordance tool and other sources. This created a *post-hoc* ground truth to benchmark against.

The second complication results from the fact that a dictionary entry, even in CHDICT's simplistic format, is not a flat list of target-language equivalents. Entries are structured into senses, which in turn may contain multiple alternatives, plus meta-information in parentheses. *Figure 3* illustrates this. Evaluating flat candidate lists against a structured gold standard is not straightforward.

Figure 3: A sample entry from CHDICT, as it appears to end users in the live website.

7.1 Quantitative evaluation

Because of the complications outlined above, I had to resort to custom definitions in order to measure recall and precision. This makes the figures somewhat difficult to

compare to other efforts. They are, however, quite useful for comparing the different approaches within this paper's context. *Table 3* presents the results.

Method	**R-XS1**	**R-XSX**	**R-ASX**	**P-1SX**
Bilingual VSM	31%	40%	21%	31%
MT-char-char	28%	50%	28%	28%
MT-char-stem	34%	55%	35%	34%
MT-jie-char	32%	58%	31%	32%
Colloc/log-likelihood	11%	12%	4%	11%
Colloc/mutual-information	14%	25%	10%	14%

Table 3: Recall and precision of the investigated vocabulary extraction methods. Recall measures: R-XS1: At least one alternative in the manually compiled entry is at the top of the candidate list. R-XSX: At least one alternative appears somewhere on the list. R-ASX: All alternatives appear somewhere on the list. Precision measure: The top candidate appears among the alternatives in the manually compiled entry.

The table's numbers are based on 390 manually prepared CHDICT entries; 10 Chinese tokens were rejected as dictionary headwords altogether. The retained entries contain a total of 576 senses, 784 alternatives, and 156 parenthesized remarks or labels.

VSM grossly outperforms the two conventional collocation extraction methods. The three NMT engines appear to have different strengths and weaknesses, depending on the metric, but they significantly outperform even VSM as a source of actual translations.

7.2 Benefits and limitations in the lexicographical process

Figure 4 shows how candidates are presented in the lexicographical workbench during the compilation of entries. All the Hungarian words from the lists on the right are also added to an auto-complete dictionary to speed up typing.

I did not include word alignments among the hints; instead, the concordance tool itself can be invoked with a shortcut.

The workbench logs the time spent compiling each headword. It is beyond this paper's scope to analyze these logs, but it appears the enriched information does not affect the speed of lexicographical work. It contributes greatly, instead, to the confidence and breadth of the Hungarian glosses produced.

Figure 4: Information shown in the lexicographical workbench for the translation of a single Chinese headword.

Often, but not always, the candidates from the different methods include the words eventually selected for the Hungarian glosses, or otherwise help explore the Chinese headword's uses and meanings. A human can normally identify the remaining noise on the lists, and the relevant items condense what would otherwise be the result of hours of corpus discovery and hunting for attestations.

They candidates are not, however, reliable enough to be included in the dictionary without human judgement. Using only candidates with a very high score would leave an unacceptably small number of reliable matches, and miss important but less frequent senses. Lowering the threshold, in turn, would result in excessive noise or a proliferation of candidates. Finally, unsupervised methods obviously fail when a Chinese lexical item can only be paraphrased and when the target equivalent needs disambiguating remarks.

7.3 Augmented dictionary

The word-aligned bilingual search tool has proven to be the most versatile approach. Its value in the lexicographical process is evident, as it allows researching real-life contexts in which a headword has been attested.

But why should such research be limited to lexicographical work? The dictionary's end users benefit equally from a chance to browse headwords in context, discovering autonomously a word's translations along with typical collocations and associations.

When integrated in a dictionary, the search tool is a substitute for example phrases, which are particularly labor-intensive to compile.

Additionally, as a fallback when a word is not found in the dictionary, the tool enables end users to discover its meaning from the translated sentence pairs. The coverage of dictionaries is limited by the person-years needed to compile them,

especially for rare combinations like Chinese-Hungarian. A large bilingual corpus inevitably encodes more knowledge than is humanly possible to compile.

References

1. Brown, Peter F., Cocke, John, Della Pietra, Stephen A., Della Pietra, Vincent J., Jelinek, Fredrick, Lafferty, John D., Mercer, Robert L., and Roossin, Paul S.: A Statistical Approach To Machine Translation. In Computational Linguistics, Volume 16 Issue 2, June 1990, pp. 79-85.
2. Cai, Q., Brysbaert, M.: SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. PLoS ONE 5(6): e10729. doi:10.1371/journal.pone.0010729 (2010)
3. Crego, Josep et al.: SYSTRAN's Pure Neural Machine Translation Systems. arXiv:1610.05540
4. Dubin, David: The Most Influential Paper Gerard Salton Never Wrote. In: Library Trends 52(4) Spring 2004: 748-764.
5. Dyer, C., Chahuneau, V., and Smith, N. A.: A simple, fast, and effective reparameterization of IBM model 2. In Proceedings of the 2013 Conference of the North American Chapter of the ACM: Human Language Technologies, pp. 644-648, Atlanta, GA, USA (2013)
6. Halpern, J. and Kerman, J.: The Pitfalls and Complexities of Chinese to Chinese Conversion. In: *Proceedings of the 14th International Unicode Conference*, Cambridge, MA, March 1999
7. Klein, Guillaume; Kim, Yoon; Deng, Yuntian; Crego, Josep; Senellart, Jean; Rush, Alexander M.: OpenNMT: Open-source Toolkit for Neural Machine Translation. In Proceedings of ACL 2017
8. Lindén, K., Silfverberg, M., Pirinen, T.: HFST tools for morphology – an efficient open-source package for construction of morphological analyzers. In: Mahlow, C., Piotrowski, M. (eds.) State of the Art in Computational Morphology, Communications in Computer and Information Science, vol. 41, pp. 28–47. Springer Berlin. Heidelberg (2009)
9. Lison, Pierre and Tiedemann, Jörg: OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013
11. Attila Novák; Borbála Siklósi; Csaba Oravecz: A New Integrated Open-source Morphological Analyzer for Hungarian. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, pp. 1315–1322.
12. Řehůřek, Radim and Sojka, Petr: Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. Valletta, Malta: University of Malta, 2010. pp. 46-50.
13. Sass Bálint, Miháltz Márton, Kundráth Péter: Az e-magyar rendszer GATE környezetbe integrált magyar szövegfeldolgozó eszközlánca. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017) (2017)
14. Sennrich, Rico, Haddow, Barry and Birch, Alexandra: Neural Machine Translation of Rare Words with Subword Units Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany.
15. Turney, Peter D., Patrick, Pantel: From Frequency to Meaning: Vector Space Models of Semantics. In: Journal of Artificial Intelligence Research 37 (2010) pp. 141-188.

16. Ugray Gábor: Egy vakmerő digitális lexikográfiai kísérlet: a CHDICT nyílt kínai-magyar szótár. Poster presentation. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017) (2017)
17. Vulić, Ivan, Moens, Marie-Francine: Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In: Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP (Vol. 2), July 2015, pp. 719-715.
18. Zhang Hua-Ping, Yu Hong-Kui, Xiong De-Yi Xiong and Liu Qun: HHMM-based Chinese Lexical Analyzer ICTCLAS, proceedings of 2nd SigHan Workshop, July 2003, pp.184-187.

Lexikai erőforrások automatikus előállítása kisebbségi finnugor nyelvekre

Simon Eszter, Mittelholcz Iván, Ferenczi Zsanett

MTA Nyelvtudományi Intézet
{VEZETÉKNÉV.KERESZTNÉV}@nytud.mta.hu

Kivonat A cikkben bemutatott projekt célja, hogy kisebbségi finnugor nyelvek számára nyelvi erőforrásokat állítson elő, melyek segítik ezeket a veszélyeztetett nyelvi közösségeket a revitalizálási folyamatokban. A projekt során kétnyelvű protoszótárakat állítottunk elő, melyeket anyanyelvi beszélők és nyelvész szakértők ellenőriztek. Az ellenőrzött fordítási párok különféle nyelvi információkkal kiegészítve kerülnek feltöltésre a Wiktionarybe. A cikk bemutatja a kétnyelvű szótárak automatikus előállításának és kézi validálásának menetét, valamint azt a munkafolyamatot, amivel a Wiktionary-szócikkeket állítottuk elő teljesen automatikusan.
Kulcsszavak: kétnyelvű szótárak, lexikai erőforrások, finnugor nyelvek, kisebbségi nyelvek, számítógépes nyelvészet, Wiktionary

1. Bevezetés

Cikkünkben egy olyan projektet mutatunk be, amelynek célja, hogy segítse a veszélyeztetett finnugor nyelvű közösségeket a digitális revitalizációban online tartalmak létrehozásával. A projekt során kétnyelvű szótárakat állítottunk elő, melyekkel kisebbségi finnugor nyelvek revitalizációját próbáljuk támogatni úgy, hogy az addig csak szegényes digitális tartalommal rendelkező nyelvek számára újabb szópárokkal gazdagítjuk az interneten fellelhető fordítási párok számát. A feltöltendő szópárokat nyelvi információkkal bővítjük ki, és a kész szócikkeket elérhetővé tesszük a Wiktionary keretein belül.

A szótári elemek a Wiktionary különböző nyelvű változataiban összekapcsolhatók, az interwiki linkek pedig a Wikipédia felé biztosítják az átjárást. Ez lehetővé teszi, hogy a nyelvközösségek gazdag lexikai anyaghoz férjenek hozzá. Emellett olyan új adatok is elérhetők lesznek a lexikai elemekhez, mint a szófaji információ vagy a fordítási megfelelők. Ezzel támogatni kívánjuk a veszélyeztetett finnugor nyelvű közösségeket a digitális revitalizációban, amivel – reményeink szerint – hozzájárulunk a nyelvi sokszínűség fenntartásához. Szabadon elérhető online többnyelvű lexikai erőforrás a sok beszélővel rendelkező nyelvekre is kevés van – kivétel ez alól a BabelNet [1] és a szabadon elérhetővé tett többnyelvű wordnetek, mint például a MultiWordNet [2] –, vagyis a megbízható lexikai erőforrások előállítása minden nyelvre kiemelten hasznos.

A projekt során hat kisebbségi finnugor nyelvvel dolgozunk forrásnyelvként: komi-permják, komi-zürjén, udmurt, mezei mari, hegyi mari és északi számi. A

fordítások célnyelve négy olyan, sok beszélővel rendelkező nyelv, melyek ezen nyelvközösségek szempontjából fontos szerepet játszanak: az angol, a finn, a magyar és az orosz. Egy forrásnyelv négy célnyelvvel alkothat párt, így összesen 24 nyelvpárral dolgozunk.

Az Expanded Graded Intergenerational Disruption Scale (EGIDS) [3] egy olyan skála, amely meghatározza a világ nyelveinek helyzetét, és segítségével osztályozható a nyelvek vitalitásának mértéke. A legmagasabb szint a 0., az ide tartozó nyelveket nemzetközi szinten, számos funkcióban használják, míg a 10. szinten levők kihalt nyelveknek számítanak. Az 1. táblázat összefoglalja az említett finnugor nyelvek jellemzőit. Minden nyelv mellett szerepel az ISO-639-3 nyelvkódja (a további táblázatokban is ezeket használjuk), az EGIDS-szintje, a beszélőinek a száma, az ország(ok), ahol beszélik és az írásrendszer, amit használ. Érdemes megfigyelni, hogy a legkevésbé veszélyeztetett nyelv az északi számi, annak ellenére, hogy a beszélőinek a száma a legkevesebb a felsorolt nyelvek közül. Ezek a számok az északi számi nyelvet célzó revitalizációs törekvések sikerességéről tanúskodnak.

nyelv	**ISO**	**EGIDS**	**népesség**	**terület**	**írás**
északi számi	sme	2	26.000	Norvégia, Finnország, Svédország	latin
mezei mari	mhr	4	470.000	Oroszország	cirill
hegyi mari	mrj	5	30.000	Oroszország	cirill
komi-zürjén	kpv	5	156.000	Oroszország	cirill
komi-permják	koi	5	63.000	Oroszország	cirill
udmurt	udm	5	340.000	Oroszország	cirill

1. táblázat. A kisebbségi finnugor nyelvek jellemzőinek összefoglalója.

2. Automatikus szótárépítés

2.1. Kitekintés

A kétnyelvű szótáraknak nem csak a nyelvtanulásban és a lexikográfiában van fontos szerepük, hanem olyan nyelvtechnológiai alkalmazásokban is, mint a gépi fordítás [4] és a nyelvközi információ-visszakeresés [5]. A kétnyelvű szótárak kézzel való előállítása időigényes feladat, mely nagy fokú hozzáértést és precizitást igényel. Ezért a kevés beszélővel rendelkező nyelvek számára nem gazdaságos ez a fajta szótárépítés. Komplett kétnyelvű szótárak teljesen automatikusan történő előállítását a jelenlegi technológia nem teszi lehetővé, ezért automatikus módszerekkel ún. protoszótárakat hoztunk létre, melyek fordítási jelölteket tartalmaznak, és kézi ellenőrzést igényelnek.

Az automatikus szótárépítés sztenderd megközelítése párhuzamos vagy összevethető korpuszokból történő kontextushasonlóság-számításon alapul [6]. Az elmúlt években a forrás- és célnyelvi szavakat reprezentáló vektorokat jellemzően szóbeágyazást alkalmazó módszerekkel nyerik ki [7]. Ezeknek a módszereknek az a hátránya, hogy nagy mennyiségű szöveget igényelnek. Viszont kivételes esetnek számít, ha egy adott nyelvpárra elég nagy párhuzamos vagy összevethető korpusz áll rendelkezésre; általánosnak inkább az tekinthető, ha nincs ilyen. Mivel az általunk vizsgált finnugor nyelvekre nincs kellően nagy korpusz, alternatív módszerekkel kísérleteztünk.

2.2. Az alkalmazott módszerek

A fent leírt okokból a protoszótárak előállításához két, közösség által épített nyelvi erőforrást használtunk fel, a Wikipédiát és a Wiktionaryt.

A Wikipédia[1] többféle módon is felhasználható kétnyelvű szótárak létrehozására. Mi Erdmann et al. [8] és Mohammadi és GhasemAghaee [9] módszerét követve kétnyelvű szótárakat hoztunk létre Wikipédia-címszópárokból a nyelvközi linkek segítségével.

A Wikipédia mellett a Wiktionary[2] egy másik, szintén nyílt, közösség által szerkesztett tudásbázis, amely forrásul szolgálhat kétnyelvű szótárak létrehozásához. A Wiktionary egy olyan többnyelvű szótár, amelynek célja, hogy minden nyelv minden szavát tartalmazza. Sok nyelven elérhető, és a definíciók, leírások mindig az adott Wiktionary nyelvén szerepelnek. Például a Wiktionary magyar kiadását Wikiszótárnak hívják, és a magyar szavak mellett tartalmaz más nyelvű szavakat is, de a hozzájuk tartozó definíciók és nyelvi információk magyarul szerepelnek. Nyelvenként eltérő, hogy mi az a minimális információ, amit tartalmaznia kell egy szócikknek, de maga a címszó, valamint az, hogy milyen nyelvű és milyen szófajú, kötelező információ minden Wiktionary-kiadás esetében. Nem kötelező elem, de bizonyos nyelvű Wiktionarykben szerepel egy fordítási tábla is, ami a címszó különböző nyelvű fordításait tartalmazza; szerepelhetnek továbbá a címszó IPA-átirata, ragozási paradigmája, szinonimái és egyéb lexikai információk is az egyes szócikkekben.

Bár a Wiktionary elsősorban emberi felhasználásra készült, a benne található adatok kinyerése bizonyos fokig automatizálható. Ács et al. [10] minden címszóhoz tartozó fordítási megfelelőt kinyert a szócikkekben található fordítási táblákból. Az általuk fejlesztett `Wikt2dict`[3] eszközzel feldolgoztuk az angol, finn, orosz és magyar Wiktionary-oldalakat, így szinte minden szóban forgó nyelvpárra sikerült fordítási párokat kinyernünk.

Ács [11] a szópárok halmazát újabbakkal bővítette úgy, hogy háromszögeléssel új kapcsolatokat hozott létre a már meglévő fordítási párokból. A háromszögelés azon a feltételezésen alapul, hogy két elem nagy valószínűséggel fordításpár abban az esetben, ha mindkettő egy harmadik nyelv szavának fordítása. A

[1] `https://www.wikipedia.org/`
[2] `https://www.wiktionary.org/`
[3] `https://github.com/juditacs/wikt2dict`

`Wikt2dict` háromszögelési technikájával protoszótárainkat tovább tudtuk bővíteni.

2.3. Kiértékelés

Első lépésként a különböző módszerekkel előállított protoszótárakat célnyelvenként összevontuk, majd az ismétlődő szópárokat kiszűrtük. Az összevont szótárak kézi kiértékelését az adott nyelvek anyanyelvi beszélői és nyelvész szakértői végezték. Az instrukciók, melyek szerint dolgoztak, az alábbiak: a forrás- (S) és célnyelvi (T) szónak is az adott nyelven létező szónak kell lennie, szótári alakban kell állnia, és a két szónak egymás fordításának kell lennie. Ha a forrásnyelvi szó nem létező szó, a szópárt hibásnak kell jelölni. Ha a forrásnyelvi szó létező szó, de nem szótári alakban áll, meg kell adni a helyes szótári alakot. Ha a célnyelvi szó jó fordítása a forrásnyelvi szónak, de nem szótári alakban áll, meg kell adni a helyes szótári alakot. Ha a célnyelvi szó nem jó fordítás, új fordítást kell megadni.

A kiértékelés során bevezettünk olyan kategóriákat, melyek azt jelzik, hogy egy adott szópár megfelel-e ezeknek az instrukcióknak, vagy sem. A kategóriák a következők:

ok-ok: az S és T szavak létező, szótári alakban álló szavak, és egymás fordításai;
ok-nd: az S és T szavak létező szavak, egymás fordításai, de a T szó nem szótári alakban szerepel;
nd-ok: az S és T szavak létező szavak, egymás fordításai, de az S szó nem szótári alakban szerepel;
nd-nd: az S és T szavak létező szavak, egymás fordításai, de sem az S, sem a T szó nem szótári alakban szerepel;
ok-wr: az S szó létező, szótári alakban szereplő szó, de a T szó vagy nem létező szó, vagy nem helyes fordítása az S szónak;
nd-wr: az S szó létező, de nem szótári alakban szereplő szó, a T szó pedig vagy nem létező szó, vagy nem helyes fordítása az S szónak;
wr-xx: az S szó az adott nyelven nem létező szó.

Az automatikusan létrehozott protoszótárak kézi kiértékelése és javítása több célt is szolgál. Egyrészt lehetőséget ad az általunk használt szótárépítési módszerek összehasonlítására – ezt ismertetjük ebben a fejezetben. Másrészt megadja azoknak a szópároknak a számát, amelyeket feltölthetünk a Wiktionarybe – erről lásd a 3.5. fejezetet. A kézi kiértékelésnek az a lépése, amelynek során új fordítást kellett megadni, ha a célnyelvi szó nem volt megfelelő, ez utóbbihoz járul hozzá. Ennek célja ugyanis a Wiktionarybe feltölthető szópárok számának növelése volt – az automatikus módszerek kiértékelésében a megadott új fordítások nem játszottak szerepet.

A 2. táblázat tartalmazza azokat az adatokat, amelyek lehetőséget adnak az automatikus szótárépítő módszerek összehasonlítására. A táblázat első három sorában a fent leírt módszerek szerepelnek: W2D ext: a `Wikt2dict` alkalmazása fordítási jelöltek kinyerésére a fordítási táblákból, W2D tri: a `Wikt2dict` alkalmazása háromszögelésre, WikiTitle: Wikipédia-címszópárok kinyerése. A negyedik sorban (KDE4) olyan szótárak szerepelnek, amelyeket az Opus korpuszból

[12] töltöttünk le. A letöltés idejében az általunk vizsgált nyelvpárok közül csak az északi számi–{angol, finn, magyar} nyelvpárokra találtunk szótárakat. Ezek a szótárak a 2.1. fejezetben említett sztenderd szótárépítő módszerekkel készültek, amelyekről azt feltételeztük, hogy nem lesznek megfelelőek az általunk vizsgált kevés erőforrással rendelkező nyelvek esetében. A 2. táblázatban látható adatok ezt alátámasztják, hiszen a 27,57%-os ok-ok kategóriájával ez a módszer volt a legkevésbé pontos.

módszer	**össz**	**ok-ok**	**ok-nd**	**nd-ok**	**nd-nd**	**ok-wr**	**nd-wr**	**wr-xx**
	(#)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
W2D ext	1.965	71,76	1,22	5,75	15,17	4,63	0,36	0,76
W2D tri	23.066	56,61	1,79	2,98	3,06	30,38	1,10	3,82
WikiTitle	16.854	54,11	2,97	5,57	32,50	2,92	0,49	0,75
KDE4	8.401	27,57	3,99	10,40	18,64	13,99	14,57	10,69

2. táblázat. Az egyes módszerek összehasonlítása.

A táblázat első oszlopában az adott módszerrel létrehozott protoszótárak összes szópárának a számát látjuk. A további oszlopokban a kézi kiértékelésnél az egyes kategóriákba sorolt szópárok százalékos arányát látjuk, vagyis hogy az összes előállított szópárból hány esett az ok-ok, ..., wr-xx kategóriába. A pontosságot szigorúan értelmezzük, és csak azt tartjuk pontosnak, ami az ok-ok kategóriába tartozik. Eszerint a fordítási megfelelőknek a Wiktionary fordítási tábláiból való kinyerése bizonyult a legpontosabb módszernek, ami nem annyira meglepő, hiszen ezek az adatok szerkesztők kézi munkájával álltak elő. A második legpontosabb módszer a háromszögelés volt: itt egy 15%-os visszaesést látunk, aminek az lehet az oka, hogy ez a módszer nem közvetlenül az emberi szerkesztőmunkára épít, hanem olyan pontok között feltételez kapcsolatot, amelyek között nem feltétlenül van, ami elsősorban a poliszémiának köszönhető. Némileg meglepő módon a Wikipédia-címszópárokból építkező módszer csak a harmadik lett, pedig erre is igaz az, hogy a megbízhatónak gondolt kézi szerkesztői munkára támaszkodik. Vegyük észre a kiemelkedően magas nd-nd számot, ami valószínűleg annak köszönhető, hogy a Wikipédia-címszavak közt sok a nem szótári alak, különösen az állat- és növénynevek körében. Például a *мераг-влак* ∼ *nyúlfélék* szópár esetében mindkét nyelvi megfelelő a biológiai besorolást követve többes számban van, de a kézi validálás során ezek le lettek javítva az egyes számú alakjaikra: *мераг* ∼ *nyúl*.

A nyelvpáronként összevont protoszótárak kiértékelése és a hasznos szópárok száma a 3. táblázatban látható. Ebben az esetben azt a szópárt tekintjük hasznosnak, amely az automatikus szótárépítés és a kézi kiértékelés teljes munkamenetének a végén tartalmaz egy létező forrásnyelvi és egy létező célnyelvi szót, és ezek jó fordításai egymásnak.

nyelvpár	össz (#)	hasznos (%)	ok-ok (%)	ok-nd (%)	nd-ok (%)	nd-nd (%)	ok-wr (%)	nd-wr (%)	wr-xx (%)
koi–eng	1.251	82,81	74,82	0,16	7,83	0,00	13,67	0,16	3,36
koi–fin	592	79,05	65,20	3,04	9,97	0,84	19,59	0,17	1,18
koi–hun	540	79,45	70,19	3,33	4,63	1,30	13,52	0,19	6,85
koi–rus	611	86,58	65,47	2,95	16,69	1,47	11,62	0,65	1,15
kpv–eng	902	97,23	66,30	0,22	0,55	30,16	2,55	0,22	0,00
kpv–fin	577	98,79	51,13	9,88	0,69	37,09	1,04	0,17	0,00
kpv–hun	523	96,18	49,90	1,34	0,96	43,98	3,63	0,00	0,19
kpv–rus	544	96,69	63,60	8,64	9,93	14,52	3,31	0,00	0,00
mhr–eng	2.549	73,40	44,41	2,55	4,04	22,40	26,09	0,51	0,00
mhr–fin	2.565	75,90	50,80	1,05	3,31	20,74	23,63	0,47	0,00
mhr–hun	1.647	84,64	52,70	0,97	5,89	25,08	14,21	1,15	0,00
mhr–rus	1.707	63,68	40,01	2,11	4,28	17,28	35,56	0,76	0,00
mrj–eng	2.334	96,40	44,09	0,17	9,04	43,10	3,08	0,51	0,00
mrj–fin	1.013	90,03	20,24	7,70	9,77	52,32	8,59	1,38	0,00
mrj–hun	942	93,20	34,18	4,99	12,95	41,08	5,20	1,59	0,00
mrj–rus	835	79,29	27,07	11,26	9,58	31,38	16,89	3,83	0,00
sme–eng	6.041	65,23	47,57	3,77	7,33	6,56	21,65	5,08	8,03
sme–fin	7.100	63,42	42,00	3,44	5,42	12,56	19,94	7,65	8,97
sme–hun	4.969	63,41	48,48	1,67	6,62	6,64	17,05	10,28	9,26
sme–rus	4.373	74,54	71,30	0,50	2,56	0,18	20,10	0,75	4,60
udm–eng	2.087	81,46	77,10	3,16	0,91	0,29	17,59	0,10	0,86
udm–fin	1.700	71,06	49,12	2,06	1,06	18,82	28,06	0,53	0,35
udm–hun	1.204	83,55	57,14	1,74	1,50	23,17	15,45	0,50	0,50
udm–rus	1.226	76,83	8,56	2,04	0,98	65,25	20,64	1,31	1,22

3. táblázat. A protoszótárak kiértékelése az egyes nyelvpárokra.

3. Wiktionary-szócikkek generálása

A Wiktionary-szócikkek alapjául a validált protoszótárak szolgálnak. Például az északi számi–angol nyelvpár esetén az északi számi szó az angol Wiktionary egy új címszava lesz, míg az angol megfelelője a definíció lesz. A szócikkek kötelező elemei és a kiegészítő információk is teljesen automatikusan lettek előállítva.

Minden Wiktionary-kiadásnak megvannak a maga szabályai a szócikk felépítését illetően. Nem csak a formára vonatkozó szabályokat írják le, hanem azt is, hogy milyen nyelvtani információkat kell tartalmaznia egy szócikknek. A négy célnyelvi Wiktionary leírásai alapján sikerült egy olyan általános felépítést meghatározni, mely tartalmazza a címszót, a címszó nyelvét, annak szófaját és a fordítási megfelelőjét. Ezen kötelező elemek közül csak a szófaji címke hiányzik a szótárainkból. Kiegészítő információkkal is lehet bővíteni az egyes szópárokat, például IPA-átírással vagy etimológiával, azonban ezek nem kötelező elemei a szócikkeknek.

3.1. A szófaji kategória meghatározása

Az új Wiktionary-szócikkek létrehozása sablonokkal történik, melyekhez szükséges az adott szó szófaji címkéje. Ahhoz tehát, hogy egy szóból Wiktionary-szócikk jöhessen létre, meg kell állapítanunk a szófaját. Ezt az információt morfológiai elemzők segítségével lehet előállítani, amelyek azonban egy szóhoz több elemzést is adhatnak. A morfoszintaktikai egyértelműsítés jellemzően kontextuális információk alapján történik, de mivel itt elszigetelt szavakról van szó, más módszert kellett találnunk.

Mind a forrásnyelvek, mind a célnyelvek számára létezik morfológiai elemző, amelyet a szófaji kategória megállapításához használhattunk. A Giellatekno[4] elemzőit használtuk az összes forrásnyelvre, valamint a célnyelvek közül a finnre és az oroszra. A magyar szavakat az `emMorph` [13] segítségével elemeztük, míg az angol szavak elemzésére a `hunmorph` elemzőt [14] használtuk a `morphdb` [15] angol nyelvű lexikon- és szabályfájljaiból előállított `aff` és `dic` fájlokkal. Mivel a morfológiai elemzők különböző kimeneti formátummal rendelkeznek, a szófaji címkéket le kellett képeznünk egy közös címkekészletre.

A morfológiai elemzők csak szavakra adnak elemzést, ezért a többszavas kifejezéseket külön kellett kezelni. Ezekben az esetekben a kifejezés az utolsó elemével lett átmenetileg helyettesítve. Emögött az a feltételezés áll, hogy a finnugor nyelvek általában fejvégűek, vagyis ha az utolsó elemhez megfelelő szófaji címkét kapunk, azt a teljes kifejezéshez hozzá tudjuk rendelni. Az angol és az orosz nyelvek viszont inkább fejkezdetűként vannak számon tartva, így ez a megközelítés nem minden esetben működik jól.

A kézi ellenőrzés során a validátorok az angol igék elé sokszor egy 'to' partikulát illesztettek. Ez a partikula a morfológiai elemző bemenetéből el lett távolítva, de megőriztük – több okból is. Egyrészt később az egyértelműsítéshez fel tudjuk használni, másrészt az angol Wiktionary előírja, hogy az igék előtt szerepelnie kell a 'to'-nak, így azok visszakerülnek az igék elé a szócikkek generálásakor.

Ennek a lépésnek a kimenete egy öt oszlopos táblázat a forrásnyelvi szóval, annak lehetséges szófaji címkéivel, a célnyelvi szóval, annak lehetséges szófaji címkéivel és egy ötödik oszloppal, mely a 'to' partikulát tartalmazza.

3.2. A szófaji címkék egyértelműsítése

A szófaji címkék egyértelműsítésének három fázisa van. Először csak az elemzőtől kapott morfológiai információk alapján szűrjük a lehetséges kategóriákat. Ezután egy horizontális összehasonlítást végzünk, melynek során a forrás- és a célnyelvi szó címkéit hasonlítjuk össze, és így szűkítjük a halmaz lehetséges elemeit. Végül vertikálisan is megvizsgáljuk egy adott forrásnyelvi szóhoz tartozó szófaji címkék halmazát, mikor az több fordítási pár forrásnyelvi megfelelőjeként is szerepel.

Morfológiai információk alapján történő szűrés. Az elemzők nem csak a szófaji címkét bocsátják ki, hanem a lemmát is, valamint információt kapunk

[4] `http://giellatekno.uit.no`

az esetről és a számról is, amiket fel lehet használni az egyértelműsítés során. Mivel a validált protoszótárak csak szótári alakokat tartalmaznak, a bemenetként adott szónak azonosnak kell lennie a lemmával. Azok az elemzések, melyekre ez nem áll, törlésre kerülnek. A számot és az esetet arra lehet felhasználni, hogy a lehetséges elemzések közül kiszűrjük azokat, amelyek egy szótári alak elemzései lehetnek. A szótári alak a névszók esetében az egyes számú alanyesetű alak, míg igék esetében a magyarban az egyes szám harmadik személyű jelen idejű kijelentő módú határozatlan ragozású alak, angolban az infinitívusz. Ha a szám és az eset nem ezeknek megfelelően alakul, akkor azt az elemzést töröljük.

Vannak esetek, mikor egyik lemma sem egyezik meg a bemeneti szóval. A többszavas kifejezések esetében például a bemenet csak a kifejezés utolsó tagja volt, így a sztring végén várhatunk egyezést. Ha egyik elemzés sem felel meg a feltételeknek, a lehetséges szófajok halmaza üres marad.

Előfordulhat, hogy egy szó több olyan címkét is kap, melyek közül az egyik a másik részhalmaza. Ilyenkor a legszűkebb kategóriát tartjuk meg: például ha a halmazban az `N` és a `Prop` (tulajdonnév) is szerepel, akkor az utóbbi marad meg.

Horizontális összehasonlítás. Az így kapott címkék körét még tovább lehet szűkíteni úgy, hogy a forrás- és a célnyelvi szó címkéit összehasonlítjuk. A szópárnak két lehetséges szófajcímkehalmaza van, és feltételezve, hogy az egy szópárban szereplő szavak ugyanazon szófajhoz tartoznak, le lehet szűkíteni a címkék számát. Vagyis itt a két címkehalmaz metszetét vizsgáljuk meg, aminek során a következő esetek állhatnak elő.

Ha a metszet üres, el kell dönteni, melyik szó címkéit tartjuk meg. Elsődlegesen a forrásnyelvi szó címkéit kapja meg a szópár – ha nincs neki, akkor a célnyelviét. Ez utóbbi esetben a többszavas kifejezéseknél nehézségek merülhetnek fel. Az angolnál a főnévi frázis esetében láttunk arra példákat, hogy a fej mégis a frázis végén állhat, ezért hoztunk egy olyan szabályt, hogy ha a célnyelvi szó többtagú, a célnyelv angol, és a lehetséges szófaji címkék között nem szerepel az `N`, a szópár lekerül a leendő Wiktionary-szócikkek listájáról.

Ha több címke is van a lehetséges szófaji címkék metszetében, különböző szabályok alapján próbáljuk meg tovább szűrni a címkéket. Az egyik ilyen a finnugor nyelvek igéinek végződésein alapul. Például ha egy északi számi szó -*t*-re végződik, és a közös szófaji címkék között megtalálható a `V` címke, akkor a szófaja ige lesz. A korábban megtartott 'to' partikula az angol igéknél segíthet.

Ha a két halmaz metszete csak egy címkét tartalmaz, ez a helyes szófajnak tekinthető, melyet rögzítünk.

Vertikális összehasonlítás. Az előző fázisokban egyértelműsített és rögzített szófaji címkéket fel lehet használni a címkék további egyértelműsítéséhez abban az esetben, ha egy forrásnyelvi szó több szópárban is megjelenik. A vertikális összehasonlítás azon a megfigyelésen alapul, hogy egy forrásnyelvi szó akkor jelenik meg több szópárban, ha a célnyelvi szavak egymás szinonimái, vagyis feltételezhető, hogy ugyanolyan szófajúak.

Vannak azonban esetek, amikor az ilyen szópároknak több mint egy címkéje van. Például a komi-permják *ань* szónak három megfelelője van az angolban: *female*, *mother* és *woman*. Ezek a szópárok különböző címkehalmazokkal rendelkeznek: a *female* `N` és `A` (melléknév), míg a *mother* és a *woman* `N` és `V`. A három halmaz metszete megadja a komi-permják szó címkéjét, ami `N`.

Természetesen vannak olyan esetek is, amikor egy szóhoz nem csak egy helyes szófaji címkét lehet rendelni. Például a mezei mari *нарынче* ('sárga') szó melléknév és főnév is, mint a magyarban. Ha további egyértelműsítés már nem lehetséges, minden szófaji címkét megtartunk.

A teljes egyértelműsítés kimenete három oszlopból áll: a forrásnyelvi szó, a célnyelvi szó, és az így kapott szófaji címke. Ha egy fordítási párhoz több címkét rendeltünk, az első két oszlop ismétlődik.

3.3. Fonetikai átírás

A következő lépésben fonetikai átírást rendeltünk a forrásnyelvi szavakhoz, hogy ezzel is gazdagítsuk a Wiktionary-szócikkek tartalmát. Ehhez a Mari Web Project automatikus IPA-átírási eszközét [16] használtuk a hegyi mari, mezei mari, komi-permják, komi-zürjén és udmurt nyelvekre. Az északi számira a Giellatekno `text2ipa` forrásfájljaiból[5] kompilált FST-t használtuk.

Mivel a tulajdonnevek kiejtése nem feltétlenül követi az adott nyelv fonetikai szabályait, a nevekhez nem adtunk IPA-megfelelőt. Hasonlóan problémások a számot tartalmazó szavak, amelyekhez szintén nem rendeltünk átírást.

3.4. A szócikkek előállítása és feltöltése

Miután minden információ a rendelkezésünkre áll, a szócikkek előállítása következik. Annak ellenére, hogy minden Wiktionary-kiadás más-más szabályokkal rendelkezik, és máshogyan határozza meg a szócikkek felépítését, lehetséges volt egy olyan általános struktúra létrehozása, mely mind a négy kiadás számára elfogadható felépítésként szolgál.

A szócikkek generálása közben a legfrissebb Wiktionary-dumpokban ellenőrizzük, hogy az adott forrásnyelvi szó létezik-e már az adott nyelvű Wiktionaryben. Ha igen, nem készül az adott szóhoz Wiktionary-szócikk. Azokban az esetekben, amikor egy forrásnyelvi szóhoz több célnyelvi szó tartozik, és ezek szófaja ugyanaz, a szócikkben ugyanazon fejléc alá kell besorolni őket, így szét kell válogatni a jelentéseket szófajok szerint. Amennyiben egy fordítási pár több szófajjal is rendelkezik, a célnyelvi szót minden fejléc alatt megismételjük.

Minden szócikk rendelkezik címszóval, mely esetünkben a forrásnyelvi szó. Minden szócikk legalább egy szófaji fejléccel rendelkezik, valamint legalább egy fordítással, azaz célnyelvi megfelelővel. Amennyiben a forrásnyelvi szó nem csak egy forrásnyelvben szerepel a szótárainkban, ezen szócikkeket egyesítjük.

A bejegyzések Wiktionarybe való feltöltéséhez a MediaWiki `Pywikibot`[6] nevű keretrendszerét használtuk. A `Pywikibot` sima szöveges állományokból generál

[5] `https://victorio.uit.no/langtech/trunk/langs/sme/src/phonetics/`
[6] `https://www.mediawiki.org/wiki/Manual:Pywikibot`

wiki oldalakat, amiket automatikusan feltölt a megadott nyelvű Wiktionarybe. A botnak megadható, hogy a már létező szócikkeket ne írja felül – vagyis itt is elvesztünk néhányat a feltölthető cikkekből.

A botok használatát mindazonáltal erősen szabályozza a Wiktionaryk szerkesztősége. Mivel az engedélyek beszerzése még folyamatban van, ezért csak a frissen letöltött Wiktionary dumpok[7] alapján tudunk kiértékelést adni az általunk létrehozott és feltöltött új Wiktionary-szócikkekről.

3.5. Kiértékelés

Ebben a fejezetben a Wiktionary-szócikkek generálásának kiértékelését mutatjuk be. A 2.3. fejezetben a szótárak pontosságát írtuk le. Egy szótár fedését kiszámolni viszont közel sem ennyire triviális feladat. Az egyik lehetséges megközelítés, ha a szótárban szereplő szavak számát egy másik – ideális esetben kézzel készített – szótár szavainak számával vetjük össze. Mivel a szópárokat a Wiktionarybe töltjük fel, a fedés kiszámításához is azt használtuk fel. Jóllehet a Wiktionary nem kizárólag kézzel készült, mégis szerkesztők ezrei ellenőrzik a szócikkeket, vagyis megbízható adatforrásnak tekinthető.

Az 4. táblázatban láthatók az eredmények. A nyelvek az ISO 639-3 kódjuk alapján azonosíthatók. A Wiktionary több szerkesztője nem tesz különbséget egyes nyelvek között, hanem a megfelelő makronyelv kódját használja (a mari nyelv esetében chm, míg a komi nyelv esetében kom). Emiatt a két mari, valamint a két komi nyelv szótárait össze kellett vonni. Az itt látható eredményeket a cikk írásának idején legfrissebb Wiktionary-dumpok alapján határoztuk meg, és nem a ténylegesen feltöltött elemek alapján. Mivel a Wiktionary egy állandóan frissülő és bővülő szótár, valódi kiértékelést csak a feltöltést követően tudunk csinálni.

A táblázat első oszlopában (‘hasznos’) a hasznosnak bizonyuló szópárok száma található. Ez minden olyan szópárt tartalmaz, ahol a forrásnyelvi szó létező szó, mivel a helyes szótári alak és a pontos, szótári alakban álló fordítás is meg lett adva a kézi ellenőrzés során. A hasznos szavak száma annyival csökken, ahány forrásnyelvi szóhoz nem sikerült szófaji címkét előállítani. A szócikkek feltöltése előtt ellenőriztük azt is, hogy az adott szó létezik-e a megfelelő Wiktionary-kiadásban. Amennyiben igen, a feltölthető szavak száma ezzel is csökken. A ‘maradék’ oszlopban a feltöltésre kész szópárok száma található. A ‘wikt’ oszlopban a már a Wiktionaryben szereplő szavak száma látható, ennek és a ‘maradék’ oszlopnak a metszetét a ‘közös’ oszlopban láthatjuk. A ‘wikt’ és a ‘közös’ oszlopokból kiszámolható az általunk létrehozott szócikkek száma (‘új’). Ezen számok birtokában az egyes szótáraknak egyfajta fedése is kiszámolható: a ‘fedés’ oszlopban található szám a ‘közös’ és a ‘wikt’ oszlopok számainak a hányadosa. Az utolsó oszlop (’növ’) az általunk létrehozott új és a már létező szócikkek számának a hányadosa, mely az adott célnyelvi Wiktionary-kiadásban található adott forrásnyelvi szócikkek számának a növekedését mutatja.

[7] A kiértékeléshez használt Wiktionary dumpok dátuma: eng: 2017. november 6., fin: 2017. november 5., rus: 2017. november 7., hun: 2017. november 6.

nyelvpár	hasznos (#)	maradék (#)	wikt (#)	közös (#)	új (#)	fedés (%)	növ (%)
kom–eng	2.111	656	54	25	631	46,30	1.168,52
kom–fin	1.162	687	42	27	660	64,29	1.571,43
kom–hun	1.025	699	152	35	664	23,03	436,84
kom–rus	1.148	673	465	223	450	47,96	96,77
chm–eng	4.883	1.671	347	53	1.618	15,27	466,28
chm–fin	3.578	1.905	443	213	1.692	48,08	381,94
chm–hun	2.589	1.634	34	12	1.622	35,29	4.770,59
chm–rus	2.542	1.497	848	202	1.295	23,82	152,71
sme–eng	5.556	2.531	4.073	882	1.649	21,65	40,49
sme–fin	6.463	2.862	817	422	2.440	51,65	298,65
sme–hun	4.509	2.392	206	146	2.246	70,87	1.090,29
sme–rus	4.172	2.034	306	237	1.797	77,45	587,25
udm–eng	2.069	754	32	15	739	46,88	2.309,38
udm–fin	1.694	828	55	45	783	81,82	1.423,64
udm–hun	1.198	739	128	69	670	53,91	523,44
udm–rus	1.211	578	644	247	331	38,35	51,40

4. táblázat. A létrehozott szócikkek kiértékelése az egyes nyelvpárokra.

4. Összegzés

Cikkünkben egy olyan projektet mutattunk be, amelynek célja, hogy olyan protoszótárakat hozzunk létre automatikus módszerekkel, amelyeknél a forrásnyelv a komi-zürjén, komi-permják, mezei és hegyi mari, udmurt, valamint északi sámi nyelvek egyike, míg a célnyelv az alábbi, sok beszélővel rendelkező nyelvek közül kerül ki: angol, finn, magyar, orosz. Mivel a sztenderd szótárépítési módszerek nagy mennyiségű szöveget igényelnek, és mi kevés erőforrással rendelkező finnugor nyelvekkel dolgoztunk, alternatív szótárépítési megoldásokat kellett találnunk, melyekkel protoszótárakat hoztunk létre minden nyelvpárra.

Az automatikus szótárépítési módszereket és a kézi validálás során előállt szótárakat is kiértékeltük. A kiértékelés során a legpontosabb fordítási párokat eredményező megoldásnak a Wiktionary-alapú módszerek bizonyultak. De a Wiktionary nem csak forrásként szolgál számunkra, hanem ide töltjük fel az automatikusan generált szócikkeinket is. A szócikkek feltöltéséhez még nem kaptunk meg minden engedélyt, ezért csak a legfrissebb Wiktionary-dumpokkal összehasonlítva tudunk kiértékelést adni, de az eredményekből már most látszik, hogy a célnyelvi Wiktionarykben szereplő forrásnyelvi lexikai elemek számát megsokszoroztuk.

Köszönetnyilvánítás

A projektet az Országos Tudományos Kutatási Alapprogram támogatja (szerződésszám: FNN 107885).

Hivatkozások

1. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193** (2012) 217–250
2. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: Developing and Aligned Multilingual Database. In: Proceedings of the First International Conference on Global WordNet, Mysore, India (2002) 293–302
3. Lewis, M.P., Simons, G.F.: Assessing endangerment: Expanding Fishman's GIDS. Revue Roumaine de Linguistique **55**(2) (2010) 103–120
4. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. Computing Research Repository **1–10.**(abs/1309.4168) (2013)
5. Levow, G.A., Oard, D.W., Resnik, P.: Dictionary-based techniques for cross-language information retrieval. Information Processing & Management **41**(3) (2005) 523–547
6. Fung, P., Yee, L.Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In: 17th ACL. (1998) 414–420
7. Vulić, I., Moens, M.F.: Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In: 53rd ACL. (2015) 719–725
8. Erdmann, M., Nakayama, K., Hara, T., Nishio, S.: An Approach for Extracting Bilingual Terminology from Wikipedia. ACM Transactions on Multimedia Computing, Communications, and Applications **5**(4) (2009) 1–17
9. Mohammadi, M., Ghasem-Aghaee, N.: Building Bilingual Parallel Corpora Based on Wikipedia. In: 2nd International Conference on Computer Engineering and Applications. (2010) 264–268
10. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: 6th Workshop on Building and Using Comparable Corpora, Sofia, ACL (2013) 52–58
11. Ács, J.: Pivot-based multilingual dictionary building using Wiktionary. In: 9th Language Resources and Evaluation Conference, Reykjavik, ELRA (2014)
12. Tiedemann, J.: News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In: Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007. John Benjamins, Borovets (2009) 237–248
13. Novák, A., Rebrus, P., Ludányi, Zs.: Az `emMorph` morfológiai elemző annotációs formalizmusa. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017), Szeged (2017) 70–78
14. Trón, V., Gyepesi, Gy., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: Open Source Word Analysis. In: Proceedings of the ACL Workshop on Software, Ann Arbor, Michigan, ACL (2005) 77–85
15. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., Simon, E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of LREC'06. (2006) 1670–1673
16. Bradley, J.: Transcribe.mari-language.com. Acta Linguistica Academica **64**(3) (2017) 369–382

Towards cross-lingual utilization of sparse word representations

Gábor Berend

University of Szeged,
Department of Informatics
2. Árpád tér, 6720, Szeged, Hungary
berendg@inf.u-szeged.hu

Abstract. In this paper, we introduce our approach for defining sparse word representations that are comparable across languages. The proposed approach is designed to require as little linguistic resources as possible. Our experimental results suggests that it is sufficient to rely on an artificially created highly noisy dictionary to map sparse word representations across languages.

1 Introduction

Distributed word representations [1,2] aim at representing symbolic word forms as some relatively low dimensional continuous vectors. Such word representations have been found to be extremely useful in a variety of natural language processing tasks. In fact most state-of-the-art approaches in NLP rely on their utilization [3].

Over the past few years, there has also been an emerging research interest in constructing word representations which are applicable over a variety of languages, see e.g. [4] for an empirical comparison. Providing word representations which are comparable across languages have huge potential because they could enable the effortless creation of machine learning models which can operate across languages. As an illustrative example, if one has access to comparable word representations for both English and French, it becomes possible to train a document classifier purely on English-written documents and then reliably apply it on French documents as well.

A further line of research which has gained attention is related to the application of sparse word representations [5,6]. Such sparse word representations have been shown to often outperform the application of dense word representations in a monolingual setting. However, to the best of our knowledge, there has not been any work conducted towards the creation of sparse word representations which are comparable across languages. In this paper, we present our proposed solution for solving that task by creating and evaluating sparse word representations for English and Hungarian which are intended to be used interchangeably.

2 The proposed solution

In this section we introduce our proposed solution for determining sparse word representations which are comparable across languages. Throughout the rest of the paper we shall denote the dense word embeddings of the source (English) and target (Hungarian) languages by $S \in \mathbb{R}^{k \times |V_s|}$ and $T \in \mathbb{R}^{k \times |V_t|}$, respectively, with V_s and V_t indicating the vocabulary of the two languages. Recall that in the followings we shall denote some symbolic word form as x and the vectorial representation that gets assigned to it in boldface, i.e. $\mathbf{x}$.

2.1 Preprocessing of continuous word embeddings

Our algorithm takes as input a pair of "traditional" dense word embeddings for both the source and target languages trained independently. Pre-trained word embeddings for a variety of languages are already accessible and we relied on the 64-dimensional pre-trained polyglot word embeddings [7]. Utilizing fasttext embeddings [8] would have been a viable alternative. In fact, the proposed approach naturally generalizes to the application of word embeddings trained in any other kinds of ways. We also conducted experiments with dense word embeddings trained according to the popular continuous bag-of-words and skip-gram models but experienced no significant difference in the results.

We performed experiments with the embeddings in their unmodified form as well as experiments that modified the pre-trained word embeddings in ways commonly met in the literature [9,10]. The two kinds of preprocessing steps we performed are making the word embeddings unit length and centralizing the word representations across all the dimensions.

2.2 Mapping of word embeddings

In order to map word representations between English and Hungarian that are trained independently, we perform a linear mapping which brings Hungarian word representations as close to semantically similar English word representations as possible. As firstly proposed in [11] such a linear mapping M can be defined by minimizing the objective function $\sum_{i=1}^{n} \|M\mathbf{s_i} - \mathbf{t_i}\|$, with $\{(s_i, t_i)\}_{i=1}^{n}$ being the seed set of word pairs which are cross-lingual equivalents of each other.

Multiple studies have showed recently that ensuring the additional constraint for M to be orthonormal can significantly improve the quality of the mapping of word embeddings from one language to the embedding space of another one [12,9,10]. Requiring M to be orthonormal can be especially useful if the set of seed word pairs used for determining the linear mapping contains a high fraction of noise, i.e. erroneously aligned word pairs.

Since our goal is to design an approach for mapping sparse word representations which relies on as little human labor and external resources (such as parallel text) as possible, we created a pseudo-dictionary similar to [12]. We aligned word forms that are present in their exact same surface forms in the

vocabulary of both languages. Recall that the pseudo-dictionary created in such a manner undoubtedly contains substantial noise, e.g. the English noun "*hat*" (referring to the clothing accessory) gets aligned to the Hungarian word with the same surface form which – on the other hand – can either refer to a numeral and a verb as well.

The pseudo-dictionary constructed that way consists of 20,292 entries. The full size of the English and Hungarian vocabularies of the polyglot vectors include 100K and 150K word forms. The large number of aligned word pairs relative to the vocabulary sizes also implies that this automatically generated dictionary is not really reliable for which reason the orthonormal constraint for the linear mapping matrix M is expected to provide important gains over its unconstrained counterpart.

2.3 Determining the cross lingual representations

Previous studies have shown that trying to reconstruct dense word embeddings as a sparse linear combination of an overcomplete set of basis vectors can provide useful word representations [5,6]. More formally, for some word embedding matrix $X \in \mathbb{R}^{k \times |V|}$, such approaches seek to find a decomposition for X such that $\|X - D\alpha\|_F + \lambda\|\alpha\|_1$ gets minimized, with $D \in \mathbb{R}^{k \times l}$ containing the set of overcomplete basis vectors and $\alpha \in \mathbb{R}^{l \times |V|}$ containing the sparse linear coefficients for the individual word forms, respectively. In the previous expression λ is the regularization coefficient which controls for the amount of sparsity emerging in the sparse coefficient matrix α.

The kind of decomposition we employ in this work differs both from [5] and [6] as here we not only require D to be a member of the convex set of matrices comprising of unit norm column vectors, but also enforce the coefficients in α to be non-negative. We used the SPAMS package accompanying [13] for performing the matrix decompositions for our experiments.

Putting the various steps together, our approach can be summarized in the following:

1. (Optionally) pre-process the embedding matrices S and T by making the embeddings unit long and centered at the origin
2. Create a (pseudo-)dictionary $\{(s_i, t_i)\}_{i=1}^{n}$
3. Find M for which $\sum_{i=1}^{n} \|M\mathbf{s_i} - \mathbf{t_i}\|$ is minimized (with the optional constraint for M to be orthonormal).
4. Find D_s and α_s such that $\|S - D_s\alpha_s\|_F + \lambda\|\alpha_s\|_1$ gets minimized
5. Find α_t by relying on D_s such that $\|MT - D_s\alpha_t\|_F + \lambda\|\alpha_t\|_1$ is minimized.

Since target word vectors are mapped to the embedding space of source word vectors and the same dictionary matrix D_s is used for decomposing the embedding matrices S and T, the nonzero coefficients in α_s and α_t provide a sparse representation being comparable across languages. Note that if we choose M to be the identity matrix, we can keep the target language embeddings intact.

3 Experiments

During our experiments, we treated English as the source language and Hungarian as the target language. Throughout our experiments, we relied on pre-trained polyglot embeddings [7] which are publicly available for a variety of languages.[1] In order to conduct comparable experiments with previous studies, we applied the same number of basis vectors (i.e. 1024) which was previously utilized in [5]. When setting the value for λ, we chose the regularization coefficient from $\{0.1, 0.3, 0.5\}$ in order to see how different sparsity levels influence results. The sparse word representation that we define for some word form w_i is simply taken by the indices of the positive coefficients in α for the particular word, that is $\phi(\mathbf{w}_i) = \{j \mid \boldsymbol{\alpha}_i[j] > 0\}$, where $\boldsymbol{\alpha}_i$ refers to the vector of sparse coefficients determined for word w_i during the sparse decomposition procedure.

3.1 Evaluation on the Swadesh word list

Swadesh lists are collections of English words of varying cardinalities that are collected as part of a universal basic vocabulary[14]. The elements of the lists are expected to be found in the majority of languages. Consequently, the lists have been translated into various languages, including Hungarian as well.

There exist multiple Swadesh lists with different amount of words included in them. During our experiments we utilized the Swadesh list translated for Hungarian[2] which consists of 207 word forms.

Unlike the pseudo-dictionary that we created for learning the linear mapping between the embedding space of the two languages, the quality of the translation pairs is much higher in this case. We should also note that none of the translation pairs originating from the English and Hungarian Swadesh lists are included in the automatically created pseudo-dictionary.

For the 207-element Swadesh list there are 161 words for which Hungarian translations are given in a one-to-one manner, thus we perform our evaluation over these word pairs alone. During this evaluation phase, we calculate for every word pair (s_i, t_i) – located in the filtered Swadesh list – the extent of overlap in the sparse representations of words s_i and t_i. We define precision and recall as $P = \frac{|\phi(s_i) \cap \phi(t_i)|}{|\phi(t_i)|}$ and $R = \frac{|\phi(s_i) \cap \phi(t_i)|}{|\phi(s_i)|}$, respectively. In order to get an aggregated score for quantifying the overlap between the sparse representation of semantically equivalent words in different languages, we take the harmonic mean (F-score) of the precision and recall scores.

Table 1 includes the results quantifying the extent to which source and target level sparse features overlap for mapped word pairs based on the 161-element subset of the Swadesh list. Setting λ to $0.1, 0.3$ and 0.5 resulted in approximately $20, 5$ and 2 non-zero coefficients per word on average. Performances in the 'No mapping' columns are extremely low. This is not surprising, however, since when

[1] https://sites.google.com/site/rmyeid/projects/polyglot
[2] https://hu.wikipedia.org/wiki/Swadesh-lista

no mapping between the source and target language embeddings is performed, overlap between the sparse representation can happen only due to chance.

Table 1 further reveals that the overlap in the sparse representation of the semantically equivalent word forms substantially increase due to any kind of mapping between the languages. The largest improvement can be observed in the case when the mapping matrix M is constrained to be orthonormal and the input word embeddings are made unit length and centered prior to the decomposition phase. As illustrated by the results, the orthogonality constraint mostly help to improve the recall of the mapped sparse word representations. Interestingly, when input word embeddings are left intact (i.e. no preprocessing is performed over them), requiring M to be orthonormal can slightly hurt performance.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.043/0.036/0.039	0.139/0.112/0.124	0.114/0.112/0.113
$\lambda = 0.3$	0.023/0.012/0.016	0.169/0.122/0.141	0.117/0.109/0.113
$\lambda = 0.5$	0.008/0.004/0.006	**0.179/0.135/0.154**	0.138/0.127/0.132

(a) No preprocessing step performed on the input embeddings.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.023/0.024/0.024	0.170/0.117/0.139	0.098/0.137/0.114
$\lambda = 0.3$	0.001/0.001/0.001	0.345/0.118/0.176	0.167/0.208/0.185
$\lambda = 0.5$	0.000/0.000/0.000	0.600/0.009/0.018	**0.271/0.202/0.232**

(b) Input embeddings made unit long and centered as a preprocessing step.

Table 1: The amount of overlap between source and target non-zero coefficients expressed in terms of Precision/Recall/F-score values for the elements of the Swadesh word lists induced dictionary.

3.2 Evaluation related to POS tagging

It has been shown previously that sparse word representations derived from the coefficient matrix can be beneficially utilized in various sequence tagging tasks such as POS tagging [5]. That paper evaluated sparse representations in the monolingual setting, however, here we investigate their applicability in the cross-lingual regime.

For comparability with prior work, we used the very same settings for training the sequence labeler, i.e. we generated features for the word forms by simply relying on their sparse features induced by the feature function ϕ. The linear-chain CRF models [15] we use are also trained using the crfsuite library [16] without modifying any of its default hyperparameter settings. Finally, evaluation was performed on the v1.2 Universal Dependencies treebank in terms of accuracy over the 17 element coarse-grained POS tag inventory it introduces.

The CRF models are trained on the English training data and evaluated on the Hungarian test set of the Universal Dependencies dataset without altering

any of the learned weights of the English model. The various CRF models trained on English data alone and evaluated on the English test have accuracies slightly below 0.9.

When judging the quality of applying the cross-lingually comparable sparse word representation, we can compare the evaluation scores achieved on the Hungarian test set to this mono-lingual setting. We should add, however, that performance drop during testing on the Hungarian test set not only originates from the insufficiencies of the cross-lingual representation, but also from the fact that we utilize the state transition features that are optimized to fit English texts. By defining the cross-lingual sparse representations, we are only able to adapt our test data in terms of the state features to the trained model.

Table 2 illustrates that applying the orthonormality constraint for the mapping matrix M improves POS tagging accuracies on the target language independent of the preprocessing step. This is a difference compared to the results of Section 3.1 where the orthonormality constraint only helped for the case when input embeddings were preprocessed prior to performing their decomposition.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.111	0.300	0.315
$\lambda = 0.3$	0.034	0.283	**0.395**
$\lambda = 0.5$	0.096	0.204	0.384

(a) No preprocessing step performed on the input embeddings.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.169	0.191	0.295
$\lambda = 0.3$	0.097	0.048	0.428
$\lambda = 0.5$	0.138	0.011	**0.446**

(b) Input embeddings made unit long and centered as a preprocessing step.

Table 2: Cross-lingual POS tagging accuracy with the pseudo-dictionary used to learn mapping M between languages.

Since this time we are not evaluating the quality of the cross-lingual sparse embeddings relative to the Swadesh lists, it is now possible to use them as a replacement for the noisy pseudo-dictionary for the calculation of the mapping matrix M. For that reason, we repeated our experiments on cross-lingual POS tagging such that the dictionary to construct M gets determined based on the 161-element dictionary induced from the English and Hungarian Swadesh lists. The results of this experiment can be seen in Table 3.

The most important observation to take when comparing Table 2 and Table 3 is that there is no substantial difference in the results whether the noisy or the more reliable word alignment is used to determine the inter-lingual mapping matrix M provided that the orhtonormality constraint is enforced for M. In fact,

there is a slight decrease of performance noticeable for using the more reliable word alignments when preprocessing of the word embeddings is also performed.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.111	0.413	0.414
$\lambda = 0.3$	0.034	0.327	**0.431**
$\lambda = 0.5$	0.096	0.322	0.430

(a) No preprocessing step performed on the input embeddings.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.169	0.372	0.338
$\lambda = 0.3$	0.097	0.310	**0.407**
$\lambda = 0.5$	0.138	0.310	0.356

(b) Input embeddings made unit long and centered as a preprocessing step.

Table 3: Cross-lingual POS tagging accuracy with the Swadesh-lists induced dictionary used to learn mapping M between languages.

For comparative purposes, we trained such CRF-models that make use of dense word embeddings. These results are included in Table 4. Comparing POS accuracies of the models which rely on sparse and dense word representations we can conclude that models relying on sparse features are dominantly better (but at least comparable) to those models which rely on dense word representation. Furthermore, the relative gain observed for mapping dense word embeddings in any way is less pronounced compared to those in the case of sparse word representations.

	No mapping performed	Unconstrained M	Orthonormal M
Pseudo-dictionary	0.156	0.039	0.144
Swadesh-based dictionary	0.156	0.346	0.264

(a) No preprocessing step performed on the input embeddings.

	No mapping performed	Unconstrained M	Orthonormal M
Pseudo-dictionary	0.201	0.010	0.328
Swadesh-based dictionary	0.201	0.294	0.368

(b) Input embeddings made unit long and centered as a preprocessing step.

Table 4: Cross-lingual POS tagging accuracy when relying on dense word representations.

4 Related work

The comprehensive survey in [4] enumerates a variety of approaches to map multiple independently trained (dense) word embeddings into the same embedding space. The approaches in the survey differ significantly in the amount of supervision they assume to have access to. In our work, we aimed at the utilization of as minimal external resources as possible by not relying on any parallel (or comparable) text resources between languages.

To this end, we created a highly noisy pseudo-dictionary similar to [12]. It has been showed that mapping noisily aligned signals close to each other is often worth to be performed by applying an orthonormal mapping [12,9,17,10].

The kind of decomposition we employ in this work differs both from [5] and [6] as here we not only require D to be a member of the convex set of matrices comprising of unit norm column vectors, but also enforce the coefficients in α to be non-negative.

5 Conclusions

In this paper, we introduced our approach for determining sparse word representations which are comparable across languages. We have shown that these cross-lingual sparse word representations can provide a substantially overlapping representation for word pairs with similar meaning of different languages. Our experiments demonstrated that such well behaving sparse cross-lingual representations can be obtained in the absence of parallel data across the languages, i.e. applying an automatically derived noisy dictionary performed on par to the scenario when a more reliable (but smaller) parallel dictionary was used for mapping word representations to the same embedding space.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
2. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Association for Computational Linguistics (2014) 1532–1543
3. Dozat, T., Qi, P., Manning, C.D.: Stanford's graph-based neural dependency parser at the conll 2017 shared task. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics (2017) 20–30

4. Upadhyay, S., Faruqui, M., Dyer, C., Roth, D.: Cross-lingual models of word embeddings: An empirical comparison. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Association for Computational Linguistics (2016) 1661–1670
5. Berend, G.: Sparse coding of neural word embeddings for multilingual sequence labeling. Transactions of the Association for Computational Linguistics **5** (2017) 247–261
6. Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., Smith, N.A.: Sparse overcomplete word vector representations. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, Association for Computational Linguistics (2015) 1491–1500
7. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: Distributed word representations for multilingual nlp. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, Association for Computational Linguistics (2013) 183–192
8. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association of Computational Linguistics **5** (2017) 135–146
9. Xing, C., Wang, D., Liu, C., Lin, Y.: Normalized word embedding and orthogonal transform for bilingual word translation. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, Association for Computational Linguistics (2015) 1006–1011
10. Artetxe, M., Labaka, G., Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2016) 2289–2294
11. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. CoRR **abs/1309.4168** (2013)
12. Smith, S.L., Turban, D.H.P., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: Proceedings of the 5th International Conference on Learning Representations (ICLR 2017). (2017)
13. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09, New York, NY, USA, ACM (2009) 689–696
14. Swadesh, M.: The origin and diversification of language: Edited post mortem by Joel Sherzer. Aldine, Chicago (1971)
15. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289
16. Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007)
17. Artetxe, M., Labaka, G., Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics (2017) 451–462

Gépi fordító rendszerek kombinálása minőségbecslés segítségével

Laki László János[1,2,3], Yang Zijian Győző[1,2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport,
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
1083 Budapest, Práter utca 50/a
[3] MorphoLogic Lokalizáció Kft.
1012 Budapest, Logodi utca 54
e-mail:{laki.laszlo, yang.zijian.gyozo}@itk.ppke.hu

Kivonat A gépi fordítás minőségének megállapítása mind a kutatásban, mind a cégek számára fontos feladat. A cégek esetében egy megbízható minőségi mutató nagyban segítheti az emberi munkát. Gyakran bevett módszer, hogy jobb minőségű fordítást kapjunk, több gépi fordító rendszert használunk a háttérben és egy kiértékelő rendszer segítségével kiválasztjuk a legjobb minőségű fordítást. A hagyományos kiértékelési módszerek referenciafordítással állapítják meg a gépi fordítás minőségét, ami időigényes és nagy költséggel jár. Ezzel szemben, a minőségbecslés módszere referenciafordítás nélkül tud valós időben becsülni, amivel hatékonyan lehet a különböző gépi fordító rendszer kimenetét kombinálni. A kutatásunk során kettő statisztikai és egy neurális gépi fordító rendszer kimeneteit kombináltuk a minőségbecslő módszerével. Bemutatjuk, hogy a minőségbecslő modell pontosabb eredményt ad hagyományos módszereknél, és nem használ referenciafordítást. Kutatásunkban bemutatjuk, hogy a kombinált rendszer kimenete minden esetben magasabb minőséget eredményez, mint a gépi fordító rendszerek külön.

Kulcsszavak: minőségbecslés, statisztikai gépi fordítás, kifejezés alapú statisztikai gépi fordítás, hierarchikus statisztikai gépi fordítás, neurális gépi fordítás

1. Bevezetés

A gépi fordító (MT) rendszerek jelentős változásokon mentek keresztül az elmúlt évtizedekben. Az éppen aktuális technológiákat követve több különböző módszer alkalmazásával törekedtek a kutatók a lehető legjobb minőségű gépi fordítás előállítására, mint például a szabályalapú MT, statisztikai MT, szintaktikaalapú MT, neurálishálózat-alapú MT. Ezen rendszerek eltérő viselkedésük miatt számos eltérő előnnyel és hátránnyal rendelkeznek. A gépi fordítórendszerek kombinációjával lehetőségünk nyílik, hogy a különböző módszerek előnyeit egyesítsük, valamint a rendszerekben rejlő problémákat gyengítsük, így a létrehozott rendszer minőségében meghaladja a kombinációban résztvett rendszerekét [2,12,14,17].

A rendszerek kombinációját két nagy csoportra oszthatjuk az úgynevezett „glass-box”, valamint a „black-box” kombinációkra. A „glass-box” kombináció lényege, hogy a legjobb fordítás kiválasztásához a rendszer figyelembe veszi a felhasznált fordítórendszerek részletes dekódoló információkat is. Ezzel szemben a „black-box” kombináció esetén csak és kizárólag a legjobb fordítási hipotézisek ismeretében születik meg ez a döntés.

A leggyakoribb kombinációs módszerek helyettesítési gráf (confusion network) létrehozásával valósítják meg a rendszerek kimenetének egyesítését. A helyettesítési gráfot általában egy vázmondat köré építik [14] [20], amihez egynyelvű szóösszekötéssel kapcsolják a hipotéziseket. Az így létrehozott helyettesítési gráfalapú dekódoló segítségével választják ki a legvalószínűbb fordítást. Rosti és társai [19], valamint Heafield és társai [7] a TER algoritmust, míg Okita és társai [16] a BLEU metódust választották.

Rosti és társai. [19] a gépi fordítórendszerek kimenetéből hipotézisalapú helyettesítési gráfot (confusion network) építettek. Ez azt jelentette, hogy az egyik rendszer legjobb fordítását használták vázként, amit szószinten kiegészítettek a többi rendszerből származó alternatív fordításokkal. Ebből a hálózatból az általuk létrehozott dekódoló választotta ki a legjobb fordítást. Ezt Heafield et al. [7] úgy fejlesztette tovább, hogy a helyettesítési gráfot nem szószinten, hanem kifejezés szinten építették.

Huang és Papineni [8] egy hierarchikus rendszerkombinációt hoztak létre, ahol a rendszer igény szerint képes szó, frázis és mondat szintű kombinációra.

Kutatásunk során, egy kifejezésalapú és egy hierarchikus statisztikai gépi fordítórendszer kimeneteit, valamint egy neurális gépi fordító kimeneteit kombináltuk a minőségbecslő módszer segítségével.

A minőségbecslés egy gépi tanulással betanított statisztikai modellen alapul. A fordításokból különböző minőségi mutatószámokat nyerünk, majd a minőségi mutatókat betanítjuk valamilyen (emberi vagy gépi) kiértékeléssel. Az így betanított modellel lehet később új fordításokra predikálni minőséget. A minőségbecslő modell segítségével a különböző fordításokra meg tudjuk becsülni a minőségét. A kutatásunkban különböző nyelvekre alkalmaztunk különböző gépi fordító rendszereket, majd a minőségbecslő modell által jobbnak ítélt fordítást választottuk ki végső kimenetként.

A jelen cikkünkben először kitérünk a kutatásunkhoz használt gépi fordító rendszereket, majd röviden bemutatjuk a minőségbecslés módszerének működését, ezután részletesen kifejtjük a méréseinket és a módszereinket. Végül leírjuk az eredményeinket és levonjuk belőlük a következtetéseinket.

2. A felhasznált gépi fordítórendszerek bemutatása

A minőségbecslés módszerének köszönhetően lehetőségünk nyílt előállítani különböző típusú minőségi mutatókat a különböző gépi fordító rendszerekből. A kutatásunk során, a különböző nyelvekre változóan, egy kifejezés alapú statisztikai gépi fordítót (PBSMT), egy hierarchikus statisztikai gépi fordítót (HBSMT) és egy neurálishálózat-alapú gépi fordítót (NMT) használunk.

A PBSMT [13] rendszer statisztikai megfigyelésen alapszik, amelyek kifejezések összekötéseiből származnak. Az összekötött kifejezéseket egy párhuzamos korpusz segítségével állítjuk elő automatikus statisztikai módszerekkel. A legfőbb oka annak, hogy ezt a módszert a mai napig használják az az, hogy nyelvfüggetlen. Továbbá azoknál a nyelveknél, ahol szintaktikailag, illetve szórendileg hasonló, a rendszer magas pontossággal tud fordítani. A PBSMT továbbá képes lokális átrendezéseket végezni, de a távolabbi átrendezéseket már nem kezeli jól. A HBSMT [3] ezt a problémát igyekszik megoldani.

A HBSMT a PBSMT bővített változata. Amíg a PBSMT kifejezésalapú dekódert használ, addig a HBSMT környezetfüggő nyelvtant használó dekódert. Ez a módszer segít a HBSMT rendszernek megtanulni komlexebb és szintaktikailag távolobb átrendezési szabályokat. Példa: az angol-francia tagadást *don't* $X \rightarrow$ *ne* X *pas* formába menti el, ahol az X helyére bármilyen igei szerkezet behelyettesíthető.

A GPU technológia nagyütemű fejlődésének, valami a megfizethető árú videókártyáknak köszönhetően a mélytanulás-alapú rendszerek gyakorlatilag a legtöbb tudományterület esetén legyőzték teljesítményben az addig létező legjobb rendszereket. Nem volt ez másként a gépi fordítás területén sem. Az NMT [23] rendszerek közvetlenül valósítják meg a feltételes valószínűségi modell alapú szöveg fordítást. A fordítási valószínűségeket leggyakrabban egy „enkóder-dekóder" architektúrájú neurális hálózat segítségével állítjuk elő. [4,9,23]. A módszer előnye az SMT-vel szemben, hogy sokkal folyékonyabb és olvashatóbb fordításokat hoz létre, viszont ezt képes a tartalmi pontosság rovására tenni.

Ha összehasonlítjuk a rendszereket, megfigyelhetjük, hogy a teljesítményük erősen függ az adott nyelvpártól és a korpusz szövegének témájától. Ezért a NMT nem képes minden esetben felülmúlni az SMT rendszereket és a HBSMT rendszer sem képes minden esetben felülmúlni a PBSMT rendszert (lásd. 2. táblázat). Következtetésképpen sokkal robosztusabb gépi fordító rendszert hozhatunk létre azáltal, hogy egy minőségbecslő módszerrel mindig ki tudjuk választani a jobb minőségű kimenetet a különböző rendszerekből.

3. Minőségbecslés

A minőségbecslő módszer (lásd 1. ábra) különböző minőségi mutatókat nyer ki a forrás és gép által lefordított mondatokból, majd a minőségi mutatókkal gépi tanulással betanítjuk a minőségbecslő modellt. A minőségi mutatókat különböző jegyek segítségével nyerjük ki, amelyeket négy csoportra bonthatjuk: minőségi mérőszámokat tudunk kinyerni magából a forrásmondatból, a gép által lefordított mondatból, a forrásmondat és a gépi fordítás viszonyából, valamint a gépi fordító rendszerből. Majd a kinyert minőségi mutatókat betanítjuk valamilyen kiértékelési mérőszámra. Általában a cél az emberi kiértékeléssel való magas korreláció, ezért emberi kiértékelésre szokták betanítani, de sok esetben ennek előállítására nincsen lehetőség, ezért szokás még a hagyományos kiértékelési módszerekre betanítani a rendszert. Az így betanított modell segítségével tudjuk megbecsülni az új ismeretlen mondatok minőségét.

1. ábra. A minőségbecslő modell

Egy másik szempontból közelítve, a jegyeket két nagyobb csoportra osztthatjuk: A „black-box" és a „glass-box" jegyek. Azok a jegyeket, amelyek a gépi fordító rendszerből tudunk kinyerni, „glass-box" jegyeknek hívjuk, mivel ezeknek a jegyeknek a kinyeréséhez ismernünk kell a gépi fordító rendszer belső működését. De az esetek többségében sajnos nincsen hozzáférésünk a rendszerek belső szerkezetéhez. Azokat a jegyeket, amelyek függetlenek a gépi fordító rendszerektől, „black-box" jegyeknek hívjuk. Ezek a jegyek kizárólag a forrás- és a fordított mondatból nyerik ki a minőségi mutatókat.

A kutatásunk során különböző gépi fordító rendszereket alkalmaztunk, ezért csak „black-box" jegyeket használtunk.

A regressziós analízis módszerének köszönhetően a minőségbecslő rendszer a jegyek segítségével képes bármilyen mérőszámra betanítani a becslő modellt. Az ideális eset, hogy emberi kiértékelésre tanítjuk be, de ennek előállítása drága és időigényes, valamint cégek esetében továbbra is a hagyományos módszereket (pl.: TER, BLEU) szokták használni. Ezért a kutatásunk során, mi is hagyományos mértékeket használtuk (pl.: BLEU [18], OrthoBLEU [6], TER [21] stb.) a fordítás minőségének becslésére. A legnagyobb előnye a rendszernek, hogy csak a tanítás műveletében van szüksége referenciára, de miután felépítette a minőségbecslő modellt, a predikció során már nincsen szüksége referenciaértékekre. Ezért a rendszer képes valós időben automatikusan becsülni minőséget, aminek segítségével kombinálni tudjuk különböző gépi fordító rendszerek kimenetét.

4. Módszerek és Mérések

Sajnos még nincsen megbízható és elfogadható jó minőségű minőségbecslő rendszer, ezért a cégek többségében a hagyományos kiértékelési módszereket használják. Az emberi utómunkát végző szakemberek számára a BLEU, TER mértékek a mai napig meghatározó támogatást nyújtanak. Azonban magát a tényleges

BLEU és TER értéket csak a gépi fordítás javítása után határozhatóak meg, de a minőségbecslő segítségével ezt megjósolhatjuk közvetlenül a gépi fordítás elvégzése után is. Sőt ha meg tudjuk mondani a BLEU és TER értékeket, akkor ennek segítségével kombinálni tudjuk a különböző gépi fordító rendszer kimenetét, hogy a jobb minőségű fordítást adjuk tovább az utómunkát végző szakember számára. Ezzel a módszerrel egy jobb minőségű kombinált rendszert kaphatunk.

Kutatásunk első feladataként felépítettük a minőségbecslő modelleket a BLEU, az OrthoBLEU és az OrthoTER (karakteralapú TER) metrikákkal. Majd a betanított minőségbecslő modellek segítségével végeztünk egy tesztanyag szegmensein minőségi predikciót. A modell által predikált értékek alapján választottuk ki a jobb minőségű fordítást.

4.1. Kutatáshoz használt korpuszok

A kutatásunkhoz négy nyelvpárt alkalmaztunk. A forrásnyelv angol, a célnyelv magyar, német, olasz és japán volt. A különböző nyelvpárok különböző témájúak is voltak: autóipari dokumentum, európai parlamenti szövegek, IT-s dokumentumok és ipari termékleírások. Az IT és az ipari szövegek rövid szegmensekből álltak, míg a jogi szövegek szegmensenként átlagosan több mint 20 szóból álltak. Ezek a nyelvek szerkezetileg is igen eltérőek egymástól.

A korpuszok nagy részét egy fordító cég biztosította, ezért ezek nem nyilvánosak. De annak érdekében, hogy a kutatásunk reprodukálható legyen, az egyik korpusz nyilvános volt: The Acquis Communautaire multilingual parallel corpus[4] angol-német nyelven.

A kutatásunkhoz használt korpuszok méreteit az alábbi táblázatban foglaltuk össze (lásd. 1. táblázat). Az egyes szegmensek egyediek, valamint nincsen átfedés a tanító és a teszt anyag között.

1. táblázat. Kutatáshoz használt korpuszok

Nyelvpár	Téma	MT tanítóanyag mérete	QE tanítóanyag mérete	QE tesztanyag mérete
angol - magyar (en-hu)	autóipar	240,000	6,000	1,500
angol - német (en-de)	jog	1,000,000	5,250	1,300
angol - olasz (en-it)	termékleírás	800,000	3,143	785
angol - japán (en-ja)	IT	1,000,000	3,169	790

4.2. Gépi fordító rendszerek beállítása

A statisztikai gépi fordításhoz a Moses [11] keretrendszert használtuk. A tanítóanyagot az előfeldolgozás műveleteinek vetettük alá (pl. tokenizálás, truecase-

[4] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

ing). A szóösszekötéshez a GIZA++ [15] rendszert, míg a nyelvi modell előállításához az IRSTLM [5] eszközt használtuk.

Az neurális fordításhoz az OpenNMT [10] nevű szabadon elérhető keretrendszert használtuk. Az OpenNMT keretrendszerben a gépi fordításhoz szükséges többféle enkóder-dekóder architektúra van implementálva, valamint a szövegek feldolgozásához szükséges programok is elérhetőek. Munkánk során enkódoláshoz egy LSTM-alapú (Long Short-Term Memory) kétirányú RNN architektúrát használtunk, míg dekódolónak egy „attention" modellt [1] alkalmaztunk. Mindegyik rendszerünk 18 epochig tanult, valamint az SGD (Stochastic Gradient Descent) optimalizáció beállítást használtuk.

4.3. Minőségbecslő rendszer beállítása

A kutatásunkhoz automatikus mértékeket használtunk a minőségbecslő modell tanításához: BLEU, OrthoBLEU és OrthoTER mértékek. Ez azt eredményezi, hogy a minőségbecslő modellünk egy 0 és 1 közötti értéket becsül attól függően melyik mértékre tanítottuk be. Az OrthoBLEU (oBLEU) és OrthoTER (oTER) karakteralapú módszer, ami a ragozó nyelvek esetén (pl. magyar) pontosabb becslést ad a szóalapú BLEU módszerhez képest. A BLEU módszer egyik hátránya ilyen esetben, hogy ha egy szó csak toldalékban különbözik, de a szótő megegyezik, két különböző szóként kezeli.

Ahogy a 3. fejezetben írtuk, a minőségbecslő modell építéséhez jegyekre van szükségünk. Ehhez a feladathoz a QuEst [22] keretrendszert használtuk és kizárólag nyelvfüggetlen „black box" jegyeket.

A kutatásunkhoz 67 jegyet alkalmaztunk, melyeket Specia és társai [22] fejlesztettek. Ezek a jegyek tartalomhűségre és nyelvehelyességre vonatkozó jegyeket is tartalmaz (pl. tokenek száma a forrás és a célnyelvi mondatban, célnyelvi mondat perplexitása stb.).

A minőségbecslő modell tanításához több különböző módszerrel kísérleteztünk: többféle regressziós módszer, szupport vektor gép, döntési fák és szabályok, Gauss eljárás stb. A Gauss eljárás (Gaussian process - GP) RBF kernellel és a szupport vektor regresszió (SMOreg) RBF kernellel érte el a legjobb eredményt, ezért az eredményeinkben csak a GP és a SMOreg értékeket mutatjuk.

A magyar nyelvre optimalizálást is végeztünk (az eredmények szekcióban *en-hu-opt*-ként hivatkozunk rá). A Yang és társai [25,24] kutatás alapján a 67 jegy mellett további 60 jegyet próbáltunk ki, melyekből 53 saját fejlesztésű jegy. Ezek a jegyek nyelvspecifikus (pl. igék és főnevek aránya a célnyelvi mondatban, igekötők aránya stb.), n-gram (pl. célnyelvi mondat perplexitása, célnyelvi mondat nyelvmodell valószínüsége stb.), hiba (pl. ismeretlen szavak aránya a célnyelvi mondatban, XML címkék aránya a mondatban stb.) és szemantikai (pl. WordNet jegyek, szótár jegyek stb.) jegyeket tartalmaznak. Az eredmények fejezetben összehasonlítjuk az optimalizált jegyhalmazzal tanított modellt a nem optimalizált jegyekkel tanított modellel.

4.4. Kompozít fordító rendszer

A kompozit fordító rendszer a különböző gépi fordító rendszerek (PBSMT, HBSMT és NMT) kombinálásával jön létre. A rendszer architektúrája az 2. ábrán látható. A rendszer első körben feldolgozza a forrásnyelvi bemenetet a különböző gépi fordító rendszerek segítségével, majd a minőségbecslő modellek segítségével megbecsüljük a lefordított mondatokat. Végül a rendszer a becsült minőség alapján kiválasztja a legjobb fordítást és ezt adja a rendszer végső kimeneteként. A kutatásunkban angol-olasz és angol-japán nyelvpárokra csak PBSMT és HBSMT rendszereket kombináltunk, míg angol-német és angol-magyar nyelvpárok esetén PBSMT, HBSMT és NMT rendszerek kimeneteit kombináltuk.

2. ábra. Kompozít rendszer architektúrája

5. Eredmények és kiértékelés

A kutatásunk során a tanító és a teszt anyagokat a különböző gépi fordítóval lefordítottuk. Majd ezután a minőségbecslő modellünk a forrás- és a lefordított mondatokból kinyerte a minőségi mutatókat. A minőségi mutatók és az előre betanított minőségbecslő modell segítségével előállítottuk a teszt halmazban lévő fordításokra a minőségi becsléseket. A kutatásunkban 4 nyelvpárt és 3 különböző kiértékelési mértéket használtunk. A 2. táblázatban láthatjuk a BLEU, OrthoBLEU és OrthoTER mértékekre betanított modellekkel becsült értékeket a különböző gépi fordító rendszerek és a kompozit rendszer (CoMT) által lefordított mondatokra. A fent bemutatott kiértékelés során az összes vizsgált esetben

az általunk létrehozott kombinált rendszer jobb eredményt ért el, mint a vizsgált rendszerek önmagukban.

2. táblázat. Kombinált rendszerek kiértékelése

		en-hu	en-hu-opt	en-de	en-it	en-ja
BLEU átlag ↑	PBSMT		0.5156	0.6288	0.7513	0.5945
	HBSMT		0.6157	0.4808	0.6998	0.6044
	NMT		0.6281	0.4364	-	-
	CoMT	**0.6926**	**0.6978**	**0.6662**	**0.7525**	**0.6057**
	max MT		0.7614	0.7330	0.7660	0.6458
oBLEU átlag ↑	PBSMT		0.7381	0.6757	0.8202	0.5361
	HBSMT		0.7679	0.6221	0.7993	0.5536
	NMT		0.7252	0.6751	-	-
	CoMT	**0.7729**	**0.7734**	**0.6855**	**0.8246**	**0.5553**
	max MT		0.8698	0.7509	0.8374	0.5832
oTER átlag ↓	PBSMT		0.2903	0.3574	0.1669	0.4281
	HBSMT		0.2193	0.4170	0.1995	0.4075
	NMT		0.2101	0.2653	-	-
	CoMT	**0.1892**	**0.1871**	**0.2649**	**0.1662**	**0.4055**
	max MT		0.0996	0.2083	0.1542	0.3769

Az eredmények mélyebb vizsgálata során szembetűnt, hogy az NMT rendszer minősége eltért a kiindulási hipotézisünktől. A neurális gépi fordító rendszer bevezetésével azt vártuk, hogy a statisztikai rendszerek érvényüket vesztik. Igaz, hogy a neurális rendszer az esetek többségében átlagosan jobb minőséget eredményez, mint a statisztikai rendszerek, de a mondat szintű vizsgálat során, az esetek egy részében mégis a statisztikai rendszerek fordításai bizonyultak jobbnak. Az, hogy bizonyos mondatokat az SMT, míg másokat az NMT rendszer fordít jobban megerősít minket a fordítórendszerek kombinációjának hasznosságában.

Felvetődik a kérdés, hogy mekkora potenciál van a rendszerek kombinációjában. Ennek kiderítére megvizsgáltuk az ideális becslő rendszer minőségét olymódon, hogy a teszthalmazon kiválasztottuk az adott metrika alapján a legjobb fordítási javaslatot, úgy mintha a minőségbecslő modellünk tökéletesen osztályozna. Az így létrehozott ideális rendszer eredményei a 2. táblázatban a max MT sorban olvashatóak. Érdemes megfigyelni, hogy a max MT rendszer minősége jelentősen jobb, mint az alrendszereké külön külön. Pédául mind a szóalapú, mint a karakteralapú BLEU esetén is egyaránt 5-15% javulás figyelhető meg. Ez azt jelenti, hogy viszonylag sok mondat esetén van jelentős eltérés az SMT és az NMT rendszerek fordításai között.

Ennek ellenére a 2. táblázatból az is kiolvasható, hogy a kombinált rendszer eredménye messze elmarad a max MT rendszer eredményeitől, és ehhez képest csak viszonylag kis méretű javulást lehetett kimutatni a legjobb fordító alrendszerhez képest. Ez azt jelenti, hogy még jelentős tartalék rejtőzik az minőségbecslő rendszer becslési pontosságában.

3. táblázat. Angol-magyar modellek optimalizálása

		en-hu	en-hu-opt
Korreláció	PBMT	0.6667	**0.6884**
	HBMT	0.5926	**0.6199**
	NMT	0.5926	**0.6199**
MAE	PBMT	0.1809	**0.1730**
	HBMT	0.1953	**0.1888**
	NMT	0.1953	**0.1888**
RMSE	PB	0.2266	**0.2196**
	HB	0.2402	**0.2341**
	NMT	0.2402	**0.2341**

Végül a 2. és a 3. táblázatokban bemutattuk, hogy az angol-magyar nyelvre az általunk fejlesztett jegyekkel optimalizálva további eredményjavulást tudtunk elérni. Az angol-magyar modellek összehasonlítására és kiértékeléséhez a MAE (mean absolute error - átlagos abszolút eltérés), az RMSE (root mean square error - átlagos négyzetes eltérés gyöke) és a Pearson-féle korreláció mértékeket használtuk.

6. Összegzés

A kutatásunk során egy kompozit gépi fordító rendszert hoztunk létre, amely több gépi fordító rendszer kimenetét kombinálva rendszerszinten jobb eredményt ért el, mint a kombináláshoz felhasznált gépi fordító rendszerek önmgukban. A rendszerek kifejezésalapú statisztikai, hierarchikus statisztikai és neurális gépi fordítókat használnak, de képesek több gépi fordító rendszert is kombinálni. A fordítások kiválasztásához a mondatszintű minőségbecslés módszerével végzett predikciót használtuk. A minőségbecsléshez csak „black-box” jegyeket használtunk. A kompozit rendszereket 4 különböző nyelvpáron teszteltük. Az eredmények alapján rendszerszinten a kompozit gépi fordító rendszerünk minden esetben jobb minőséget adott, mint a PBSMT, a HBSMT és a NMT rendszerek önmagukban. Angol-magyar nyelvpár esetében nyelvfüggő jegyekkel tovább növeltük a rendszer minőségét.

7. Köszönetnyilvánítás

Köszönetünket fejezzük ki a MorphoLogic Lokalizáció Kft. támogatásáért, hogy lehetővé tette számunkra a korpusz használatát.

Hivatkozások

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473 (2014), http://arxiv.org/abs/1409.0473

2. Bangalore, S., Bordel, G., Riccardi, G.: Computing consensus translation from multiple machine translation systems. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 350–354. Madonna di Campiglio, Italy (2001)
3. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 263–270. ACL '05, Association for Computational Linguistics, Stroudsburg, PA, USA (2005)
4. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR abs/1406.1078 (2014), http://arxiv.org/abs/1406.1078
5. Federico, M., Bertoldi, N., Cettolo, M.: IRSTLM: an open source toolkit for handling large scale language models. In: INTERSPEECH. pp. 1618–1621 (2008)
6. FTSK: Orthobleu – mt evalution based on orthographic similarities @ONLINE (May 2014)
7. Heafield, K., Hanneman, G., Lavie, A.: Machine translation system combination with flexible word ordering. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. pp. 56–60. Association for Computational Linguistics, Athens, Greece (March 2009)
8. Huang, F., Papineni, K.: Hierarchical system combination for machine translation. In: EMNLP-CoNLL. pp. 277–286 (2007)
9. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. Association for Computational Linguistics, Seattle (October 2013)
10. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: Open-Source Toolkit for Neural Machine Translation. ArXiv e-prints
11. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL. pp. 177–180 (2007)
12. Kumar, S., Byrne, W.: Minimum bayes-risk decoding for statistical machine translation. In: Susan Dumais, D.M., Roukos, S. (eds.) HLT-NAACL 2004: Main Proceedings. pp. 169–176. Association for Computational Linguistics, Boston, Massachusetts, USA (2004)
13. Lopez, A.: Statistical machine translation. ACM Comput. Surv. 40(3), 8:1–8:49 (Aug 2008)
14. Matusov, E., Ueffing, N., Ney, H.: Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In: McCarthy, D., Wintner, S. (eds.) EACL. The Association for Computer Linguistics (2006)
15. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
16. Okita, T., van Genabith, J.: Minimum Bayes Risk Decoding with Enlarged Hypothesis Space in System Combination, pp. 40–51. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
17. Okita, T., Rubino, R., Genabith, J.v.: Sentence-level quality estimation for mt system combination. In: Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT. pp. 55–64. The COLING 2012 Organizing Committee, Mumbai, India (December 2012)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)

19. Rosti, A.V., Zhang, B., Matsoukas, S., Schwartz, R.: Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In: Proceedings of the Third Workshop on Statistical Machine Translation. pp. 183–186. Association for Computational Linguistics, Columbus, Ohio (June 2008)
20. Sim, K.C., Byrne, W.J., Gales, M.J.F., Sahbi, H., Woodland, P.C.: Consensus network decoding for statistical machine translation system combination. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. vol. 4, pp. IV–105–IV–108 (April 2007)
21. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: In Proceedings of Association for Machine Translation in the Americas. pp. 223–231 (2006)
22. Specia, L., Shah, K., de Souza, J.G., Cohn, T.: Quest - a translation quality estimation framework. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 79–84. Sofia, Bulgaria (2013)
23. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. CoRR abs/1409.3215 (2014), http://arxiv.org/abs/1409.3215
24. Yang, Z.G., Laki, L.J.: Minőségbecslő rendszer egynyelvű természetes nyelvi elemzőhöz. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 37–49. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged, Hungary (2017)
25. Yang, Z.G., Laki, L.J., Siklósi, B.: Quality estimation for english-hungarian with optimized semantic features. In: Computational Linguistics and Intelligent Text Processing. Konya, Turkey (2016)

VI. Poszterek

Evaluation of Universal Dependency parsers for Hungarian

Evelin Ács[1], Gábor Recski[2]

[1] Department of Theoretical Linguistics
Eötvös Loránd University
[2] Department of Automation and Applied Informatics
Budapest University of Technology and Economics

Abstract. We present the results of manual error analysis performed on sample output of three recent dependency parser systems that achieved top scores at the CoNLL 2017 Shared Task: *Multilingual Parsing From Raw Text To Universal Dependencies*. We separate error classes that are sensitive to the technicalities of UD annotation and evaluation principles, such as the treatment of punctuation and coordination, from core issues of syntactic analysis, such as identifying the main predicate in a sentence and resolving structural ambiguity.

1 Introduction

This paper evaluates the performance of three state-of-the-art neural dependency parsers on Hungarian data. The 2017 CoNLL Shared Task on multilingual dependency parsing [1] provided an opportunity to test multiple language-independent parsers on data released as part of the Universal Dependencies v2.0 corpus. Quantitative analysis has been performed as part of the shared task, here we perform a qualitative analysis on a small subset of the Hungarian test data to identify the most important causes of parser errors. We briefly describe the Universal Dependencies project, the shared task, and some related work on the parsing of Hungarian in Section 2. Section 3 presents the results of our manual evaluation and Section 4 draws some conclusions.

2 Background

2.1 Universal Dependencies

Dependency parsing has recently become one of the most investigated technologies in natural language processing, dependency structures are among the most common types of syntactic representations used by downstream NLP applications. In an effort to develop a cross-linguistically consistent annotation system, the Universal Dependencies (UD) project[1] brought together hundreds

[1] http://universaldependencies.org/

of researchers and has resulted in the publication of over 100 UD treebanks in 60 languages (as of version 2.1, released in November 2017) [2].

As a response to the heightened interest in UD and dependency parsing, the 2017 edition of the Conference on Natural Language Learing (CoNLL) organized a shared task on *"Multilingual Parsing from Raw Text to Universal Dependencies"* [1]. The training data – based on version 2.0 of the Universal Dependency dataset – consisted of 64 treebanks for 45 languages. Test treebanks contained at least 10,000 words for each language in the training set and an additional 4 surprise languages.

33 research groups submitted solutions to the task, their systems were ranked based on the macro-average of labeled attachment F-scores (LAS) achieved on each language. LAS matches require that a dependency is assigned to the correct pair of tokens in a sentence and with the correct label. In contrast, unlabeled attachment score (UAS) is more lenient in that it disregards edge labels. A third metric commonly used to evaluate dependency parsers is Content-word Labeled Attachment Score (CLAS), which only considers relation between content words and not function words or punctuation.

In Section 3 we shall analyze errors made by the top three parsers in the competition. The Stanford [3] and C2L2 (Cornell) [4] teams submitted neural parsers that use LSTMs for representing input sentences; both of these systems leverage character-level representations to handle languages with rich morphologies. The Stuttgart IMS team's solution [5] uses CRFs for POS/morphological tagging and a neural tagger for predicting supertags. Overall scores and scores for Hungarian data achieved by each of these three systems is presented in Table 1. Note that the gap between these three systems and the next teams is quite large so that Stanford, C2L2, and IMS are the top three systems based on any of the metrics presented here, and in particular for the Hungarian data.

	Overall			Hungarian		
	LAS	CLAS	UAS	LAS	CLAS	UAS
UnstableParser (Stanford)	76.30	72.57	81.30	77.56	76.08	82.35
C2L2 (Cornell)	75.00	70.90	80.32	76.55	74.36	82.07
IMS (Stuttgart)	74.42	70.18	79.90	73.55	70.87	79.90

Table 1. LAS, CLAS and UAS scores of all three parsers

2.2 Dependency parsing of Hungarian

The Hungarian section of the Universal Dependencies dataset has been created using the Szeged Dependency Treebank [6], challenges of the conversion process are described in [7]. A manual error analysis similar to ours has been performed on Hungarian data before: [8] inspects 200 sentences from the output of Bohnet's parser [9] trained on the Szeged Dependency Treebank. A meaningful comparison of our analysis and theirs is not possible due to the differences between the two tasks: most error classes are specific to the respective annotation systems.

3 Evaluation

We inspected manually the analyses given by each of the three parsers on the first 50 sentences of the Hungarian test data. We grouped errors both by the types of dependency relations they involved and by the types of errors, i.e. the way in which the parsers misinterpreted the structure of a phrase, a clause, or an entire sentence. The number of erroneous edges in each output is similar in all three outputs: the Stanford data contained 208, C2L2 245, and IMS 261. Table 2 lists the top errors by edge type.

UnstableParser		C2L2		IMS	
punct	43	punct	46	punct	44
cc	13	cc	17	cc	17
det	11	det	16	det	11
advmod	9	conj	6	advmod	11
amod	7	conj-nmod	5	amod	7
conj	7	cc-advmod	5	amod-conj	6

Table 2. Types of erroneous edges

As we shall also see when grouping errors by their possible cause, punctuation is the single largest error class for each of the three systems. It has been questioned whether edges in a dependency graph that connect punctuation symbols to some word in the sentence are relevant to dependency structure, in fact the UD community is currently experimenting with the CLAS score as a means to disregard these edges when evaluating dependency parsers [1, p.7]. The `cc` relation is also ignored by CLAS scoring: it is responsible for connecting conjuncts such as *és* ('and'), *de* ('but'), etc. to some other word in the sentence.

3.1 Error types

We shall now describe the most common classes of errors, based on a close observation of each misinterpreted sentence. Besides punctuation and conjuncts we shall discuss 4 additional problem classes that are each responsible for between 2 and 7% of all observed errors (see Table 3 for counts).

Root elements In nearly a fifth of all sentences observed, parsers assigned the `root` dependency to the wrong word, i.e. they failed to identify the main predicate of the sentence. These errors are worthy of attention not only because of their frequency but because they are usually responsible for several further erroneous edges – if the parser misses the main predicate, it is likely to miss relations of each of its dependents. An example of this phenomenon is shown in Figures 1 and 2, which show the gold and erroneous dependency analyses of the sentence in (1).

	UnstableParser	C2L2	IMS
punct, cc	59	63	61
root	9 (15)	9 (17)	10 (19)
conj	9	9	7
modifier POS	8	5	13
structural ambiguity	6 (8)	6 (8)	5 (7)

Table 3. Number of occurrences of each error type (number of edges affected, if different)

(1) – *Azért* *nem* *lehetett* *olyan* *rossz* *közelről* *élvezni* *a*
– Because not be-CAN-PAST so bad near-DEL enjoy-INF the
nehézsúly *Lewis-Holyfield-csúcsrangadóját!*
heavy-weight Lewis-Holyfield-faceoff-ACC!

It can't have been that bad, enjoying the Lewis-Holyfield faceoff from so close!

Fig. 1: Gold analysis of (1)

Coordination Another group of errors involves coordinating conjunctions. In UD, conjunctions are treated asymmetrically: one of the coordinated elements is considered the head of the conjunction and others are connected only to this element (via the `conj` relation) but not to any other word in the sentence. Parser

Fig. 2: Incorrect analysis of (1) by the IMS parser

errors occur when these non-head elements of a conjunction are also connected to other words. These erroneous relations can be justified, since they reflect dependencies that actually hold between some word and *each* element of a coordinated structure – nevertheless this treatment goes against UD conventions. An example is shown in Figure 4, a partial analysis of the sentence in (2).

(2) *Ezek is leginkább csak a januári napokban, amikorra*
these too mainly just the January-ATT day-PL-INE, when-SUBL
a fa már kiszáradt és egy csillagszóró is lángba
the tree already dry-out-PAST and a sparkler too flame-INE
boríthatja – mondta az alezredes.
cover-DEF – say-PAST the colonel.

But only in the days of January, when the tree is dry and a sparkler might burn it down – said the colonel.

Modifiers The UD relations `nmod` and `amod` represent the dependencies between a noun and its nominal or adjectival modifier, respectively. Similarly, the `advmod` relation connects adverbs to predicates or modifiers. A large portion of errors were caused by parsers mixing the above three labels on edges that were otherwise correctly identified, i.e. they connected the modifiers to the right word. Since the distinction between `nmod`, `amod`, and `advmod` is based entirely on the part-of-speech (POS) categories of dependents, one may expect that each of these errors are direct results of POS-tagging mistakes. In fact, out of 26 such

Fig. 3: Partial analysis of (2)

Fig. 4: Partial analysis of (2) by both the C2L2 and the Stanford parser

errors in the three datasets (Stanford: 8, C2L2: 5, IMS: 13), only 14 (4, 2, 8) are in line with the above assumption: the output contains an incorrect POS-tag for the modifier word and the dependency label reflects the same mistake (an example is shown in Table 4). In the remaining 12 cases dependency labels were assigned incorrectly despite a correct POS-tag. In 4 cases, however, 2 made by the Stanford system and 2 by IMS, one may argue that the incorrect dependency labels are actually justified, while gold labels are a result of annotators' compliance with gold POS-tags that are linguistically questionable. An example is shown in Table 5.

	az	*Y2K*	*problémát*
	the	Y2K	problem-ACC
gold POS	DET	NOUN	NOUN
gold dependency	det	nmod	
IMS POS	DET	ADJ	NOUN
IMS dependency	det	amod	

Table 4. Gold and IMS analyses of a noun phrase

Structural ambiguity The final error group involves sentences that are structurally ambiguous and whose parses are consistent with a different constituent structure than the one reflected by the gold dependency annotation. The ambiguous phrase of one such sentence is shown in (3), with English paraphrases for both possible readings. The two dependency structures are shown in Figure 5 and Figure 6.

	türelmetlenül	újra	tárcsáz
	impatient-ESS	again	dial
	'dials again impatiently'		
gold POS	ADJ	ADV	VERB
gold dependency	amod	advmod	
IMS POS	ADJ	ADV	VERB
IMS dependency	advmod	advmod	

Table 5. Gold and IMS analyses of a noun phrase

(3) *a Péterfy kórház sürgősségi belgyógyászati és klinikai toxikológiai osztálya*
the Péterfy hospital emergency internal-medicine and clinical toxicology department-POSS

The department of emergency internal medicine and clinical toxicology
The emergency department of internal medicine and clinical toxicology

Fig. 5: Gold analysis of (3).

3.2 Comparison

[8] includes an error analysis on the output of Bohnet's parser trained on Hungarian data from the Szeged Dependency Treebank. Their method is similar to ours and involves manual inspection of 200 parser errors on the news section of the Szeged dataset.

4 Conclusion

We have presented the results of manual error analysis of three dependency parsers on a small sample of Hungarian data. We have identified several error

Fig. 6: UnstableParser's analysis of (3).

classes that are in some ways technical: those concerning punctuations and conjuncts have little relevance to the dependency structure of content words and underline the necessity of alternative evaluation metrics like CLAS, while those involving coordinating conjunctions introduce edges that may be justifiable and might challenge UD's current treatment of coordination. Modifier relations have brought to light errors in POS-tagging and some possible inconsistencies in the gold standard data. Finally, we have seen examples of structural ambiguity, which remains one of the most challenging problems in syntactic analysis.

References

1. Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C.D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H.F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., Li, J.: CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, Association for Computational Linguistics (2017) 1–19
2. Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M.J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Bauer, J., Bengoetxea, K., Bhat, R.A., Bick, E., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Burchardt, A., Candito, M., Caron, G., Cebiroğlu Eryiğit, G., Celano, G.G.A., Cetin, S., Chalub, F., Choi, J., Cinková, S., Çöltekin, Ç., Connor, M., Davidson, E., de Marneffe, M.C., de Paiva, V., Diaz de Ilarraza, A., Dirix, P., Dobrovoljc, K., Dozat, T.,

Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Erjavec, T., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hajič jr., J., Hà My, L., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Ion, R., Irimia, E., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Laippala, V., Lambertino, L., Lando, T., Lee, J., Lê Hong, P., Lenci, A., Lertpradit, S., Leung, H., Li, C.Y., Li, J., Li, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Moskalevskyi, B., Muischnek, K., Müürisep, K., Nainwani, P., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguyen Thi, L., Nguyen Thi Minh, H., Nikolaev, V., Nurmi, H., Ojala, S., Osenova, P., Östling, R., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Popel, M., Pretkalnia, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Ramasamy, L., Rama, T., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rinaldi, L., Rituma, L., Romanenko, M., Rosa, R., Rovati, D., Sagot, B., Saleh, S., Samardžić, T., Sanguinetti, M., Saulīte, B., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Washington, J.N., Wirén, M., Wong, T.s., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhu, H.: Universal dependencies 2.1 (2017) LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

3. Dozat, T., Qi, P., Manning, C.D.: Stanford's graph-based neural dependency parser at the conll 2017 shared task. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, Association for Computational Linguistics (2017) 20–30
4. Shi, T., Wu, F.G., Chen, X., Cheng, Y.: Combining global models for parsing universal dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, Association for Computational Linguistics (2017) 31–39
5. Björkelund, A., Falenska, A., Yu, X., Kuhn, J.: IMS at the CoNLL 2017 UD Shared Task: CRFs and Perceptrons Meet Neural Networks. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, Association for Computational Linguistics (2017) 40–51
6. Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., Csirik, J.: Hungarian dependency treebank. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA) (2010)
7. Vincze, V., Simkó, K.I., Szántó, Z., Farkas, R.: Universal dependencies and morphology for hungarian-and on the price of universality, Association for Computational Linguistics (2017)

8. Farkas, R., Vincze, V., Schmid, H.: Dependency parsing of hungarian: Baseline results and challenges. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2012) 55–65
9. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, Coling 2010 Organizing Committee (2010) 89–97

Így írtok ti

Nem sztenderd szövegek hibatípusainak detektálása gépi tanulásos módszerrel

Dömötör Andrea[1,2] , Yang Zijian Győző[1,3]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport,
[2] Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar,
[3] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
e-mail:{domotor.andrea, yang.zijian.gyozo}@itk.ppke.hu

Kivonat A szövegek automatikus minőségbecslése nem csak a gépi fordítások kiértékelésénél, hanem egynyelvű szövegek esetén is fontos feladat lehet, hiszen ezek egyrészt bemenetét képezik a különböző természetesnyelvi feldolgozó rendszereknek, másrészt korpusznyelvészeti kutatások nyersanyagául is szolgálnak. A cikkben egy olyan gépi tanuláson alapuló minőségbecslő rendszert mutatunk be, amely kifejezetten egynyelvű, emberek által létrehozott szövegekhez készült. Az eredményeink a rendszer hatékonyságának mérésén kívül arról is érdekes tanulságokkal szolgálnak, hogy mik az internetes informális szövegek nyelvi jellemzői, és hogy mennyiben térnek el az emberi szövegek minőségi problémái a gépi fordítók által generált tipikus hibáktól.

Kulcsszavak: minőségbecslés, korpusznyelvészet, természetesnyelvi elemzés

1. Bevezetés

A szövegek minőségbecslésének igénye eddig elsősorban a gépi fordító rendszereknél merült fel. A fordítóprogram által adott output minőségének becsült pontszáma hasznos információ a felhasználónak, és ennek automatikus mérésére több elterjedt módszer is létezik. A minőségbecslés azonban nem csak a gépi fordítók teljesítményének értékelésével kapcsolatban lehet releváns, hanem egynyelvű, emberek által létrehozott szövegek esetén is.

Az egyre nagyobb számban rendelkezésre álló korpuszokból nagy mennyiségben lehet hozzájutni nyersanyaghoz különböző nyelvtechnológiai, alkalmazott vagy akár elméleti nyelvészeti kutatásokhoz. A kapott szövegek azonban sok esetben nem követik a sztenderdet, és ez megnehezítheti a kutatók dolgát. Éppen ezért szükség van egy olyan eszközre, amelyik - a gépi fordítások minőségbecslő rendszereihez hasonlóan - információt tud adni az adott szövegek minőségéről és arról, hogy a feldolgozásuk előtt milyen esetleges normalizáló lépésekre lehet szükség. Ehhez nem elégséges egy meglévő, gépi fordítónak készült minőségbecslőt az adott nyelvre betanítani, hiszen az emberek által írt szövegek hibái egészen más jellegűek, mint amiket egy gépi fordító követ el. Ebben a cikkben egy olyan

gépi tanulásos rendszert mutatunk be, amely mind a tanulókorpusz, mind a jegyek tekintetében kifejezetten az egynyelvű, emberek által létrehozott szövegekre lett adaptálva.

Reményeink szerint a rendszer hasznos eszköz lesz a különböző természetesnyelv-feldolgozó alkalmazások számára azzal, hogy előzetes információt ad az input minőségéről, így például segíthet eldönteni egy automatikus mondatelemzőnek, hogy „megküzdjön"-e a mondattal, vagy jelezzen hibát. Emellett bízunk abban, hogy az egynyelvű szövegekre „szakosodott" minőségbecslő jó segédeszköz lehet a korpuszokat használó nyelvészeknek is.

2. Kapcsolódó munkák

A hagyományos minőségbecslő módszer (lásd 1. ábra) különböző minőségi mutatókat nyer ki a forrásnyelvi és a gép által lefordított mondatokból, majd gépi tanulással betanítja azokat az emberek által kiértékelt mutatókra. Az így létrehozott modell segítségével tudja megbecsülni az új, ismeretlen mondatok minőségét. Mivel a gépi tanulás modellje emberi kiértékelésen alapszik, ezért a becsült értékek magasan korrelálnak az emberi minőségítéletekkel.

1. ábra. A minőségbecslő modell

A QuEst++ [8] rendszer szó szintű elemző része tartalmaz egynyelvű kiértékeléseket, többek között nyelvimodell-jegyeket, szintaktikai jegyeket, célnyelvi kontextusjegyeket stb. De ezek csupán egy apró részét képezik a teljes rendszernek, amely nem kifejezetten egynyelvű szövegek minőségbecslésének céljával készült.

2. ábra. A πRate rendszer

3. πRate minőségbecslő rendszer

A πRate [9] (lásd 2. ábra) egy egynyelvű minőségbecslő rendszer. A módszere a gépi fordításhoz használt minőségbecslés metodikáját (lásd 1. ábra) veszi alapul. A rendszernek két fő modulja van: a tanuló modul és a kiértékelő modul. A tanuló modul legfőbb feladata, hogy betanítja a minőségbecslő modellt. Ez a modul alapvető működésében megegyezik a hagyományos gépi fordításnál használt minőségbecslő modell tanuló moduljával. A tanításhoz emberek által kiértékelt egynyelvű korpuszt használtunk. A gépi tanuláshoz különböző nyelvi és statisztikai jegyeket vettünk fel: pl.: nyelvi jegyek; n-gram jegyek; fuzzy jegyek; hibajegyek stb. A korpuszból a jegyek segítségével kinyertük a minőségi mutatókat, majd a mutatók segítségével betanítottuk a minőségbecslő modellt az emberek által kiértékelt minőségi mutatókra.

A másik fontos modul a kiértékelő modul. Ez első lépésként beolvassa a bemeneti szöveget, majd az előfeldolgozó fázisban elemzi azt. Az így elemzett szöveget adjuk tovább a jegykinyerő (feature extraction) modulnak, amely a különböző jegyek és a betanított minőségbecslő modell segítségével előállítja a minőségi mutatókat.

A bemenet lehet nyers vagy elemzett szöveg. A πRate rendszer egyik előnye, hogy feladatorientált architektúrával rendelkezik, ezért ha a bemenet már elemzett, akkor a morfológiai elemzés műveletét kihagyhatjuk az előfeldolgozó

fázisban. Így képesek vagyunk optimalizálni az erőforrást és ezáltal a teljesítményt.

A kiértékelő modulnak három fő része van: előfeldolgozó modul, ellenőrző modul és jegykinyerő modul.

A rendszer képes inkrementálisan növekvő bemenetet kezelni. Amint a szöveg beérkezik a πRate rendszerbe, az előfeldolgozó modul morfológiailag elemzi a szöveget (ha az még nem elemzett), kiszámolja az n-gram valószínűségeket stb. Majd az elemzett szöveget továbbküldi a jegykinyerő modul számára, ahol először az ellenőrző modul ellenőrzi le azt a hibajegyek segítségével. Ha a szöveg hibamértéke meghalad egy megadott küszöbértéket, akkor az ellenőrző modul felhatalmazást kaphat megszakítani a folyamatot vagy szűrni, „cenzúrázni" a szöveget. Máskülönben továbbengedi a többi jegy számára, és a minőségbecslő modell a saját hibaértékeit használja fel minőségi mutatóként.

A jegyek kinyerése után a minőségbecslő modell a minőségi mutatók alapján kiszámolja a becsült értékeket (pl.: aktuális mondat minősége, eddig beolvasott összes szöveg globális minősége stb.).

4. Módszerek és mérések

A minőségbecslő modell felépítéséhez egynyelvű jegyekre van szükségünk, amelyek segítségével kinyerjük a minőségi mutatókat. A tanítás során a jegyek egy egynyelvű korpuszból nyerik ki a szükséges értékeket. Majd gépi tanulással emberek által kiértékelt minőségi mutatókra tanítjuk be a modellt (lásd 1. ábra). A πRate rendszer felépítéséhez JAVA EE-t használtunk.

4.1. Egynyelvű korpusz

Az egynyelvű, emberek által létrehozott szövegek hibái más jellegűek, mint azok, amelyeket egy gépi fordító követ el, ezért az egynyelvű szövegek minőségbecslése másfajta tanítóanyagot és jegykészletet igényel. A tanító- és tesztkorpuszt az MNSz2-ből [5] hoztuk létre. Ehhez a beszélt nyelvi és a személyes alkorpuszokból kértünk le random adatot. Azért választottuk ezeket a szövegtípusokat, mert feltételezhetően ezekben a leggyakoribb a sztenderdtől való eltérés. A tanítókorpusz annotálása nyelvészeti alapokon, kézzel történt. Kétféle annotációt használtunk: Likert-skálát (1-5) és osztályozási modellt. A Likert-pontszámok esetén nem a szubjektív emberi értékelést vettük figyelembe, hanem azt próbáltuk pontozni, hogy a mondat elemzése várhatóan mennyire okozhat nehézséget (a normától való eltérésből adódóan) egy szabály alapú gépi eszköz számára. A pontszámokat a helyesen elemezhető összetevők és összetevős szerkezetek arányából számoltuk ki. Az elemzendő összetevők alatt az NP-ket és névutós szerkezeteket, az igék, igekötők és vonzatok kapcsolatát, illetve a tagmondatokat és végül a teljes mondatot értjük. Az (1a)-ban és (1b)-ben látható példákban []-lel jelöljük a felismerhető, és []!-lel a nem felismerhető összetevőket.

(1) a. *[Emberünk ugyanis [állateledel és kisállat kereskedést [tart fenn]]! .]!*

b. *[Ugyanis [törvénysértést [követett el]] [az erkölcsrendész ismerosöm szerint]!]!*

Az osztályozás során a hibátlan mondatok kaptak 5 pontot, a 20%-nál kevesebb felismerhetetlen összetevőt tartalmazók 4-et, a 20 és 39% közötti hibaarányúak 3-at, a 40 és 59% közöttiek 2-t, a legrosszabb 1-es pontszámot pedig azok a mondatok kapták, ahol az összetevők legalább 60%-a minősült elemezhetetlennek. Ez az annotálási rendszer bonyolultnak és indokolatlanul időigényesnek tűnhet, de például (1a)-ban láthatjuk, hogy az *állateledel- és kisállatkereskedés* rossz helyesírásán az emberi értelmezés ugyan könnyen túllendül, egy számítógépes elemző számára azonban gyakorlatilag értelmezhetetlen (vagy félreérthető) így a mondat. Az értékelés tehát azért nem pusztán emberi ítélettel történt, mert a pontszámokkal elsősorban gépi eszközöket szeretnénk informálni, így a pontozásnál ki kellett iktatni az ember „természetes normalizáló képességét".

Az így kapott érték információt adhat az elemző rendszernek az input megbízhatóságáról. Ugyanakkor a hiba típusa még sokkal informatívabb lehet egy gépi eszköz számára: ha a minőségbecslő megbízhatóan detektálja a hiba jellegét, az eszköz alkalmazni tudja a megfelelő normalizáló modult (például: helyesírásellenőrző, ékezet-visszaállító). Az osztályozási modellünk öt hibatípust tartalmaz, amelyek a beszélt nyelvi és az informális internetes szövegekre jellemzők. Ezek a következők:

1. Központozás hibái (hiánya), nagybetűk elhagyása
2. Elírások, helyesírási és nyelvi hibák
3. Idegen nyelvű, idegen szavakat tartalmazó szövegek
4. Ékezetek hiánya
5. Nehezen elemezhető beszélt nyelvi vagy informális szövegek (ismétlések, elakadások, szleng, rövidítések, emotikonok stb.)

A fenti osztályokon kívül meghatároztunk még egy szegmentálásihiba-osztályt is, ide azokat az eseteket soroltuk, amikor a korpuszból kapott adat valójában nem volt mondat, vagy nem egy mondat volt. Továbbá a hibaosztályok mellett természetesen szerepelt egy osztály a hibátlan mondatok számára is.

Ha egy mondat több hibatípusba is besorolható, az egycímkés tanításhoz kiválasztottunk egy fő hibát, azt, amelyik a szöveg nagyobb részét lefedi. Például (2a)-ban elírás is szerepel (*gyeppet*), de jobban jellemzi a szöveget az ékezetek hiánya. (2b)-ben pedig, bár egy mondatközi írásjel is hiányzik, de jelentősebb a helyesírási hibák előfordulása, hiszen a mondat nyolc szavából három is hibás.

(2) a. *De gyeppet sem siettek el a dolgot, es mindharman a jarda kozepen tanyaztak.*
Ékezetek hiánya

b. *Valamelyik ujságban olvastam hogy állitolag fel akarják újitani.*
Helyesírási hibák

A többcímkés osztályozási modellhez minden mondat 3 címkét kapott, ennél több hibaosztályba egyik adat sem volt besorolható. Ahol háromnál kevesebb hibatípus fordult elő, ott a hiányzó helyekre a „helyes” címkét kapták a mondatok.

4.2. Egynyelvű jegyek

A minőségbecslő modellünk 36 különböző típusú jegyet használ, amelyeket jellegük alapján az alábbi kategóriákba soroltuk:

- nyelvi jegyek:
  - főnevek, igék, igekötők, melléknevek, határozószók, kötőszók, névmások, névelők, indulatszók aránya a mondatban;
  - főnevek és igék aránya a mondatban;
  - főnevek és melléknevek aránya a mondatban;
  - igék és igekötők aránya a mondatban;
  - főnevek és névelők aránya a mondatban;
  - tokenek száma;
  - átlagos szóhossz a mondatban;
- n-gram jegyek:
  - a mondat nyelvmodell valószínűsége;
  - a mondat nyelvmodell perplexitása;
  - a mondat szótöveinek, szófaji címkéinek nyelvmodell valószínűsége;
  - a mondat szótöveinek, szófaji címkéinek nyelvmodell perplexitása;
- neurális nyelvmodell jegyek:
  - 1-gram, 2-gram és 3-gram perplexitás;
- hibajegyek:
  - ismeretlen szavak aránya a mondatban;
  - ékezetes szavak aránya a mondatban;
  - írásjelek aránya a mondatban.

Az n-gram modell felépítéséhez (az n-gram jegyekhez) szintén az MNSz2 egy részkorpuszát használtuk, amely 98500 lemmatizált és elemzett mondatot tartalmaz.

A neurális nyelvmodell tanításához a Pázmány Korpuszból [2] vettünk 1 millió mondatot. A nyelvmodell felépítéséhez egy GRU-alapú (Gated recurrent unit) RNN architektúrát használtunk, amely 6 epochig tanult. Továbbá a modellünkhöz szóbeágyazást (word embedding) is alkalmaztunk, amelyhez a [4] által készített magyar nyelvű szóbeágyazási modellt használtuk fel.

4.3. Mérések

A kutatásunk során kétféle mérést végeztünk: osztályozást és regressziót. A Likert értékek segítségével regressziós modelleket építettünk, a hibaosztályok segítéségével pedig osztályozási modelleket készítettünk. Az osztályozás esetében végeztünk egycímkés (a fő hibaosztállyal) és többcímkés osztályozást is. A regresszióhoz és az egycímkés osztályozáshoz a WEKA [3] szoftvert, míg a többcímkés osztályozáshoz a MEKA [7] rendszert használtuk.

Többféle tanuló algoritmust is kipróbáltunk, melyek közül az egycímkés osztályozáshoz a szupport vektor gép, a regresszióhoz a szupport vektor regresszió, a többcímkés osztályozáshoz pedig a véletlen erdő (random forest) alapú „Classifier Chains" [6] módszer érte el a legjobb eredményt, ezért az eredmények fejezetben csak az ezekkel a módszerekkel betanított modellek eredményeit mutatjuk. Az annotált tanítóanyag segítségével az alábbi három minőségbecslő modellt építettük:

- LS modell: regressziós minőségbecslő modell a Likert értékeket felhasználva. A Likert értékek 1-től 5-ig terjedő egész számok.
- CS modell: egycímkés osztályozási minőségbecslő modell a fő hibaosztályokat felhasználva. Összesen 6 hibaosztály és a helyes mondat osztályozási címkéje.
- CCS modell: többcímkés minőségbecslő modell az osztályzási értékeket felhasználva. Minden mondathoz 3 db osztályt rendeltünk, az első osztály a főcímke, ami vagy a fő hibaosztály, vagy a helyes mondat címkéje. A második és a harmadik osztály a mellékhibá(ka)t tartalmazza, ha nincs ilyen, a helyes mondat címkéjét viselik.

Továbbá végeztünk optimalizációt is. A gépi fordítás kiértékelésének optimalizációja alapján [1] ha kivesszük a kevésbé releváns jegyeket a rendszerből, kevesebb jeggyel magasabb minőséget tudunk elérni. A „forward selection" [10] módszerével az alábbi optimalizált modelleket hoztuk létre:

- OptLS modell: optimalizált LS modell.
- OptCS modell: optimalizált CS modell.

5. Eredmények és kiértékelések

5.1. A πRate rendszer

A πRate rendszer kiértékeléséhez az MAE (mean absolute error - átlagos abszolút eltérés), az RMSE (root mean square error - átlagos négyzetes eltérés gyöke), a Pearson-féle korreláció, a helyesen osztályozott egyed (Correctly Classified Instances - CCI), a Hamming-veszteség (Hamming loss) és a pontosság (accuracy) mértékeket használtuk. A teszteléshez minden esetben tízszeres keresztvalidálást használtunk. Az 1. és 2. táblázatokban látható, hogy a 36-os jegykészlet ∼77,1%-os korrelációt és ∼64,48% helyesen osztályozott egyedet ért el.

Az optimalizálást a „forward selection" módszerrel végeztük. Az optimalizálás utáni eredményeket a 1. és 2. táblázatok második sorai tartalmazzák.

- Az OptLS jegykészlet, amelyik 15 jegyet használ, nagyjából ugyanolyan korrelációt ért el, mint a teljes jegykészlet, de lényegesen kevesebb munkával.
- Az OptCS jegykészlet, amelyik 28 jegyet használ, nagyjából ugyanolyan eredményt ért el, mint a teljes jegykészlet, de kevesebb munkával.

Amint látható, a Likert-skála modell megfelelő korrelációval működik, ám az egycímkés osztályozási modell kevesebb eredményességet mutat. Ennek magyarázata részben az annotálási módszer is lehet. Amint 4.1-ben említettük, egy

	Korreláció	MAE	RMSE
LS modell - 36 jegy	0.7712	0.7121	1.0047
OptLS készlet - 15 jegy	0.7777	0.7226	0.9625

1. táblázat. Az LS modell és az OptLS jegykészlet értékelése

	CCI	MAE	RMSE
CS modell - 36 jegy	64.48%	0.214	0.3171
OptCS készlet - 28 jegy	65.17%	0.2137	0.3167

2. táblázat. A CS modell és az OptCS jegykészlet értékelése

mondat csak egy hibaosztályba tartozhat az annotáció szerint, holott a valóságban többféle hibát is tartalmazhat. Azaz, az osztályozó feladata valójában az, hogy meghatározza az elsődleges hibatípust, ami igencsak nehéz is lehet, ha a mondatban egyéb, kevésbé releváns hibák is megtalálhatók.

A sikertelenség okainak pontosabb feltérképezésére elkészítettük a méréshez tartozó tévesztési mátrixot (3. táblázat). Ebből látható, hogy a rendszer jól boldogul az ékezet nélküli és a hibátlan szövegek besorolásával, a többi osztálynál viszont gyengébb eredményeket mutat. Különösen figyelemre méltó, hogy az idegen nyelvű és a beszélt nyelvű szövegekre egyszer sem tippelt a program, úgy látszik, ezek jellemezhetők a legkevésbé a rendszer által használt jegyekkel.

	Ékezet	Hí.	Id. nyelv	Beszélt nyelv	Központozás	Szegm.	Helyes	F-score
Ékezet	107	0	0	0	1	0	7	0.915
Helyesírás	2	30	0	0	40	1	72	0.267
Idegen nyelvű	1	0	0	0	1	3	3	0.0
Beszélt nyelvi	0	6	0	0	28	0	19	0.0
Központozás	3	23	0	0	127	1	83	0.547
Szegmentálás	5	10	0	0	14	12	21	0.304
Helyes	1	3	0	0	16	0	385	0.774

3. táblázat. Tévesztési mátrix

5.2. Az optimalizált jegykészlet

A 4. és 5. táblázatokban láthatók az optimalizált jegykészletek, a jegyek relevanciája szerint rendezve.

A Likert-skála szerinti minőségbecslés szempontjából legrelevánsabb nyelvi és hibajegyek az ékezetekkel, a tokenszámmal és az írásjelekkel függnek össze. Ez az eredmény nyelvészeti szempontból nézve nem meglepő. Az informális írásbeli kommunikációban (azaz legjellemzőbben az internetes szövegekben) az ékezetek és írásjelek elhagyása a legtipikusabb sztenderdtől való eltérés. Utóbbi a gépi

Jegy
Ékezetes karakterek száma
Ékezetes karaktert tartalmazó szavak / összes szó
A tokenek n-gram perplexitása (ismeretlen szavakkal együtt)
A szövegben előforduló elemzési címkék n-gram perplexitása (ismeretlen szavakkal együtt)
1-gram perplexitás (neurális nyelvmodell)
A szövegben előforduló elemzési címkék n-gram perplexitása (ismeretlen szavak nélkül)
Ismeretlen szavak száma
A szótövek n-gram perplexitása (ismeretlen szavak nélkül)
A szövegben előforduló szófajcímkék n-gram perplexitása
A szótövek n-gram perplexitása (ismeretlen szavakkal együtt)
Indulatszók aránya
Mondatközi írásjelek száma / mondatvégi írásjelek száma
Tokenek száma
Főnevek száma / igék száma
A szófajcímkék n-gram perplexitása (ismeretlen szavakkal együtt)

4. táblázat. 15 jegyre optimalizált jegykészlet a Likert-modellhez

feldolgozásban szegmentálási problémákat okozhat, ezzel magyarázható a sokszor extrém magas tokenszám. Ha tehát egy mondat (vagy amit az elemző egy mondatnak hisz) nagyon hosszú, a minőségbecslő rendszer gyanakodhat, hogy egy szerkesztetlen szövegről van szó, ami gyakran együtt jár az alacsony minőséggel. Az optimalizált jegykészletben előfordul még az ismeretlen szavak és az indulatszók aránya is. Ezen címkék magas száma szintén az informális szövegek sajátossága.

Az egycímkés osztályozási modell optimalizált jegyei között már több olyan jegy is megjelenik, amely valódi nyelvi hibára utalhat, ilyenek például a főnevek és névelők, az igék és igekötők aránya vagy a névmások száma. Ezek szükségesek ahhoz, hogy a modell a nyelvtani helyességre (helytelenségre) vonatkozó hibaosztályokat is detektálni tudja. Az ilyen típusú jegyek azonban kevésbé tűnnek relevánsnak a Likert-modell esetén, ami azt mutatja, hogy az egynyelvű korpuszokból származó szövegek minőségi problémái nagyrészt nem nyelvi természetűek a szó szoros értelmében véve.

5.3. A hibatípusok összefüggései

A 6. táblázat az egyes hibaosztályok átlagos Likert-pontszámát mutatja. Eszerint a normalizálást tekintve az ékezet-visszaállítás, a nyelvfelismerés és a megfelelő mondatokra és tagmondatokra bontás (praktikusan írásjel-visszaállítás) jelentheti a legnagyobb segítséget egy gépi feldolgozó eszköz számára.

A hibatípusok összefüggéseit főkomponens analízissel vizsgáltuk meg. A kapott eredményeket a 7. táblázat tartalmazza. Ebből az derül ki, hogy a vizsgált

Jegy
Ékezetes karaktert tartalmazó szavak / összes szó
Írásjelek aránya
Ékezetes karakterek száma
Névmások száma
A mondat szótöveinek n-gram valószínűsége
Igék aránya
Főnevek száma / névelők száma
Főnevek száma / igék száma
Tokenek száma
A szövegben előforduló elemzési címkék n-gram valószínűsége
Kötőszavak aránya
írásjelek aránya
Határozószók aránya
A szófajcímkék n-gram valószínűsége
A tokenek n-gram perplexitása (ismeretlen szavakkal együtt)
Átlagos szóhossz a mondatban
A mondat szótöveinek n-gram perplexitása (ismeretlen szavakkal együtt)
Melléknevek aránya
A tokenek n-gram valószínűsége
A szófajcímkék perplexitása (ismeretlen szavakkal együtt)
2-gram perplexitás (neurális nyelvmodell)
Számnevek aránya
Igék száma / igekötők száma
1-gram perplexitás (neurális nyelvmodell)
3-gram perplexitás (neurális nyelvmodell)
Főnevek száma / melléknevek száma
A mondat szótöveinek n-gram perplexitása (ismeretlen szavak nélkül)
Determinánsok aránya

5. táblázat. 28 jegyre optimalizált jegykészlet az egycímkés osztályozási modellhez

	Átlagos pontszám
Ékezetek hiánya	1.16
Idegen nyelvű szövegek	1.63
Szegmentálási hibák	1.74
Elütések, helyesírási és nyelvi hibák	2.69
Írásjelek hibái (hiánya), nagybetűk elhagyása	3.20
Nehezen elemezhető beszélt nyelvi szövegek	3.28

6. táblázat. A hibaosztályok átlagos Likert-pontszáma

informális műfajokban a kevésbé gondos szövegalkotás több jelenségben is megnyilvánulhat egyszerre. Így például számíthatunk arra, hogy az ékezeteket nem tartalmazó szöveg jó eséllyel nyelvi szempontból sem fog megfelelni a sztenderdnek (l. 1. faktor). A szegmentálási hibák az elemzés szerint gyakran az idegen nyelvű szövegekkel járnak együtt (l. 2. faktor), az írásjelek elhagyása viszont úgy tűnik, hogy a többi szempontból kifogástalan szövegekre is nagy arányban jellemző (3. faktor). Mivel az egycímkés osztályozási modell tanuló korpusza jelenleg csak egy (elsődlegesnek tekintett) hibatípust rendel a mondatokhoz, fontos ezeknek az összefüggéseknek az ismerete.

1. faktor	beszélt nyelvi szöveg	nyelvi hibák	ékezetek hiánya
2. faktor	szegmentálási hibák	idegen nyelvű szövegek	
3. faktor	központozási hibák		

7. táblázat. A hibatípusok összefüggései főkomponens analízissel

5.4. Többcímkés osztályozási modell

A többcímkés osztályozási modell (8. táblázat) a fő hibaosztály detektálásában hasonló eredményességet mutat, mint az egycímkés modell. A második és harmadik címke pontosságának javulása annak is köszönhető, hogy ezek egyre nagyobb arányban tartoztak a hibátlan osztályba. Mindemellett az 56.6%-os pontos egyezés, a feladat komplexitását és a tanuló adatok kis számát tekintve, viszonylag jó eredménynek mondható.

	Fő hibaosztály	2. hibaosztály	3. hibaosztály
Pontosság címkénként	0.652	0.835	0.964
Pontos egyezés		0.566	
Hamming veszteség		0.183	

8. táblázat. A többcímkés osztályozási modell eredményei

6. Összegzés

A cikkben egy olyan egynyelvű szövegek minőségbecslésére tervezett rendszer felépítését és működését mutattuk be, amely jól alkalmazható lehet a korpusznyelvészetben vagy a természetesnyelvi elemző rendszerek előfeldolgozó moduljában. Az eredmények azt mutatták, hogy az emberek által létrehozott egynyelvű szövegek, a gépi fordítók által generáltakkal ellentétben, nagyobbrészt nem nyelvtani

hibákat tartalmaznak. Ezen szövegek minőségi problémái sokkal inkább az internetezők írási szokásaiból adódnak, mint például az ékezetek vagy az írásjelek elhagyása.

A rendszer értékeléséből az derült ki, hogy a hibatípusok detektálása még fejlesztésre szorul, egyelőre csak a hibátlan és az ékezet nélküli szövegek elkülönítésében működik megbízhatóan. Az általános minőségi mutató (Likert-pontszám) megbecsülésében azonban jó eredményt értünk el.

További általános észrevétel még, hogy a jegykészlet-optimalizáció nagy jelentőséggel bír. Az eredményeink szerint a jegykészlet csökkentésével javítható a teljesítmény, és az erőforrás-felhasználás is kevesebb lesz. Az egynyelvű szövegek optimális jegykészlete eltért attól, ami a korábbi kutatásokban a fordítások esetén tapasztalható volt, elmondható tehát, hogy az optimális jegykészlet feladatfüggő.

Hivatkozások

1. Beck, D., Shah, K., Cohn, T., Specia, L.: Shef-lite: When less is more for translation quality estimation. In: Proceedings of the Workshop on Machine Translation (WMT) (2013)
2. Endrédy, I., Prószéky, G.: A pázmány korpusz. Nyelvtudományi Közlemények (112), 191–206 (2016)
3. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. 11(1), 10–18 (Nov 2009)
4. Novák, A., Novák, B.: Magyar szóbeágyazási modellek kézi kiértékelése. In: XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018). Szegedi Tudományegyetem, Szeged, Hungary (2018)
5. Oravecz, C., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, N., et al. (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation. ELRA, Reykjavik, Iceland (may 2014)
6. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Mach. Learn. 85(3), 333–359 (Dec 2011), http://dx.doi.org/10.1007/s10994-011-5256-5
7. Read, J., Reutemann, P., Pfahringer, B., Holmes, G.: MEKA: A multi-label/multi-target extension to Weka. Journal of Machine Learning Research 17(21), 1–5 (2016), http://jmlr.org/papers/v17/12-164.html
8. Specia, L., Paetzold, G., Scarton, C.: Multi-level translation quality prediction with quest++. In: ACL-IJCNLP 2015 System Demonstrations. pp. 115–120. Beijing, China (2015)
9. Yang, Z.G., Laki, L.J.: Minőségbecslő rendszer egynyelvű természetes nyelvi elemzőhöz. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 37–49. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged, Hungary (2017)
10. Yang, Z.G., Laki, L.J., Siklósi, B.: Quality estimation for english-hungarian with optimized semantic features. In: Computational Linguistics and Intelligent Text Processing. Konya, Turkey (2016)

A HuTongue spontán beszélt nyelvi korpusz leiratozásának és annotálásának minőségbiztosítási munkálatai

Gulyás Attila[1], Galántai Júlia[1], Szabó Martina Katalin[1, 2, 3], Szebeni Zea[1]

[1]MTA TK „Lendület" RECENS Kutatócsoport
[2]Szegedi Tudományegyetem, Szláv Intézet, Orosz Filológiai Tanszék
[3]Precognox Informatikai Kft.
{Gulyas.Attila;Galantai.Julia;Szabo.Martina;Szebeni.Zea}@tk.mta.hu

Kivonat: Jelen dolgozatban egy magyar nyelvű, spontán beszélt nyelvi korpusz, a HuTongue leiratozásának és annotálásának minőségbiztosítási munkálatairól számolunk be. A korpuszban feldolgozott szövegeket hétköznapi szituációkban, külső ingerektől teljesen elzárt környezetben keletkezett spontán nyelvi produktumok alkotják. A korpusz létrehozásának legfőbb célja, hogy megfelelő vizsgálati anyagot teremtsünk a pletyka természetének elsősorban társadalomtudományi szempontú kutatásához. A HuTongue egy egyedülálló adatbázis: tudomásunk szerint ez az egyetlen magyar nyelvű, nagy méretű, spontán szituációkban keletkezett, beszélt nyelvi korpusz, amely teljes egészében manuálisan gépelt és annotált formájú. A korpusz létrehozása – amely jelenleg is folyamatban van − több munkafázisban történik. Az előkészítés után a fájlokat egy feldolgozócsapat legépeli és annotáltatja. A munka három alapvető feladatból tevődik össze: a hanganyagon hallható verbális közlések rögzítéséből, a nem verbális hanghatások kódolásából, valamint egy, a kutatás szempontjából kardinális, szemantikai-−pragmatikai jellegű sajátság jelöléséből. Azt reméljük, hogy a korpusz a kutatási kérdés sokrétű és automatikus megoldásokkal hatékonyan támogatott vizsgálatát fogja lehetővé tenni a számunkra a jövőben. A jelen dolgozat célja, hogy bemutassuk e komplex feldolgozási munkának a minőségbiztosítási folyamatát. Szólunk a minőségbiztosítás szempontjairól, megtervezésének dilemmáiról és lépéseiről, valamint bemutatjuk az általunk alkalmazott megoldást.

1. Bevezetés

A magyar nyelvű, spontán beszélt nyelvi korpusz, a HuTongue korpusz létrehozásának legfőbb célja, hogy megfelelő vizsgálati anyagot teremtsünk a pletyka természetének elsősorban társadalomtudományi szempontú, beható vizsgálatához.

Annak ellenére, hogy az elmúlt évtizedekben számos spontán nyelvi korpusz született a világ különböző nyelvein [4], [1], [7], [6], a magyar nyelvre irányuló kutatásunkhoz nem állt rendelkezésünkre megfelelő vizsgálati adatbázis. Mindenekelőtt, a magyar nyelv legtöbb beszélt nyelvi korpusza olvasott szövegekből áll [3]. Jelenleg három hazai spontán beszélt nyelvi korpuszról van tudomásunk, az ún. BEA (Magyar spontán

beszéd adatbázis) [3], a Kivi (Korpusz az inferencialitás vizsgálatához) [5], valamint egy készülőben lévő adatbázisról, a Budapesti Egyetemi Kollégiumi Korpuszról (BEKK, bekk.elte.hu), amely társadalomtudományi céllal készül és nyelvi interakciókat tartalmaz. Ugyanakkor, saját vizsgálati céljainknak közülük egyik adatbázis sem felelt meg részletesebben [2], [8]. Olyan korpuszra volt szükségünk, amely megfelelő ahhoz, hogy spontán nyelvi környezetben keletkezett diskurzusokat vizsgálhassunk, és a pletyka jelenségére vonatkozó megállapításokat tehessünk. A fentebbieket megfontolva döntöttünk egy saját korpusz létrehozása mellett. A korpusz létrehozásának fő célja az volt, hogy mind a szövegek típusa, mind az alkalmazott annotáció tekintetében olyan korpuszt hozzunk létre, amely lehetővé teszi a pletyka jelenségének beható vizsgálatát.

Egy korábbi dolgozatunkban [8] részletesen tárgyaltuk a korpusz szövegeinek keletkezési körülményeit, a feldolgozói munka előkészítő lépéseit, a feldolgozási és az annotálási folyamat eszközét, módszereit, a megtervezés dilemmáit, valamint az alkalmazott megoldásokat. A jelen dolgozatban a korpusz leiratozásának és annotálásának minőségbiztosítási folyamatát kívánjuk ismertetni. Meg szeretnénk mutatni mindazokat a minőségbiztosítást érintő kérdéseket és dilemmákat, amelyekkel a munka során szembesültünk, valamint a feladatban alkalmazott megoldásokat és eszközöket. A tárgyalás során, a problémák szemléltetése céljából a korpuszból származó példákat is segítségül hívunk.

2. A korpusz rövid bemutatása és alapvető statisztikai adatai

A korpuszban feldolgozott szövegeket hétköznapi szituációkban, külső ingerektől teljesen elzárt környezetben keletkezett spontán nyelvi produktumok alkotják. A spontán alatt azt értjük, hogy a résztvevők azzal és arról beszéltek, akivel és amiről akartak. Emellett a beszéd mennyisége sem volt korlátozva. A környezet, amelyben a rögzítés történt, egy összesen három nagyobb társas térre osztható épület volt. A résztvevők átlagos életkora 23 volt, a legfiatalabb 21, a legidősebb pedig 26. Közülük 3 nő és 5 férfi volt. A hangrögzítés napi 24 órában történt úgy, hogy minden résztvevő rögzítő eszközt viselt.[1] A résztvevők tudatában voltak annak, hogy a hangjukat 24 órában rögzítik, és a hanganyag jogtulajdonosa hozzájárult annak kutatási célú felhasználásához.

A teljes anyagból végül nyolc nap felvételét dolgoztuk fel a korpuszban. Ez az anyag összesen megközelítőleg 500 órányi hanganyagot tesz ki.

A szövegek feldolgozásának a megkezdése előtt az anyagot a munkához elő kellett készíteni: a felvételekből a csendet, és az egyéb, nem beszéd tartalmú hangokat tartalmazó részeket a lehető legteljesebb mértékben ki kellett vágnunk, majd az anyagot 60 perces egységekre osztottuk [8].

[1] A spontán nyelvi hanganyagot, amelyet egy szórakoztatóipari cég rögzített, kizárólag tudományos célokra adták át és használjuk fel, teljes titoktartási kötelezettségvállalás mellett. A felvételen résztvevő önkéntesek teljes körű tájékoztatásban részesültek a hangfelvételek elkészüléséről.

3. A feldolgozói munka rövid bemutatása

A feldolgozói munkát a résztvevők kiválasztása, majd betanítása előzte meg. A kiválasztott feldolgozók először egy nagyon részletes ún. útmutatót kaptak, amely pontos leírását adta az elvégzendő feladatnak, valamint a feldolgozáshoz használni kívánt eszköz kezelésének. Az útmutatóban rendre példákat (szöveges és hangzó) is elhelyeztünk. A leírás megismerése után a teljes folyamatot megbeszéltük a feldolgozókkal.

Annak céljából, hogy a munka minőségét a lehető legmagasabb szinten tarthassuk, a feldolgozókkal folyamatos kapcsolatot tartunk, a kérdéseket, észrevételeket megbeszéljük, és a megbeszélteket a teljes csoport számára elérhető formában rögzítjük.

A feldolgozáshoz az f4transkript (https://www.audiotranskription.de/english/f4.htm) elnevezésű szoftvert alkalmaztuk. A szoftvert gyakran alkalmazzák társadalomtudományi kutatások alkalmával, mert a használata nagyban elősegíti a leiratozás és a taggelés gyors és egyszerű, egy időben történő elvégzését. Mivel a jelölés során nem fonetikai annotációt készítettünk, a számunkra kiválóan megfelelt a szemantikai és pragmatikai jellemzők rögzítésére.

A feldolgozók online kapták meg a hangfájlokat, valamint az egyes hangfájlokhoz tartozó szegmensek időbélyegeit - amelyek a kivágott részek jelölésére szolgálnak, és útmutatóként a fájlok összeillesztéséhez, txt formátumban. Ez utóbbi tette lehetővé számunkra azt, hogy a későbbiekben az elkészült, írásos formájú szegmenseket az időbélyegek alapján egymáshoz illesszük, valamint útmutatóként szolgálnak a kivágott szöveganyag jelzésére. A feldolgozók e fájlt töltötték ki a program segítségével a hanganyag írásos rögzítésével. A alábbi ábra egy részletet mutat a program működéséről.

Fig. 1. Részlet az F4-programról

A feldolgozók munkája a jelen fázisban három alapvető feladatból tevődött össze: a hanganyagon hallható verbális közlések rögzítéséből, a nem verbális hanghatások kódolásából, valamint egy, a kutatás szempontjából kardinális, szemantikai--pragmatikai jellegű sajátság jelöléséből. A munka során tehát nem csupán a verbális információk rögzítését céloztuk, hanem olyan, nem verbális információk tagelését is, amelyek véleményünk szerint fontos segítségül szolgálhatnak majd számunkra a korpusz

jövőbeli felhasználása során a kutatási probléma mélyebb összefüggéseinek feltárásában [8]. A nem verbális hanghatások tageinek kiválasztásánál és meghatározásánál fontos szempont volt, hogy a kutatási kérdéseink és hipotéziseink megválaszolásához megfelelő elemzési szempontokat és magyarázó változókat szolgáltassunk. Az annotációs jelölések kiválasztásánál tehát fontos volt az olyan érzelmi töltetű megnyilvánulások jelölése, amelyek korrelációja az általunk jelölt pragmatikai sajátságok meghatározásánál fontos többlettartalmat jelölhet (pl. gunyoros vagy zavarodott nevetés, köhögés stb.)

Általános, a teljes leiratozásra vonatkozó sajátságként mondható el, hogy a feldolgozóknak a szoftver segítségével időbélyegeket kell elhelyezniük a gépelt szövegben, annak megfelelő helyein (részletesebben l. [8]). Ez nem csak a fájlok összeillesztését teszi lehetővé, hanem a későbbiekben időintervallum alapú mérési eszközök kidolgozására ad lehetőséget, és megőrzi a hanganyag kapcsolatát a szöveges fájllal.

(1) #00:00:46-0# (Sanyi) Persze. #00:00:46-4#
#00:00:46-4# (Gabi) Sietünk, hátha #00:00:49-3#

Azt, hogy az adott diskurzus verbális és nem verbális hanghatásai kitől származnak, ugyancsak jelöljük, például

(2) (Gabi) Hol a kávém?

Tekintettel arra, hogy minden egyes résztvevő mikroportjának hanganyagát külön feldolgoztuk, nem szükséges az, hogy az anyagok teljes tartalmát leiratoztassuk, elég csupán azt, ahol a mikroport viselője további szereplőkkel részt vesz egy diskurzusban. Minden diskurzusnál jelöljük azonban, hogy a verbális és előre meghatározott nem verbális hanghatások kitől származnak.

Amennyiben érthetetlen a teljes megszólalás vagy annak egy része, a következő jelet alkalmazzuk a diskurzus megfelelő helyén: (?). Abban az esetben pedig, ha a leiratozó nem biztos benne, hogy jól értette, amit hallott, az adott szövegrészt a következő nyitó- és zárótaggel jelöli meg: (( )) Ha olyan megszólalót hall, aki nem tartozik a csoport tagjai közé, így jelöli: (k?)

Két típusú hanghatást kódolunk, a pillanatnyit, valamint a hosszabb ideig tartót. Az előbbiek közé tartozik például a köhögés, a nevetés vagy az ásítás. A hosszabb ideig tartó hanghatások közé tartozik például az, ha valaki sírva vagy nevetve mond valamit. Ilyenkor nyitó- és zárótaget használunk, annak érdekében, hogy beazonosíthassuk, milyen hosszú egy-egy hanghatás. Az alábbi példák egy nevetést, valamint egy nevetve mondott szövegrészt tartalmaznak:

(3) a. (Gabi) Szerintem (~) elég jó csaj!
b. (Gabi) Szerintem (~) elég jó csaj! (~)))

Amennyiben több hanghatás is történik egyszerre, akkor azt a megfelelő sorrendben és feltétlenül ugyanabban a sorban, azaz ugyanazzal az időbélyeggel jelöljük.

(5) #23:18# (Gabi) Mikor lesz kész a vacsora? (Éva) (~) #23:33#

Bár nem jegyezzük le a verbális és a nem verbális hanghatásokat azokban az esetekben, amikor azok nem az adott diskurzus részét képezik, azonban bizonyos, a kutatás szempontjából releváns információkat ilyenkor is jelölni kell. Ezekben az esetekben három különböző megoldást alkalmazunk annak a jelölésére, hogy mit érzékelünk a háttérben levőkről [8]. Amennyiben kivehetők a nem az adott diskurzus résztvevőinek a verbális és nem verbális közlései, és be is tudjuk azonosítani a beszélőket, akkor a résztvevők neveinek a kezdőbetűivel jelöljük a jelenlétüket, például Gabi esetében (G) taget alkalmazunk. Ha, bár a közlések kivehetők, a forrásukat azonban már nem tudjuk beazonosítani, akkor megpróbáljuk megbecsülni a számukat, például (4). Amennyiben még ez sem lehetséges, a következő annotációs jelet alkalmazzuk: (t?).

A leiratozási munka egy, a kutatás szempontjából kardinális lépése az, hogy a verbális közlések meghatározott tartalmait (p) jelzéssel látjuk el. A (p)-tartalmat a következőképpen definiáljuk: (p) jelzéssel jelöljük azokat a szövegrészeket, amelyek során a beszélgetésben egy olyan, a csoporthoz korábban vagy jelenleg is tartozó harmadik személyről esik szó, aki valószínűleg nincs jelen a diskurzus közben. Amennyiben tudjuk, hogy az adott közlés kire vonatkozik, azt is jelölni kell a megfelelő módon.

(6) (Gabi) jól alszik Éva a másik szobában munka helyett! (p-É)

4. A minőségbiztosítás szempontjai

Az 1. táblázat tartalmazza a jelen munka szempontjából fontos kódokat, és azok, a minőségbiztosítás szempontjából lényeges jellemzőit.

A táblázatban szereplő *szegmens* terminus alatt a két időbélyeg közötti szövegrészeket értjük, melyek több mondatból állhatnak és több jelölőt is tartalmazhatnak. A különböző jelölők közül a verbális tartalmaknál, valamint az azokhoz rendelt jelölők (hanghatások) esetében a szegmensen belüli elhelyezkedés is fontos, a többi jelölő esetén csupán az a lényeges, hogy jelen vannak-e, vagy sem.

A gépelt kimenettől, valamint a manuális annotációtól azt reméljük, hogy mind a társadalomtudományi, mind a nyelvészeti kutatási kérdéseink megválaszolásában lényegi segítséget nyújthatnak. Ehhez azonban elengedhetetlen egyrészt az, hogy az elvégzett munka belső érvényessége megfelelő legyen, másrészt, hogy a hibás gépelés vagy annotálás ne okozhasson fennakadást az elemzés során. A feldolgozók produktumait illetően tehát komoly minőségbiztosításra van szükség a folyamat lefolytatásához.

1. Táblázat: A jelen munka keretében ellenőrzött kódok, és azok lényegi sajátságai

Jelölő alkalmazása	Jelölő	Kategória	Leírás	Pozíció fontos
Szegmens eleje	(<név>)	Beszélő	A beszélő keresztneve	Nem
Szegmensrészlet	(<hanghatás>)	Hanghatás	<hanghatás> lehet: "s" - sóhajtás, "~" - nevetés, "*" - sírás, "gn" - gunyoros nevetés, "zn" - zavarodott nevetés	Igen

Szó	(<hang-hatás>)	Hang-hatás	<hanghatás> lehet: "sik" - sikítás, "k" - köhögés, "á" - ásítás, "pi" - pisszegés, "ujj" - ujjongás, "tor" - torokköszörülés, "f" - fütyülés, "é" - éneklés	Igen
Szegmens	(t?)	Tisztázó	távolabbi beszélők jelenléte	Nem
Szegmens vége	(k?)	Tisztázó	nem szereplő beszéde	Nem

5. A minőségbiztosítás lépései és eredményei

Tekintettel a feldolgozói munka komplexitására, egy alaposan átgondolt és részletes minőségbiztosítási folyamatot kellett megterveznünk a leiratozók munkaminőségének ellenőrzéséhez, illetve a minőség folyamatos biztosításához.

A HuTongue korpusz létrehozásának a folyamata során több minőségbiztosítási módszert is alkalmaztunk. Első lépésben a feldolgozott fájlok szúrópróba-szerű, kvalitatív átnézését végeztük minden egyes feldolgozó munkájának az ellenőrzése céljából. A szúrópróba-szerű ellenőrzés lehetőséget nyújtott ahhoz, hogy egyes feldolgozók szisztematikus hibáit azonosítsuk, és visszajelezzünk arról az egyes feldolgozók számára. Az ilyen módon feltárt hibákat, hiányosságokat a feldolgozók javították, azonban túl nagy minőségbeli eltérés esetén, abban az esetben, ha az első visszajelzés sem hozott eredményt a feldolgozó munkájának javulása terén a feldolgozó nem folytathatta a korpusz feldolgozását. Az ilyen típusú, rontott fájlokat később újra feldolgoztattuk egy másik feldolgozó munkatárs segítségével.

A korpusz minőségbiztosítása érdekében azonban szükségünk volt egy olyan átfogó mérőeszközre, amely több dimenzióban is képes mérni a munka létrehozásának eredményességét, pontosságát. Ezért a feldolgozók munkáját egymáshoz viszonyítva, és egy referenciafeldolgozó munkájához képest is ellenőriztük. A referenciagépelő kiválasztását a leiratozott fájlok elsődleges, kézi ellenőrzése alapján végeztük. Így a kézi ellenőrzés során, a konzisztensen legjobban teljesítő feldolgozót választottuk ki a kvantitatív ellenőrzés referenciagépelőjének.

A kvantitatív minőségbiztosítási folyamatot ezért a következő részfeladatokra bontottuk: Szövegegyezés, annotáció egyezése és az időbélyegek elhelyezésének helyessége.

A továbbiakban bemutatjuk a minőségbiztosításunk fókuszában álló jellemzőket, az erre kidolgozott mutatót, azaz az IRI indexet (Intercoder Reliability Index), valamint a mutató által szolgáltatott eredményeket egy 60 percet átfogó hangfájl részletén.

5.1 A mérni kívánt sajátságok

A HuTongue korpusz egyik legkülönlegesebb jellemzője az általános nyelvi korpuszokhoz képest a benne alkalmazott komplex és széles körű annotáció. Ebbe beletartozik a korábbiakban ismertetett, több dimenziós annotáció, illetve a szövegek időbélyegekkel történő kiegészítése. A minőségbiztosítás célja tehát a következő

sajátságok pontosságának a mérése volt: a gépelt szöveg helyessége, a jelölők, valamint az időbélyegek megfelelő használata.

A szövegek helyessége jelen kutatásunkban nem a magyar helyesírás szabályainak való megfelelést jelenti, hanem – élőnyelvi korpusz lévén – az elhangzottak lehető legpontosabb rögzítését, bizonyos esetekben fonetikus átírással. Ennek megfelelően a legépelt szöveg nyelvtani helyessége számunkra nem befolyásolja a kapott eredményeket.

A tagek tartalmazzák kutatásunk szempontjából a legfontosabb információt, ezért ezekre helyeztük a legnagyobb hangsúlyt a minőségbiztosítás során. A legfontosabb szempont a különböző jelek egy szegmensen belüli megléte vagy hiánya volt, valamint néhány jelölő esetében fontos volt azok mondaton belüli elhelyezkedése is. A legtöbb tag esetében a meglét ellenőrzése elégséges, azonban, különösen a hanghatásokat jelző jelölők esetében a helyes elhelyezés is fontos szempont. További nehézséget jelentett a nyitó-, illetve zárótagek jelölésének mérése is egyes kódok esetében, ezeknek a jelölőknek a pontos bemérése, összehasonlítása a szöveg egyes részein.

Végül, az időbélyegek használatát illetően mind a pontosság, mind a meglét ellenőrzése fontos volt. A feldolgozók egzakt kritériumok szerint tehetik ki az időbélyegeket a sorokvégén (a szegmenseket természetesen mindig időbélyeg zárja), így ezek konzisztens használata is nagyon fontos. Az időbélyegek adják a szöveg természetes szegmentálását, ezért ezek pontossága nagy jelentőséggel bír. Az, hogy egy adott feldolgozó mekkora szegmenseket alkot, tehát milyen gyakran tesz ki egy időbélyeget, praktikus szempontból fontos számunkra, de követnie kell a beszéd természetes meghatározottságát is, ezért a nem megfelelő helyen való tagolást is hibának tekintettük.

A fentebb ismertetett három szempont együttesen adja meg azt a feltételrendszert, amely mentén egy adott feldolgozó munkája kiértékelhető. Fontos azonban, hogy az értékelés ne csupán kvalitatív módon (azaz humán ellenőrzéssel), hanem kvantifikálhatóan is megtörténhessen. A minőségbiztosítást ugyanis iteratív folyamatként kell értelmeznünk, tehát meghatározott periódusonként ismételt minőségbiztosítási fázisokat kell beiktatnunk a feldolgozói munkába, amely feladat manuálisan egyrészt nem idő és költséghatékony, másrészt a kvantitatív értékelésre sem ad lehetőséget.

5.2 A megvalósítás eszköze és módszere

Az eddigiekben ismertetett három szempontot egyaránt lehetséges külön-külön mérőszámokkal és kompozit mérőszámokkal is mérni. A jelenlegi munkánk során egy kompozit mérőszám előállítására törekedtünk, amelynek segítségével általános képet kaphatunk a feldolgozók munkájáról, azonban, amennyiben ez szükséges, a három sajátság eredményességét külön-külön is fel tudjuk mérni minőség javítási céllal.

5.2.1 Az IRI index

A minőségbiztosítás során az egyik legfontosabb szempont volt az alkalmazott mérőszám(ok) egyszerű értelmezhetősége, illetve egyúttal a részletekbe menő információgyűjtés. Ahhoz, hogy mindezt egyben elérhessük, definiáltuk az Intercoder Reliability Indexet, melyre a továbbiakban az IRI rövidítéssel hivatkozunk. Ez a mérőszám két szöveg összehasonlítására alkalmas, amelyben a T2 a referenciaszöveget, a

T1 pedig a kiértékelt szöveget jelöli. A lehetséges értékei 0 és egy között mozognak, ahol 1 jelöli a referenciaszöveggel történő teljes egyezést (azaz, hogy a feldolgozó munkája megbízható), 0 pedig az abszolút különbözőséget (azaz, hogy a feldolgozó munkája nem megbízható). Az IRI-t az alábbi összefüggéssel definiáljuk, amelyet a három lényeges minőségbiztosítási szempont mérőszámaiból összeállított kompozit indexként értelmezünk:

$$(7) \qquad IRI = 1 - \left[w_L L(T_1, T_2) + w_{Ta} I_{Ta}(T_1, T_2) + w_{Ti} I_{Ti}(T_1, T_2) \right]$$

A kifejezés első eleme $L(T_1,T_2)$ a szövegek közti eltéréseket magába foglaló mutató, amely célra a legmegfelelőbbnek az egyszerű relatív Levenshtein távolságot találtuk[2]. Ez a mutató nem veszi figyelembe a helyesírás szabályait, és kifejezetten a szövegek közti eltérés mértékét adja meg, tehát a számunkra lényeges információt mutatja.

A szövegek közti eltérések mutatóját a jelölők pontosságát leíró mérőszám követi, melyet $I_{Ta}(T_1,T_2)$-vel jelölünk. Ez már önmagában egy kompozit mérőszám, amelyeket jelölőkategóriánként is számolunk, majd összegzünk. A mérőszám magában foglal egy a jelölők meglétéből és elhelyezkedéséből számolt pontszámot, az alábbi módon: egy adott jelölő hiánya (akár a referenciaszövegben, akár a kiértékelt szövegben) 1 pont, egy adott jelölő kategórián belüli eltérése 0,5 pont, továbbá − ha az adott kategória esetén ez releváns − egy jelölő rossz helyre történő beillesztése 0,5, és túl nagy távolságra (legalább három szó) történő rossz helyre való beillesztése pedig 1 pont. Az elemzett szövegrészen végighaladva a pontokat az összegzés után a jelölők számával osztjuk, a felső korlátjaként pedig (a többi mérőszám felső értékét figyelembe véve) 1-et adtunk meg.

Végül, a kifejezés utolsó tagja $I_{Ti}(T_1,T_2)$ az időbélyegek pontosságát összehasonlító mérőszám, melynek értékét az időbélyegek átlagos, egymástól való eltérése adja. A munka jellegéből adódóan nem várjuk el, hogy az időbélyegek helye teljesen egybeessen, így a három másodperc alatti eltéréseket egyező időbélyegnek tekintjük. Ezt azért engedhettük meg, mert az egyes szövegrészeket az időbélyegek elválasztják, és adott esetben a mondatok befejezését követően a következő beszélő mondandójának megkezdése közé bárhová eshet időbélyeg. Az efeletti pontatlanságokhoz tartozó pontszámot az alábbi összefüggéssel számítjuk (alább láthatjuk a kifejezést szemléltető diagramot):

$$(2) \qquad I_{Ti}(t_1.t_2) = 1 - e^{-0.25(|t_2 - t_1| - 3)}$$

[2] A Levenshtein távolság a szövegrész teljes hosszának az arányában.

Fig. 2. Az időbélyegek eltéréséből számolt hibapontszám

Ez a kifejezés az egyedi időbélyeg értékek összehasonlítását adja meg. Azaz eszerint az időbélyegek esetében sem "nézzük el" az alacsony hibákat, hiszen a pontszám már 10 másodperces hibáért (egy egyórás szövegben ez egyszerűen bekövetkezhet, hiszen a szöveg előrehaladtával az időbélyegekben vétett hiba kumulálódik) is igen magas értéket vesz fel. Mivel azonban az egyszeri csúszás a gépelés során gyakorlatilag egy állandóan magas értéket eredményezne, hiszen a csúszást követően minden időbélyeg torzulhat, figyelembe kell vennünk az adott szegmensek hosszát is. A szegmenshosszok hibáit az időbélyeghez rendelt hibapontszámnál figyelembe véve tehát az alábbi összefüggést használtuk.

$$I_{Ti}(T_1.T_2) = 0.1*\left(1-e^{-0.25(|t_{k2}-t_{k1}|-3)}\right)+0.8*\left(1-e^{-0.175(|t_{h2}-t_{h1}|-3)}\right) + 0.1*\left(1-e^{-0.25(|t_{z2}-t_{z1}|-3)}\right) \quad (8)$$

Ez a súlyozott exponenciális kifejezésekből álló összeg akkor vesz fel 1-hez közeli értéket, hogyha az időbélyegekben csúszás van (3 mp-nél nagyobb, a kezdő illetve záró időbélyegeket a „t_{kx}", illetve „t_{zx}" jelöli), illetve ezzel egyúttal a szegmenshossz sem megfelelő (utóbbi szerepel a legnagyobb súllyal, a szegmenshosszokat „t_{hx}" jelöli). Ezáltal a gépelés során jelen lévő állandó csúszást jóval kisebb mértékben büntetjük, mint ha a feldolgozó egy adott szegmenst nem megfelelő hosszúságúnak rögzít. Ettől eltérő módszerekkel is meghatározható az időbélyegekre adott pontszám, azonban az itt leírt szempontok figyelembe vétele számunkra a minőségbiztosításban elegendő.

Végül az (1) kifejezés szerint az IRI különböző elemeit különböző súlyokkal vesszük figyelembe (melyek összege 1-et ad) attól függően, hogy melyik pontosságra helyezzük a legnagyobb hangsúlyt. Emellett célszerű alacsony súlyt adni az időbélyegek pontosságát leíró mennyiségnek, mivel ez a többivel ellentétben nem normált mennyiség. Ennek megfelelően, ha a többi mérőszámmal megegyező súllyal vesszük figyelembe, azzal torzíthatjuk az IRI index értékét. Jelen írásunkhoz a 0.2-0.6-0.2 súlyokat használtunk, kiemelve a jelző pontosságának számítását.

Az ismertetett megoldás egy viszonylag egyszerű lehetőséget ad a feldolgozók szövegeinek az összehasonlítására, azonban szem előtt kell tartanunk, hogy az egyszerű számként létrejövő IRI nem feltétlenül minden esetben elég informatív a munka javításához. Emellett számos probléma felmerülhet a szövegekkel kapcsolatban, amelyek az IRI-t kiegészítő megoldásokat kívánnak.

5.2.2. Praktikus megfontolások és az IRI alkalmazása

A gyakorlatban az IRI-t úgy használhatjuk, hogy egy adott szövegrészt több feldolgozóval is legépeltetünk és annotáltatunk, majd az így létrejövő szövegekre egyedileg kiszámoljuk az IRI értékét. Praktikus okokból nem csupán a kompozit index számítását végezzük el, hanem az IRI részeit külön-külön is elemezzük, hogy megértsük, hogy az egyes feldolgozók esetében mely feladatok jelentik a problémát. Ez főként a tagek alkalmazásánál, és az időbélyegek elhelyezésénél fontos.

A tagek esetében az IRI kiszámítását mindig adott jelölőkategóriákra végezzük el, amellyel aggregált értéket kapunk. Az egyes kategóriák egyedi elemzése lehetővé teszi közülük a problematikusak azonosítását, így például a különböző típusú hanghatások észlelhetőségét, a beszélgetésekben jelen lévő harmadik személy jelenlétét, stb. Továbbá külön vizsgálhatóak a hibakategóriák szerinti "pontszámok", így felismerhető, ha egy-egy feldolgozó valamely taget túl gyakran, vagy éppen túl ritkán használja, esetleg szisztematikusan rosszul helyez el a gépelt szövegben.

Az időbélyegek esetében elsőként a fentebbinél egyszerűbb mérőszámot vizsgálunk: az adott szövegben levő időbélyegek számát, és időbeli eltérését. Itt a feldolgozók számára értékes visszajelzés az, ha túl sok, vagy túl kevés időbélyeget látunk a munkájukban. Azonban az időbélyegek használatában való eltérés egyéb problémákra is rámutathat. A feldolgozók például gyakran eltérően és nem megfelelően szegmentálják velük a szöveget, vagy pedig olyan szövegrészeket is legépelnek, amelyet nem kellene (egy nem az adott diskurzusban résztvevő verbális és nem verbális közléseit). Nem ritka tehát, hogy két szövegben eltérő szegmenseket, illetve valamely szövegben indokolatlan szegmenseket találtunk. Az időbélyegek száma mellett a szegmenseket kezdő és záró időbélyegek pontossága is fontos mérőszám.

Emellett, mivel a minőségbiztosításra használt gépelt szöveg mérete megfelelt a beszélt szövegben az egy órás időtartamnak, az IRI számítási algoritmusának futásideje is igencsak megnőtt a nagyszámú természetes szegmens összerendelése miatt.

A feldolgozók munkájának összevetését tehát -- a fentebb elmondottak okán -- nem lehetett az időbélyegekre támaszkodva elvégezni, így a szöveg mesterséges szegmentálása mellett döntöttünk. A munkát az ún. horgonyszavakra támaszkodva végezzük. A horgonyszavak az összehasonlított szövegekben egyaránt maximálisan egyszer előforduló, négy betűnél hosszabb, értelmes szavak. Emellett egy olyan megkötést is tettünk, hogy csak azokban a természetes, (azaz a feldolgozók által létrehozott) szegmensekben keressük a horgonyszavakat, amelyeket az adott hangrögzítőt viselő személy mondott (azaz amelyek a beszélőhöz tartozó jelölővel kezdődtek)[3]. A mesterséges szegmentálást

[3] Ennek a megkötésnek a számítások felgyorsítása mellett a praktikus oka az volt, hogy a többi szereplő által elmondott szöveget a feldolgozók nagyon eltérő minőségben gépelték. Ebből fakadóan különböző hosszúságú szegmenseket kaptunk, és sok esetben egyes feldolgozók olyan szegmenseket is legépeltek, amit a többiek nem. Végül tehát a legépelt szövegekben a megegyező, egyértelműen összehasonlítható tartalom az adott beszélő szövege volt.

úgy végeztük, hogy az adott horgonyszavak feldolgozók által kijelölt mondatait tekintettük a szegmensek határainak – a feldolgozók jellemzően hasonló helyekre tették a mondatvégi írásjeleket a beszélt szövegben. Azaz egy adott mesterséges szegmens egy olyan mondattal kezdődött, amelyben benne volt az adott horgonyszó.

Így nagyobb számú, de kisméretű szövegszegmenst kaptunk. Vizsgálataink során azt tapasztaltuk, hogy az így létrehozott mesterséges szegmensek akkor is jól összerendezhetőek, ha az időbélyegekben, vagy akár a szegmensek hosszában komolyabb eltérés mutatkozik. Ez alapján az IRI indexeket a mesterséges szegmensek között számítjuk ki, a végső indexeket pedig a szegmenshosszal súlyozott összegként. Egy mesterséges szegmens nagyjából egy perces szöveget fogott át így.

5.3 Eredményeink az első minőségbiztosítási fázisban

Az első minőségbiztosítási fázis a korpuszépítés kezdeti szakaszában, közvetlenül a feldolgozási útmutató megismertetése és a feldolgozók betanítása után zajlott. A méréshez minden feldolgozónak ugyanazt a szöveget adtuk. A szöveg a teljes hanganyag kis része volt, amelyet úgy választottunk meg, hogy a feldolgozóknak az összes típusú jelölőt használniuk kelljen, illetve, amelyben kellő mennyiségű időbélyeg is szerepel.

Fontos lépés volt a referenciaszöveg kiválasztása is. Erről a rendelkezésre álló szövegek kézi feldolgozásával, a minőség manuális vizsgálata alapján döntöttünk, különös figyelmet fordítva a jelölők helyes alkalmazására.

A 2. táblázatban először is a szövegek struktúráját tekintjük át. Amint arra a táblázat adatai rámutatnak, a szövegek hossza feldolgozónként változik. Néhol ez az eltérés számottevő, ami annak tudható be, hogy a feldolgozók a távolabbi résztvevők mondatait eltérő mértékben érzékelik. Ez hasonlóképp igaz az időbélyegek számára, illetve az alkalmazott jelölőkre is. A táblázatban az időbélyegek és a jelölők száma mellett látható a legépelt szavak száma is. A referenciaszövegként használt szövegben található a legtöbb időbélyeg és jelölő, így a jelölők összehasonlításakor kevésbé fordulhat elő, hogy a referenciaszöveg a hiányos.

2. Táblázat: Az ellenőrzött szövegek gyakorisági jellemzői

Szöveg	Karakter-szám	Szavak száma	Szöveg	Karakter-szám
Referencia	64273	10202	1940	1357
Feldolgozó 1	64414	10276	1868	1139
Feldolgozó 2	33737	5183	911	637
Feldolgozó 3	57842	9001	1503	1105
Feldolgozó 4	43837	6482	1203	1051
Feldolgozó 5	54992	8951	1312	1132
Feldolgozó 6	48943	7326	1404	1079
Feldolgozó 7	57674	9323	1483	1306
Feldolgozó 8	60449	9760	1567	1413
Feldolgozó 9	52269	8187	1399	897

A szövegek horgonyszavas szegmentálását követően 55 szegmenst azonosítottunk, melyek átlagosan 380 karakterből állnak, azonban igen nagy szórásúak voltak (321 karakter).

Az elemzés következő lépéseként kiszámítottuk az IRI pontszámokat az egyes szövegekre a referenciaszöveghez hasonlítva. Praktikus okokból a pontszámokat kategóriánként számítottuk a végleges IRI pontszám ebből az (1) kifejezésben említett súlyozás szerint adódhat. A minőségbiztosítási folyamatban ezeket a pontszámokat külön-külön alkalmaztuk. Az eredmények azt mutatják, hogy a feldolgozók egyaránt konzisztens teljesítmény nyújtottak mind a szöveges tartalmak egyezését illetően (tehát, amelyben nem voltak benne a jelölők és az időbélyegek), mind az időbélyegek és a jelölők alkalmazásának terén. Ezért végül csupán a számunkra az annotálás szempontjából legfontosabb jellemzőt vettük alapul a feldolgozók rangsorolásában: a jelölők használatát[4].

3. Táblázat: A szövegekre számolt IRI részértékek

Szöveg	Szóbeli egyezés	Időbélyeg pontszám	Jelölő pontszám
Feldolgozó 1	0.643885701	0.425844137	0.513876
Feldolgozó 2	0.671654333	0.264923636	0.456804
Feldolgozó 3	0.641393827	0.257121049	0.531356
Feldolgozó 4	0.681724923	0.243786923	0.497431
Feldolgozó 5	0.699152208	0.212764862	0.585468
Feldolgozó 6	0.669104015	0.228939292	0.509069
Feldolgozó 7	0.640636905	0.237307279	0.519228
Feldolgozó 8	0.713745273	0.218149973	0.531762
Feldolgozó 9	0.720822972	0.167920389	0.54657

Az egyszerű rangsorolást követően vettük figyelembe az időbélyegek és a szóbeli egyezés mutatóit, majd részletekbe menően vizsgáltuk a jelölők használatát jelölőkategóriák szerint is, itt azonban nem célunk ennek ismertetése.

A minőségbiztosítási fázis végén a teljesítmény alapján hat feldolgozóval nem folytattuk tovább a munkát, a többiekkel pedig részleteiben egyeztettük a minőség-ellenőrzés eredményét. Ennek eredményeképp sikerült javítanunk a tagek használati szabályait, tehát sokkal precízebb kritériumokat megadni az annotálás részleteivel kapcsolatban.

6. Összegzés, további tervezett lépések

Dolgozatunkban a jelenleg is fejlesztés alatt lévő HuTongue, magyar nyelvű spontán beszéd korpusz minőségbiztosításában alkalmazott néhány módszert mutattuk be. A korpusz méretéből fakadóan a gépelés és az annotálás komoly humán munkaerőt igényel, így a megfelelően alapos és pontos minőségbiztosítás elengedhetetlen. Erre az

[4] Ez alól egyedüli kivétel az 1. Feldolgozó volt, ahol a többiektől elmaradó szövegi egyezésnek és a magas időbélyeg pontszámnak köszönhetően egyedileg is elemeztük a hibákat.

általánosan használt módszerek mellett egy, a HuTongue különleges jellemzőit figyelembe vevő egyedi mérőszámot alkottunk, amellyel jelentősen egyszerűbbé tettük a minőségbiztosítási fázisokban gépelt dokumentumok elemzését és a gépelést végzők teljesítményének az értékelését.

A dolgozatban röviden ismertettük az általunk definiált mérőszám részleteit, valamint az alkalmazásánál figyelembe vett gyakorlati szempontokat. Ezt követően, az első minőségbiztosítási fázis adatain bemutattuk, hogyan lehet a mutatót a gyakorlatban alkalmazni.

Az ismertetett eredmények alapján azt látjuk, hogy a mutató már jelenlegi formájában is alkalmas az egyes minőségbiztosítási fázisokban a feldolgozók értékelésére, azonban lehetséges, hogy további finomítással a mutató még érzékenyebbé tehető a feldolgozás során elkövetett hibákra. A jelenleg alkalmazott, a tagek és időbélyegek közti eltéréseket mutató mérőszámok korlátozott mértékben érzékenyek bizonyos hibákra. Az időbélyegek esetén a hibák típusának átsúlyozása érzékenyebben mutathatja a hibákat, illetve a jelölők használata esetén lehetséges, hogy nyers mutatók bevétele a mutatók számításába (például az egyes szegmensekben a beillesztett jelölők közti számszerű különbség) is javíthatja az IRI-t. Emellett a jelenleg használt Levenshtein távolságot egy jóval részletesebb mérőszámmal is lehetne helyettesíteni, amely kevésbé érzékeny a szavak sorrendjére, és inkább az adott szegmensekben szereplő szavak és azok jelentése közti eltéréseket veszi figyelembe.

Mindezt összefoglalva azt tapasztaltuk, hogy a fejlesztett megoldás fontos előnye egyrészt az, hogy folyamatosan nyomon tudjuk követni a munka minőségének szinvonalát, másrészt a kapott mérési eredmények alapján visszajelzésre vagyunk képesek a feldolgozók felé, biztosítva ezzel a munka minőségének folyamatos javítását. Ennek köszönhetően amellett, hogy a gépelés pontossága javult, az annotációs jelölők használata is jelentős mértékben fejlődött.

Köszönetnyilvánítás

A kutatást az Európai Kutatási Tanács (European Research Council), az Európai Unió Horizont 2020 kutatási és innovációs programjának keretében (ERC_CoG_2014_648693 sz. szerződésben) támogatja, a kutatás vezetője Takács Károly.

Bibliográfia

1. Crowdy, S.: Spoken Corpus Design. Lit Linguist Computing (1993) **8**(4): 259–265.
2. Galántai J., Pápay B., Kubik B., Szabó M., és Takács K.: A pletyka a társas rend szolgálatában: Az informális kommunikáció struktúrájának mélyebb megértéséért a Computational Social Science eszközeivel. Magyar Tudomány. Megjelenés előtt.
3. Gósy M., Gráczi T.E., Gyarmathy D., Váradi, T., Veresné Horváth, V.: Magyar spontán beszéd adatbázis = Hungarian Spontaneous Speech Corpus. Munkabeszámoló. OTKA (2012) (http://www.nytud.hu/adatb/bea/index.html).

4. Hemphill, C.T., Godfrey, J.J., Doddington, G.R.: The ATIS spoken language systems pilot corpus. In Proceeding HLT '90 Proceedings of the workshop on Speech and Natural Language. Hidden Valley, Pennsylvania. (1990) 96–101.
5. Kugler N.: Megfigyelés és következtetés a nyelvi tevékenységben. Budapest, Tinta. (2015)
6. Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous Speech Corpus of Japanese. In Proceedings of LREC. (2000) 947–952.
7. Oostdijk, N.: The spoken Dutch corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer eds. Proceedings of the Second International Conference on Language Resources and Evaluation 2. (2000) Paris. ELRA. 887–893.
8. Szabó, M.K., Galántai J.: Egy magyar nyelvű spontán beszélt nyelvi korpusz (HuTongue) létrehozásának tapasztalatai. In MANYE-kongresszus konfenreciakötete (2017) Megjelenés előtt.

“S azóta jól élnek és vidám dalokat énekelnek Leninről és Sztálinról.” Szovjet propagandamesék műfaji azonosításának kísérlete

Horváth Csilla

MTA Nyelvtudományi Intézet

Kivonat: Ebben a munkában a Szovjet népmesék Leninről és Sztálinról című mesegyűjtemény prózai szövegeinek (Lenin-mesék) lehetséges műfaji besorolásához próbálunk számítógépes nyelvi eszközökkel támpontokat nyújtani más doménekbe tartozó szövegeken végzett összevetés segítségével. A kétes műfaji meghatározású Lenin-meséket szófaji gyakoriság és emócióeloszlás alapján hasonlítjuk össze rokon műfajú domének korpuszával. Eredményeink alapján a Lenin-mesék szófaji gyakoriság és eloszlás tekintetében valóban a népmesékkel mutatnak rokonságot, emóciók gyakoriságában és eloszlásában azonban a szovjet propaganda-irodalom különböző doménjeihez hasonlítanak.

1 Bevezetés

Ez a munka a Szovjet népmesék Leninről és Sztálinról című mesegyűjtemény prózai szövegeinek lehetséges műfaji besorolásához próbál számítógépes nyelvi eszközökkel támpontokat nyújtani más doménekbe tartozó szövegeken végzett összehasonlítás segítségével.

Az összesen öt doménbe sorolt szövegeket először elemeztük, majd megvizsgáltuk bennük a szófajok gyakorisági sorrendjét és eloszlását, illetve az emóciók gyakorisági sorrendjét és eloszlását is. A műfaji besorolás elősegítése érdekében az eredményeket a Kendall-változó segítségével hasonlítottuk össze.

Vizsgálataink során az alábbi kérdésekre kerestük a választ:

- milyen emóció-eloszlás jellemzi a Lenin-meséket
- mutatnak-e különbséget ebben a tekintetben a Lenin-mesék egyes kiemelt részei
- a fenti eloszlások milyen hasonlóságot és különbséget mutatnak a többi vizsgált domén hasonló eloszlásaihoz képest
- az összehasonlítás eredménye alapján melyik doménbe sorolhatók a vizsgált kötet meséi

2 A vizsgált korpusz

Vizsgálatunk alapjául egyik részről a Szovjet népmesék Leninről és Sztálinról című kötetben [1] található meseszövegek (a továbbiakban: Lenin-mesék) szolgáltak. A mesék terjedelme 1348 mondat, 13 486 szó. A Lenin, Sztálin a mesében való megjelenését jelölő szöveghely alapján két részre osztott szövegek terjedelme a megjelenés előtt (továbbiakban: Lenin előtt) 645 mondat, 6648 szó, a megjelenés után (a továbbiakban: Lenin után) 703 mondat, 6589 szó.

1. Táblázat: A Szovjet népmesék Leninről és Sztálinról vizsgált korpuszának adatai

	Lenin-mesék	Lenin előtt	Lenin után
mondat	1348	645	703
szövegszó	13 486	6648	6589

A Lenin-mesék eredetéről keveset lehet tudni. A kötet mindössze a fordítók nevét tünteti fel, más hasonló gyűjteményektől eltérően sem a szerkesztőkről, sem a mesék eredeti forrásáról nem közöl információt. Bár orosz nyelven több hasonló gyűjtemény is létezik (pl. [2]), a Lenin-mesék szövegei közül ezekben csak néhány található meg, nem sikerült mindet felderíteni. A Lenin-mesék szövegének népmesei besorolása már csak a kötet címét olvasva is kétségesnek tűnik, a szövegek elolvasása után pedig ez a gyanú csak fokozódik.

Lenin és Sztálin a mesékben történő megjelenése előtt a Lenin-mesék szövege néhány szóhasználati érdekességtől eltekintve hasonló felépítésű, mint a Szovjetunió népeitől gyűjtött más népmesék magyar fordításáé, amint azt az alábbi, orosz és nivh meséből vett példán látható:

A legdrágább

Nem tudtak megegyezni a kolhozparasztok. Elhatározták hát, hogy elküldik Fjodort, Marjuskát meg Alekszejt: járják be az orosz földet, tudják meg, mi a legjobb és a legdrágább ezen a földön.

Felkészültek hát hárman: Fjodor, Marjuska meg Alekszej, de nem tudták, merre induljanak. Marjuska azt mondta, hogy mielőtt a nagyanyja meghalt, meghagyta neki: »Hogyha nem tudod, merre indulj, vedd kezedbe az orsómat, fogd meg a cérnát a végén, s guritsd el az úton. Amerre az orsó gördül, arra indulj.«

Megmondta ezt Marjuska Fjodornak meg Alekszejnek is. Elővették a kemencesutból az orsót, úgy tettek, ahogy a nagyanyó meghagyta, aztán elindultak az orsó nyomában.

Mentek, mendegéltek.

A hősök

Sokáig éltek a földön emberek, anélkül, hogy látták volna a napot. Az a gonosz és kapzsi ikiz törzs lopta el a napot, amelyik hazugság és álnokság árán birtokába

kaparintotta az egész földet. Egy feneketlen, mély szakadékban tartották fogságban az ellopott napot, nehogy egyetlen ragyogó sugár is érje a hideg, sötétségbe borult földet.

Sötét éjszaka volt a földön, de a legsötétebb a nivheknél volt, mivel ők a sötét tajgában éltek.

Lenin és Sztálin a mesékben történő megjelenése után azonban a Lenin-mesék szövegei szakítani látszanak a népmesék hagyományos struktúrájával és szóhasználatával is, fogalmazásuk inkább az agitációs propaganda termékeinek fogalmazásmódjához és felépítéséhez hasonlít, amint az az alábbi evenk és nanáj mese részlete is mutatja:

A boldogság csillaga

(...) Odament Barkaul a tűzmadárhoz és így szólt:
— Nagy sas, Sztálin hatalmas szárnya!
— Igazad van. Ülj fel a madárra, elviszünk.
— Már hogy ülnék fel? Félek.
— Ülj fel, Sztálin vendégségbe hív.
Barkaul az emberekre nézett, nem tréfálnak-e vele? De nem. Szemük komoly maradt, igazat mondtak.
— Barátom hív vendégségbe. Mennem kell. El is megyek. Hallani fogom Sztálin szavát.
A tűzmadár felzúgott és elvitte Barkault délre.
Sok folyó van a földön — s mind a tengerbe ömlik. A tengertől pedig Moszkvába vezet az út. Sok ösvény van a földön, de mind Moszkvába vezet. Moszkva hatalmas.
A tűzmadár Moszkva felett repült. És Barkaul mindent látott..
Sztálinnál volt vendégségben, s mintha csak otthon lett volna.
Együtt pipázgattak, Barkaul mesélt a tajgáról, mesélt az új életről... (...)

A nép napja

(...) Ezt a hőst Leninnek hívták. Mikor az emberiséget nagy szerencsétlenség érte — mikor meghalt Lenin — a gonosz emberek megörültek. Azt gondolták, hogy a nap ezentúl nem ragyog a népnek. De a gonosz emberek öröme nem sokáig tartott, mert a nép napját egy másik vitéz őrizte tovább, Lenin legközelebbi barátja és elvtársa — Sztálin. E hős erejével nem mérheti össze magát senki. Szeme mindent lát, ami a földön történik; fülébe minden eljut, amit a népek beszélnek; minden az eszében van, amit az emberek gondolnak; szíve befogadja minden ember örömét és bánatát; gondolatai mélyek, akár a feneketlen óceán; hangját mindenki hallja, aki a földön él. Ő a földön a leghatalmasabb. S íme, Lenin kezéből átvette a napot és magasra, igen magasra emelte. Azóta nem alszik ki a földön a boldogság, mert nem alhat ki az égen a nap sem.

A vizsgált korpusz másik részét a kontroll szövegek alkotják. Kontroll szövegeket négy doménből választottunk. Az első doménba tartoznak a Repülő hajó című

kötetben [3] található népmesék, ezek közül is azok a meseszövegek, amelyek nemzetiségi besorolása megegyezik a Lenin-mesék nemzetiségi besorolásával, vagyis az orosz, ukrán, belorusz, hanti, evenki, nanáj mesék. A szovjet népmesei domén (továbbiakban: szovjet mesék) korpusza összesen 2 843 mondatból, 29 326 szóból áll. A második doménba a tematikai hasonlóság okán, vagyis valós történelmi személy, uralkodó középpontba állítása okán a Mátyás királyról szóló mesék és mondák [4] [5] szövegei (továbbiakban: Mátyás-mesék) tartoznak. Ezek terjedelme 7768 mondat, 80 364 szó.

2. táblázat: A Szovjet népmesék Leninről és Sztálinról vizsgált korpuszával összehasonlított kontrollszövegek adatai

	szovjet mesék	Mátyás-mesék	Rákosi-mesék	szocreál irodalom
mondat	2 843	7 768	883	26 833
szövegszó	29 326	80 364	7668	353 227

A harmadik doménba tartoznak a Lenint és Sztálint dicsőítő mesékhez hasonlóan a személyi kultuszt szolgáló szövegek (továbbiakban: Rákosi-mesék), melyek a Rákosi Mátyás 60. születésnapjára összeállított ünnepi kötetben [6] találhatók. A kötetből azokat a szövegeket vontuk be a vizsgálatba, amelyek mind tematikájukban, mind stílusukban hasonlítanak a Lenin-mesék szövegére, vagyis nem egyszerűen Rákosit dicsőítik, hanem Rákosi szereplőként, főszereplőként jelenik meg bennük. A Rákosi-korpusz terjedelme 883 mondat, 7668 szó. A negyedik doménba a szocializmus (Szovjetunióban, és részben Magyarországon is) preferált irodalmi stílusát, a szocialista realizmust reprezentáló Gorkij-regények [7,8,9,10,11] tartoznak (továbbiakban: szocreál irodalom), ezek terjedelme 26 833 mondat, 353 227 szó.

3 A szófajok eloszlása

A vizsgálat első lépéseként a magyarlanc elemzővel [12] elemeztük a szövegeket. Vincze doménvizsgálathoz [13] hasonlóan kizárólag a fő szófaji információkra összpontosítottunk, minden token esetében csak a fő szófajt vettük figyelembe. A szófajok az összes doménben megfigyelhető eloszlása a 3. táblázatban látható.

A szófajok eloszlásának vizsgálatához, illetve a domének közti hasonlóságok és különbségek megállapításához a Kendall-együtthatót (W) alkalmaztuk, amely a vizsgált elemek, jelen esetben a szófajok gyakorisági rangsorát felállítva mutatja meg, mennyire homogének a vizsgált szövegek. Az eredményeket a 4. táblázat mutatja.

A Kendall-együttható értéke alapján a szövegek homogének, néhány eltéréstől eltekintve (mint pl. az adverbiumok szépirodalmi szerzők szövegeiben mért nagyobb gyakorisága) a domének között jelentős különbség nem figyelhető meg, a Lenin-mesék legnagyobb hasonlóságot a szovjet mesékkel és a Rákosi-mesékkel mutatnak, míg a korpuszból leginkább a szocialista realista szövegek doménje válik ki.

3. táblázat: A szófajok gyakorisági rangsora doménenként

	Lenin-mesék	Lenin előtt	Lenin után	Szovjet mesék	Mátyás-mesék	Rákosi-mesék	Szocreál irodalom
N	1	1	1	1	1	1	1
V	2	2	2	2	2	2	2
DET	3	3	3	3	3	4	4
ADV	4	4	4	4	4	3	5
PRON	5	5	5	5	5	6	6
ADJ	6	6	6	6	6	5	3
CONJ	7	7	7	7	8	7	7
PROPN	8	8	8	8	7	8	9
SCONJ	9	9	9	9	9	9	8
NUM	10	10	10	10	10	11	11
ADP	11	11	11	12	11	10	10
PART	12	12	12	11	13	12	12
INTJ	13	14	13	13	12	13	13
AUX	15	16	14	14	14	14	14
SYM	15	15	15	16	15	16	15
X	15	13	16	15	16	15	16

4. Táblázat: A domének hasonlósága a szófajok eloszlása terén.

	Lenin-mesék	Lenin előtt	Lenin után	szovjet mesék	Mátyás-mesék	Rákosi-mesék	szocreál irodalom
Lenin-mesék		0,993	0,991	0,9976	0,9939	0,9951	0,9865
Lenin előtt	0,993		0,9953	0,9942	0,993	0,9917	0,9868
Lenin után	0,991	0,9953		0,9944	0,995	0,9919	0,9864
szovjet mesék	0,9976	0,9942	0,9944		0,9939	0,9939	0,9853
Mátyás-mesék	0,9939	0,993	0,995	0,9939		0,9927	0,9853
Rákosi-mesék	0,9951	0,9917	0,9919	0,9939	0,9927		0,9914
szocreál irodalom	0,9865	0,9868	0,9864	0,9853	0,9853	0,9914	

4 Az emóciók eloszlása

Mivel a domének összehasonlítása során a szófajok eloszlásának vizsgálata nem hozott döntő eredményeket és nem segítette elő a Lenin-mesék műfaji besorolását, szükségessé vált a szövegek más jellemzők szerinti összehasonlítása. Az összehasonlítás alapjául az emóciók eloszlásának vizsgálatát választottuk, a vizsgálathoz használt nyolc kategóriás emóciószótár ([14, 15]) alapján felállított emóciósorrend összehasonlítását az 5. táblázat mutatja.

5. táblázat: Az emóciók gyakorisági sorrendje doménenként

	Lenin-mesék	Lenin előtt	Lenin után	szovjet mesék	Mátyás-mesék	Rákosi-mesék	szocreál irodalom
öröm	1	1	1	1	1	1	1
bánat	2	2	2	2	2	2	2
düh	3	3	3	4	4	3	3
szeretet	4	4	4	3	3	5	5
félelem	5	5	5	5	5	4	4
undor	6	6	6	6	6	7	6
meglepetés	7	7	7	7	7	6	7
feszültség	8	8	8	8	8	8	8

Az emóciók gyakorisági sorrendje hasonlóságot mutat abban, hogy minden doménben az örömöt és a bánatot kifejező elemek kerülnek az első két helyre. A továbbiakban azonban megfigyelhetők lényegi eltérések: a düh fontosabb szerepet játszik a szocialista jellegű doménekben, mint a folklórban, a szeretet kifejezésében pedig a Lenin-mesék a folklór domének és a szépirodalmi szerzők által írott szövegekből összeállított domének között helyezkedik el.

Az emóciók gyakorisági sorrendjének pontosabb értelmezésének érdekében áttekintjük az emóciók gyakorisági számait is, melyet a 6. táblázatban közlünk.

Az emóciók doménenkénti eloszlása alapján a gyakorisági sorrend alapján levont következtetések tovább pontosíthatók, ha megállapítjuk: a Lenin-mesékben a bánat és a düh a Lenin, Sztálin a mesékben történő megjelenése előtt jellemző, míg az öröm és szeretet nagyobb arányban van jelen a főhősök megjelenése (a szövegek többségében egyúttal a rendszerváltás és a szocializmus diadala) után.

A szófaji vizsgálatokhoz hasonlóan az emóciók esetében is a Kendall-együtthatót használtuk a domének hasonlóságának megállapításához, ennek eredményei a 7. táblázatban láthatóak.

Az összehasonlítás eredményei alapján elmondhatjuk, hogy míg a szovjet és magyar folklórdomének, vagyis a szovjet népmesék és a Mátyás királyról szóló mesék között teljes a hasonlóság, addig ez a szovjet és magyar propagandaműfajokról, vagyis a Lenin-mesékről és a Rákosi-mesékről nem mondható el maradéktalanul.

6. táblázat: Az emóciók eloszlása doménenként

	Lenin-mesék	Lenin előtt	Lenin után	szovjet mesék	Mátyás-mesék	Rákosi-mesék	szocreál irodalom
öröm	234	85	148	449	2642	62	4642
bánat	60	36	24	193	555	24	1978
düh	43	26	17	104	157	22	1598
szeretet	25	9	16	115	311	13	993
félelem	23	8	15	72	143	15	1038
undor	12	6	6	44	80	2	739
meglepetés	5	3	2	18	27	7	258
feszültség	0	0	0	0	0	0	0

7. táblázat: A domének hasonlósága az emóciók eloszlásának terén

	Lenin-mesék	Lenin előtt	Lenin után	szovjet mesék	Mátyás-mesék	Rákosi-mesék	Szocreál irodalom
Lenin-mesék		1	1	0,9881	0,9881	0,9762	0,9881
Lenin előtt	1		1	0,9881	0,9881	0,9762	0,9881
Lenin után	1	1		0,9881	0,9881	0,9762	0,9762
Szovjet mesék	0,9881	0,9881	0,9881		1	0,9524	0,9643
Mátyás-mesék	0,9881	0,9881	0,9881	1		0,9524	0,9643
Rákosi-mesék	0,9762	0,9762	0,9762	0,9524	0,9524		0,9881
Szocreál irodalom	0,9881	0,9881	0,9762	0,9643	0,9643	0,9881	

5 Az eredmények összehasonlítása és értékelése

Az eredmények alapján kirajzolódnak a domének közti hasonlóságok, illetve távolságok. A Lenin-mesék szófaji gyakoriság és eloszlás tekintetében a népmesékre, vagyis a szovjet mesékre és a Mátyás-mesékre hasonlítanak, míg az emóciók gyakoriságában és eloszlásában azonban a szovjet propaganda-irodalom különböző doménjeivel mutatnak egyezést, a pozitív emóciók pedig, mint az öröm és szeretet, Lenin és Sztálin a mesékben történő megjelenése után kerül túlsúlyba.

A vizsgálat pontosságát, így a Lenin-mesék műfaji besorolásának lehetőségét elősegítené a különböző domének korpuszának terjedelmi kiegyenlítése, illetve amennyiben a Lenin-mesékhez hasonló másik kötet magyar nyelven nem jelent meg, legalább a szovjet mesék és a Rákosi-mesék doménjének bővítése további szövegekkel. Érdekességgel szolgálna még ezen kívül a Lenin-mesék

karakterspecifikus emócióelemzése, vagyis annak számítógépes nyelvészeti eszközökkel történő vizsgálata, hogy milyen módon kapcsolódnak a pozitív emóciók Leninhez és Sztálinhoz, illetve milyen negatív emóciók fordulnak elő ellenfeleik szövegkörnyezetében.

6 Összefoglalás

Ebben a munkában a Szovjet népmesék Leninről és Sztálinról című mesegyűjtemény prózai szövegeinek lehetséges műfaji besorolását próbáltuk meg elősegíteni számítógépes nyelvi eszközök felhasználásával. A Lenin-meséket előbb magyarlanccal elemeztük, majd megállapítottuk a szófajok gyakorisági sorrendjét, illetve eloszlását a korpuszban, valamint megvizsgáltuk az emóciók gyakorisági sorrendjét és eloszlását is. A műfaji besorolás elősegítésére ezeket az eredményeket összehasonlítottuk hasonló domének korpuszának azonos adataival. A Lenin-mesék és a többi domén szövegeinek hasonlóságát Kendall-változóval vizsgáltuk a szófajok és az emóciók esetében is.

Eredményeink alapján a Lenin-mesék szófaji gyakoriság és eloszlás tekintetében elsősorban a népmesékkel mutatnak rokonságot, míg az emóciók gyakoriságában és eloszlásában azonban a szovjet propaganda-irodalom különböző doménjeihez hasonlítanak, a pozitív emóciók, mint az öröm és szeretet, Lenin és Sztálin a mesékben történő megjelenése után kerül túlsúlyba.

Hivatkozások

1. Szovjet népmesék Leninről és Sztálinról. Ifjúsági Könyvkiadó, Budapest (1953)
2. Шуб, Т.А. (ed.): Ленин и Сталин в твочестве народов СССР. Советский писатель, Москва (1938)
3. Ortutay, Gy – Rab, Zs. (eds.): A repülő hajó. Ifjúsági Könyvkiadó, Budapest (1955)
4. Kóka, R.: Mátyás király rózsát nyíló ostornyele. Timp Kiadó, Budapest (1990)
5. Kríza, I.: Mesék és mondák Mátyás királyról. Helikon Kiadó, Budapest (2004)
6. Gyárfás, M. Magyar írók Rákosi Mátyásról. Szépirodalmi Könyvkiadó, Budapest (1952)
7. Gorkij, M.: Az Artamanovok. Európa Könyvkiadó, Budapest (1985)
8. Gorkij, M.: Az igazi kultúra. Szikra Könyvkiadó, Budapest (1950)
9. Gorkij, M.: Gyermekéveim. Atheneum Nyomda, Budapest (1920)
10. Gorkij, M.: Három elbeszélés. Singer és Wolfner Kiadása, Budapest (én)
11. Gorkij, M.: Inasévek, az én egyetemeim. Európa Könyvkiadó, Budapest (1973)
12. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763_771
13. Vincze, V. : Domének közti hasonlóságok és különbségek a szófajok és szintaktikai viszonyok eloszlásában. In: XI.Magyar Számítógépes Nyelvészeti Konferencia, pp. 182-192 (2013)
14. Drávucz, F., Szabó M.K., Vincze V: Szentiment- és emóciószótárak eredményességének mérése emóció- és szentimentkorpuszokon. XIII.Magyar Számítógépes Nyelvészeti Konferencia, pp. 228-239 (2016)

15. Szabó M.K., Vincze V., Morvay G.: Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In: *Távlatok a mai magyar alkalmazott nyelvészetben.* Budapest: Tinta (2016)

Mozgást jelentő igék argumentumszerkezetének korpuszalapú vizsgálata

Lázár Bernadett[1], Szabó Martina Katalin[1,2,3], Vincze Veronika[4,5]

[1]Precognox Informatikai Kft.
[2]Szegedi Tudományegyetem, Szláv Intézet, Orosz Filológiai Tanszék
[3]MTA TK „Lendület" RECENS Kutatócsoport
[4]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[5]Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék

Kivonat: Dolgozatunkban az *átjön/bejön*, valamint az *átmegy/bemegy* igekötős igepárok argumentumszerkezeteit vizsgáljuk a Szeged és a Dívány korpusz adatai alapján. A kutatás célja az, hogy számítógépes eszközökkel, korpuszadatok alapján feltárjuk a kiválasztott igék argumentumszerkezeteinek kvantitatív és kvalitatív sajátságait, majd vizsgálataink eredményeit összevessük különböző szótárak szócikkeivel és a szakirodalom vonatkozó megállapításaival. A vizsgálattal tehát egyrészt az igei sajátságok megismerésére törekszünk, másrészt szeretnénk megtudni, hogy a szótárak, valamint a szakirodalmi adatok a valós nyelvhasználatot tükrözik-e. Mindezzel összefüggésben a dolgozat fontos célja az is, hogy a lexikológiai kutatásokhoz adalékul szolgáljon. A munka során azt is meg szeretnénk vizsgálni, hogy mutatkozik-e eltérés a két, szöveganyagukat tekintve jelentősen különböző korpusz adataiban.

1 Bevezetés

Dolgozatunkban az *átjön/bejön*, valamint az *átmegy/bemegy* igekötős igepárok argumentumszerkezeteit vizsgáljuk a Szeged Korpusz [16] és a Dívány korpusz [10] adatai alapján. A dolgozatot a következőképpen építettük fel:

A vizsgálat első lépéseként két szerző vonatkozó eredményeit [13, 14, 5] tekintjük át, majd azokat összevetjük a magyar értelmező szótárak adataival. Az elméleti áttekintés után számítógépes nyelvészeti eszközökkel, korpuszadatok alapján vizsgáljuk meg az igéket. Ehhez először röviden bemutatjuk a felhasznált korpuszokat, majd a lekérdezéssel kapott eredmények alapvető statisztikai adatait. Azt is megvizsgáljuk, hogy mutatkozik-e valamiféle eltérés a két, szöveganyagukat tekintve jelentősen különböző korpusz adataiban. Végül a kapott eredményeket összevetjük a szótárak adataival és az elméleti irodalom tárgyalt megállapításaival.

2 Szakirodalmi áttekintés

Az igekötős igék vizsgálatával számos tanulmány foglalkozik, az elméleti és az alkalmazott nyelvészet oldalán egyaránt.

Az egyik elméleti nyelvészeti megközelítési mód az emberközpontúság fogalma felől magyarázza az igekötős igéket. Ez azt jelenti, hogy tárgyakat, élettelen dolgokat (az elmélet terminológiája szerint az ún. *nem viselkedők*et) *viselkedők*re jellemző tulajdonságokkal ruházunk fel (vö. [5]). E megközelítés alapján a világfelfogásunkat alapvetően a mozgás határozza meg, amit a nyelvhasználat is tükröz. Radden [7] szerint egy entitást úgy érzékelünk, mint amely mozgásban van akkor, amikor lokalitása érzékelhető, látható változáson megy keresztül. A mozgásnak van egy kezdőpontja és egy végpontja, ahová a kezdőponttól egy úton keresztül jutunk el, így adott az irányultság is (vö. [7]).

A fenti elmélettel összhangban van Lakoff és Johnson [6] mozgásértelmezése, amelyben a következő elemekkel operálnak a szerzők: *forrás, ösvény, cél* és *irányultság*. Úgy vélik, hogy "bármikor elmozdulunk valamerre, van egy hely, ahová érkezünk, egy összefüggő pontok sorozata, mely összeköti a kezdeti pontot a végponttal, valamint létezik irányultság" [6].

Szilágyi N. Sándor [11] ugyancsak hasonló megállapításokat tesz a magyar nyelv vonatkozásában. Véleménye szerint "a magyarban az igekötőnek talán a legfontosabb szemantikai szerepe éppen az, hogy megjelenítse, explicitté tegye a nyelvi világ térstruktúrájában való elmozdulásokat. Alkalmasint ezzel függ össze az is, hogy az igéket befejezett aspektusúvá teszi" [11].

A térbeli elmozdulás alapján az igekötők abban is különböznek egymástól, hogy a KEZDET–ÚT–VÉG sémából mit ragadnak meg. A *be-* és a *ki-* igekötők például a kezdetre vagy a végpontra vonatkoznak, a *fel-* és a *le-* igekötők az irányultságra, az *át-* igekötő utalhat a kezdetre és a végre (1a) vagy az út egy kiemelt helyére, ami által a mozgásban levő valamilyen relációba kerül (1b).

(1) a. Átment Európából Amerikába.
b. Átszökött a kerítésen.

A továbbiakban azokat a megállapításokat tekintjük röviden át, amelyek konkrétan az általunk vizsgált igekötős igékre vonatkoznak. A vizsgálat során elsősorban Szili Katalin [13, 14], valamint Imre Attila [5] dolgozataira támaszkodtunk, majd a megállapításokat összevetettük az értelmező szótárak adataival.

Szili Katalin [13, 14] azt vizsgálta, hogy milyen jelentésmódosulásokon keresztül válik az irányjelentéssel bíró *be-* igekötő (pl. *bemegy, bejön, befut*) tiszta perfektiváló elemmé (pl. *bevall, bebizonyít*). Elemzésében egyszerre használ szintaktikai és szemantikai vizsgálati szempontokat, és a különböző fogalmi típusokra tartályokként utal. Vizsgálatai kiindulópontját olyan tárgyatlan és tárgyas igékkel álló szintagmák képezik, amelyekben az igekötő klasszikus irányjelentésben szerepel. Ezeknek az igekötős igéknek a sajátossága, hogy az ige környezetében megjelennek a *be* vonzatának tekintett *hova?* kérdésre válaszoló *-ba/-be* ragos helyhatározók, hozzájárulva az igekötő irányjelentéséhez. Mellettük még olyan argumentumok is

megjelenhetnek, mint például a *-hoz,-hez,-höz* határozóragos nyelvi elemek vagy a *honnan?* kérdésre felelő, ablatívuszi esetragos bővítmények.

A *be-* igekötős igék kapcsán Szili [13, 14] az argumentumok minősége alapján, a belső tereknek négy nagy típusát különíti el: helyiség jellegűek (pl. ház, lakás, szoba), konténerszerű tárgyak (pl. doboz, táska, zseb), élő és élettelen testek (pl. ember, fa, kenyér), valamint csoportok, halmazok. Szili [13, 14] alapján az általunk vizsgált *bemegy* és *bejön* mozgásigék leginkább az első csoport elemeivel állnak. Tekintsük az alábbi példákat! (A példákban a vizsgált elemeket félkövér szedéssel emelem ki.)

(1) a. **bemegy** a házba
b. **bejött** a szobába

Ezeket a szerző [13, 14] a helyiség jellegű helyekkel álló igék csoportjának nevezi és megállapítja, hogy környezetükben a már említett *honnan?* és *hova?* kérdésekre felelő határozók fordulnak elő. Az így létrejövő vonzatstruktúrában megmarad a *be-* igekötő az eredeti irányjelentésében, és az alanyi argumentum bekerül valamilyen helyiségszerű térbe.

A különböző nyelvekben ugyanaz a tér máshogy jelenhet meg. Az olyanokat, mint például a *tér* a magyarban sík, lapszerű területként fogjuk fel, míg a parkot, erdőt vagy az utcát, amelyben házak állnak, zárt belső területként.

A „HIVATAL, INTÉZMÉNY, KÖZPONT" típust a szerző átmeneti csoportként definiálja. Átmeneti, hiszen képzeletünkben a hivatalok, intézmények és maguk az azoknak helyet adó épületek (a konkrét zárt helyek) összemosódnak. Ezeket vagy konkrét intézményként fogjuk fel, vagy csak az azoknak helyet adó épületekként gondolunk rájuk, például

(2) **Bejön** ma a főnök?

A Magyar nyelv értelmező szótára [1] ezt a példát a következőképpen magyarázza: "a beszélő megérkezik hivatalába, belső helyiségben lévő munkahelyére, ahol a beszélő is tartózkodik". Szili [13, 14] megközelítése tehát egybevág az értelmező szótár definíciójával: mind a két megközelítés kulcsfontosságúként emeli ki azt a tényt, hogy BENT helyről beszélünk.

Szintén zárt térként raktározzuk tudatunkban a minket körülvevő dolgokat, tulajdonainkat (vö. [13, 14]), például

(3) Csinos kis összeg **jött be** a gyümölcsből.

Ugyancsak zárt térként jelennek meg az emberek kisebb-nagyobb időleges csoportosulásai is, például

(4) A legtöbb tag 6-8 holddal **jött be** a szövetkezetbe.

Az *átjön/átmegy* igekötős igék vizsgálata során Imre Attila [5] eredményeire támaszkodtunk. Imre [5] alapján ahhoz, hogy „ÁT" helyzetről beszélhessünk, észlelési feltételekre van szükség. Észlelési feltételnek a következőket tekinti:

a) Adott a kezdet–út–vég” séma és a mozgás;
b) Feltételezünk egy azonos térben elhelyezkedő, ún. stabilt és egy ún. mobilt, amelyek egymáshoz viszonyított helye változik a mozgás során;
c) A viszonyítás alapja a stabil, ehhez képest a mobil közeledik felé, kapcsolatba kerül vele, majd megszűnik a viszony. Fontos, hogy a mobil a stabil által az úton kijelölt helynek egyik oldaláról a másikra kerül.

Imre [5] a következő stabil-típusokat különbözteti meg: anyagszerű stabilok, más jellegű stabilok (a rész-egész viszonyt kialakító), valamint az emberi test, mint stabil (lehet BENNE hely is, RAJTA hely is). Tekintsük Imre Attila az *átmegy/átjön* igekötős igékre vonatkozó megállapításait e rendszer alapján!

Az igekötős igék hordozhatnak konkrét, valamint metaforikus értelmet, és ezt gyakran a stabil természete határozza meg. Például, ha a mobil a stabil határvonalai között van, és az egyik vonaltól eljut a másikig, akkor konkrét értelmű kifejezésről beszélünk:

(5) **Átment** a szobán.

A példában a mobil a stabil határai között mozog.

A határátlépés egy speciális esetének tekinthetjük a színek közötti átmeneteket. Tekintsük az alábbi példát!

(6) **Átmegy** vörösből rózsaszínbe.

A példában jól érzékelhetően elkülönül a két végpont egymástól, azonban magát az átmenetet, tehát azt, hogy pontosan mi van a két végpont között, jóval kevésbé lehet meghatározni.

Lehetséges olyan eset is, amikor a mobil a stabil egy úgynevezett virtuális határvonalán jut át úgy, hogy nem érintkeznek, például

(7) Most **megy át** a város fölött.

Dichotómiás szerkezetekben (pl. kicsi-nagy, egyik-másik stb.) két stabilról beszélünk, amelyekkel a mobil érintkezik, például

(8) **Átment** a kicsi szobából a nagyba.

Mint ahogy korábban említettük (l. fentebb), Imre [5] szerint a stabilok nem csak anyagszerűek, hanem viselkedők is lehetnek. Ez azt jelenti, hogy egyes stabilok emberi tulajdonsággal felruházottak is lehetnek. Ilyen viselkedő stabilt látunk a következő példában:

(9) A bárányhimlő **átment** a testvéreire.

3 Az argumentumszerkezetek problémaköre a számítógépes nyelvészetben

Az igei argumentumszerkezetek vizsgálata mind elméleti, mind számítógépes nyelvészeti szempontból fontos probléma. A nyelvtechnológia oldalán azért is különösen jelentős, mert az igekötős igék argumentumszerkezetének automatikus elemzése még nem teljes mértékben megoldott, bizonyos részproblémáira megoldás még nem született. (A jelen dolgozatnak nem célja, hogy ezekkel a problémákkal foglalkozzon, részletesebben l. [7]).

Az alábbiakban csupán megemlítünk néhányat a magyar nyelvre vonatkozó nyelvtechnológiai kutatások és fejlesztések közül.

A vonzatkeretek egy fontos vizsgálati eszköze a Magyar Nemzeti Szövegtárban lévő, online elérhető kutatóeszköz, melynek segítségével az igei vonzatkereteket, az igei bővítményszerkezeteket tudjuk kvantitatívan tanulmányozni (vö. [8]). A vonzatkeretek elsajátításának modellezését és az igei vonzatkeretek gépi tanulásának más nyelvekre jól működő megközelítéseit, statisztikai módszereit alkalmazva (vö. [2]), a gépi tanulási görbéket a gyereknyelvi adatokon számított tanulási görbékkel vetették össze. A kutatás fő irányvonala az argumentumstruktúrák elsajátításának számítógépes modellezése volt.

A vonzatkeretek számának, a magyar igék hosszának és gyakoriságának összefüggését vizsgálták a Szeged Dependency Treebank (vö. [16]) rövid üzleti híreket tartalmazó alkorpuszának összegyűjtött adatai alapján (vö. [17]). Feltételezték, hogy a leggyakoribb igék rendelkeznek a legtöbb vonzatkeretekkel, a legrövidebb igék a leggyakoribbak és a legrövidebb igéknek vannak a legtöbb vonzatkeretei. A vizsgálatot kiterjesztették a teljeskörű vonzatkeretekre is, majd ezeket összehasonlították egy elméleti alapon épített vonzatkeret-lexikonban található vonzatkeretekkel.

4 A vizsgálati korpuszok bemutatása

A lekérdezésekhez a Dívány [10] és a Szeged korpuszt [16] alkalmaztuk. A vizsgálatunkhoz olyan szöveganyagokat kerestünk, amelyek szignifikánsan eltérnek egymástól, és az egyiket referenciának tekinthetjük. Ebből a megfontolásból tehát a sokféle szövegtípust tartalmazó Szeged Korpuszt [16], valamint a téma- és műfajspecifikus, termékvélemény-szövegekből álló Dívány korpuszt [10] választottuk.

A Dívány korpusz egy magyar nyelvű szentimentkorpusz, amelyet termékvélemény-szövegekből hoztunk létre kutatási és fejlesztési céllal [9]. A korpusz szöveganyagát a [http://divany.hu/] honlap termékvéleményeiből állítottuk össze. A korpusz fragmentum- és aspektusszintű annotációval rendelkezik [10]. Az adatbázis összesen 154 véleményszöveget, 17 059 mondatot és 251 202 tokent (központozással) tartalmaz.

A Szeged Korpusz egy kézzel annotált, 1,2 millió szavas magyar nyelvi korpusz. A korpusz létrehozásának egyik fő szempontja volt, hogy az a lehető legkülönbözőbb műfajú és témájú szövegeket tartalmazza, így jött létre az a hat különböző témakörből

álló adatbázis, témakörönként körülbelül 200 ezer szó terjedelemben. A Szeged Korpusz szépirodalmi szövegeket, 14-16 éves tanulók fogalmazásait, újságcikkeket, számítástechnikai témájú és szövegeket, valamint gazdasági és pénzügyi rövidhíreket tartalmaz.

5 A korpuszadatok elemzése

A Dívány és a Szeged Korpuszból a *jön/megy* igéket, valamint azok igekötős változatait kérdeztük le. Az eredmények az egyes igealakokat, valamint az azokhoz tartozó argumentumokat tartalmazták, az MSD-kódrendszer szerinti jelölésekkel. Az adatok „full valency" [14] értelemben, azaz az összes bővítményükkel szerepelnek. A „full valency" azt jelenti, hogy a kötelező és szabad bővítményeket egyaránt az igei vonzatkeret részének tekintjük [3].

Az eredmények feldolgozásánál külön-külön elemeztük a *jön/megy* igék adatait igekötő nélkül, valamint az *átjön–bejön* és *átmegy–bemegy* igéket is.

A lekérdezések eredményeit az alábbi diagram szemlélteti:

BEAM venngage.com/beam

1.Ábra: Az igék gyakorisága a vizsgálati korpuszokban

Ami a Dívány korpusz adatait illeti, az összesen 251 202 szövegszóban 343 alkalommal fordult elő a *jön* és a *megy* ige. Kézzel megvizsgáltuk a kapott adatokat, és azt tapasztaltuk, hogy a igék leggyakrabban átvitt értelemben szerepelnek a korpuszban, például

(10) Várom, hogy kiérezzek gyümölcsízeket, de nem **megy**.
(11) A második vonalban **jön** az enyhe, értelmezhetetlen mellékíz.

Amennyiben a *bejön* és az *átjön* igék előfordulásait a *bemegy* és az *átmegy* előfordulásaival vetjük össze, azt látjuk, hogy az előbbiek előfordulása jóval gyakoribb az utóbbi igepár előfordulásánál (megközelítőleg hatszor gyakrabban fordul elő). Ha az igekötők szerint vetjük össze az igepárok előfordulásait, érdekes eredményt kapunk: amíg a *bejön* jóval gyakoribb a *bemegy* igénél, addig az *át-* esetében éppen az ellenkezőjét figyelhetjük meg, tehát az *átmegy* lesz a gyakoribb a *bemegy* igénél.

A lekérdezés egy meglepő tapasztalata az volt, hogy az MNSZ2 adatai gyakorinak várt *bemegy* ige mindössze 1 adattal szerepelt a korpuszban. Nem meglepő módon azonban az *átjön* ige, amely az MNSZ2 korpuszban szintén csekély számmal fordult elő, a Dívány korpuszban is mindössze 1 adattal szerepelt:

(12) Olyan mintha tök náthás lennék, és esélytelen, hogy ízhez jussak, aztán végül mégis **átjön** valami.
(13) Ha **bemegyek** a boltba, biztos nem ezt vásárolom meg.

A Szeged Korpusz 1,2 millió szavas adatbázisa a *jön* és a *megy* igéket összesen 1599 alkalommal tartalmazta. Ha a két korpusz lekérdezési eredményeit összevetjük egymással, azt látjuk, hogy a vizsgált igék találati száma összességében arányos a korpuszok méretével.

A Szeged Korpuszban a vizsgált igekötős igék egymáshoz képesti gyakorisági adatai nem egyeznek a Dívány adataival. A Szeged Korpuszban ugyanis az *átmegy* és *bemegy* igék fordulnak elő nagyobb gyakorisággal az *átjön* és a *bejön* igéknél.

A két korpusz adatai közötti talán legszembeötlőbb eltérés az, hogy amíg a Dívány korpuszban a *bejön* ige a *jön* összes előfordulásának majdnem a 15%-át teszi ki, addig a Szeged Korpuszban ugyanez mindössze 4,64%. Ezt az eltérést véleményünk szerint akár a következő sajátság is okozhatja: mivel a Dívány korpuszt termékvélemények alkotják, a *bejön* igét átvitt értelemben, pozitív értékelés kifejezésére is gyakorta használhatják a szövegekben (pl. *nekem nagyon bejön ez a telefon!*). Feltevésünk igazolására megnéztük a Dívány korpuszban a *bejön* ige összes előfordulását, és azt tapasztaltuk, hogy a 30-ból 26 esetben használták az igét a hipotézisünknek megfelelően, tehát pozitív vélemény kifejezésére, például

(14) Savanykás, ez nekem jobban **bejön**, mint az eddigiek
(15) Kicsit édesebb, mint ami nekem **bejön**

Fontos megemlíteni, hogy a szótárak szócikkeiben nem szerepel a *bejön* ige esetében ez a jelentés, ugyanakkor azt látjuk, hogy a tényleges nyelvhasználatban

gyakori. A szótárak tehát ebben az esetben nem korrelálnak a valós adatokkal, ami azok kiegészítését, revízióját sürgeti.

Ezzel ellentéteset látunk akkor, ha a *bemegy* ige adatait vizsgáljuk. Ez ugyanis a Szeged Korpuszban a *megy* összes előfordulásának több mint a 10%-át teszi ki, míg a Díványban alig fordul elő.

Az *átjön* mindkét korpuszban nagyon ritka, legfeljebb csupán 1%-át teszi ki az *jön* ige összes előfordulásának. Az *átmegy* ige ugyanakkor kétszer olyan gyakran fordul elő a Szeged Korpuszban a Dívány adataihoz képest.

A továbbiakban azt vizsgáljuk, hogy a lekérdezett igék hány argumentummal fordulnak elő az egyes korpuszokban. Azt is megfigyeljük, hogy a *jön* és a *megy* igék argumentumszámán változtat-e, ha igekötővel állnak, és ha igen, hogyan módosítja azt.

A Dívány korpuszban a *jön* ige esetében a leggyakoribb az első három pozíció kitöltése, amely az összes előfordulás 58%-át teszi ki. A *jön* ige legmagasabb argumentumszáma 8, amely csupán egyetlen esetben fordul elő, és az ige *be-* igekötős alakjának az adata:

> (16) Így pedig nagyobb az esély arra, hogy hat, de nekem hosszú távon sajnos nem **jött be**, de ez egyéni dolog.

A 8 bővítményt tehát három kötőszó, egy szuperesszívuszi esetű főnév, egy határozói bővítmény, egy tagadószó, egy kérdőszó, valamint egy alanyi bővítmény adják.

A *megy* ige leggyakrabban 3 argumentummal fordul elő (44%). Az ige legmagasabb argumentumszáma 7, amely a *jön* ige esetéhez hasonlóan mindössze egyetlen egyszer fordul elő, és a *rá-* igekötős ige adataként szerepel. A 7 bővítményt ebben az esetben egy kötőszó, három határozói bővítmény, egy adesszívuszi és egy szublatívuszi esetű főnév, és egy időhatározói bővítmény adják.

Amennyiben az igekötős igék argumentumszámait vizsgáljuk, a következőket tapasztaljuk: A Dívány korpuszban a *bejön* ige legmagasabb argumentumszáma 6, azonban mind az 5, mind a 6 argumentum nagyon ritka, az összes előfordulásnak csupán a 10%-át tesz ki. Ugyancsak ritka a 3 és a 4 argumentummal való előfordulás. 26 esetben két argumentumú az ige, ami az összes eset (30) 86%-a. Ezt tekinthetjük tehát a leggyakoribb esetnek a Dívány korpusz alapján. Az *átjön* ige legmagasabb argumentumszáma 3, amelyek az első két kitöltöttségi helyen határozói, a harmadik helyen pedig alanyi bővítmények. Azt mondhatjuk tehát, hogy a *jön* ige esetében az igekötő megléte jelentősen redukálja az argumentumok mennyiségét.

A *bemegy* ige esetében ugyanazt tapasztaltuk, mint az *átjön* igénél. Legmagasabb argumentumszáma 3, amelyek rendre illatívuszi esetű főnévi, valamint határozói szófajú bővítmények. Az *átmegy* ige legmagasabb argumentumszáma 5, ez azonban csupán 2 esetben fordul elő. A *megy* ige esetében ugyanakkor az 5 argumentum kimondottan gyakori.

Ugyanezek a vizsgálatok a Szeged Korpuszon az alábbi eredményeket hozták: A *jön* és a *megy* igék legmagasabb argumentumszáma a 7 volt. Az alapigék esetében az argumentumszámok a Szeged Korpuszban nem mutatnak jelentős eltérést a Dívány korpuszban mért eredményeinkhez képest.

A *bejön* ige legmagasabb argumentumszáma 6, amely egyetlen egyszer szerepel a korpuszban. A Szeged Korpusz *bejön* igéjének adatai az argumentumszámokat illetően hasonlóak a Díványéhoz, itt is a két argumentum a leggyakoribb, amely a 79%-át teszi ki az összes előfordulásnak. Az *átjön* ige legmagasabb argumentumszáma 3, amely ugyancsak egyszer szerepel a korpuszban. Ezek az eredmények szintén egyezést mutatnak a Dívány korpuszban kapott eredményeinkkel.

Az *átmegy* és a *bemegy* igék esetében a legmagasabb argumentumszám 5. A bemegy igénél szintén a 2 argumentum a leggyakoribb, amely 89%-át teszi ki az összes előfordulásnak.

Ha összevetjük a *jön* és *megy* igéket azok igekötős alakjaival, azt tapasztaljuk, hogy ezek argumentumai a *bejön* ige kivételével jelentősen leredukálódtak.

Összevetve a két korpusz fentebbi adatait megállapítható, hogy mindkét korpuszban az igekötő megléte, illetve annak hiánya jelentősen változtatott az argumentumok számán. Mind a két korpuszban a *jön* és *megy* igék magasabb argumentumszámmal fordultak elő igekötős változataiknál, és az igekötős alakok átlagos legmagasabb argumentumszáma (a *bejön* ige kivételével) 3 volt.

A kutatás eredményei azt mutatják, hogy a vizsgálati korpuszok szövegműfaji sajátságai jelentősen nem befolyásolták az argumentumok számát. A vizsgált igék argumentumszámai ugyanis a két korpuszban nem mutattak szignifikáns eltérést, 4 esetben teljesen meg is egyeztek az adatok.

Azt is megnéztük, hogy az általunk választott igekötős igepárok mely vonzatokkal állnak a leggyakrabban. Azokat az igéket, amelyekre valamely korpuszban nagyon kevés az adat, itt nem elemezzük.

A Dívány korpuszban a *bejön* ige esetében azt tapasztaltuk, hogy a 30 előfordulásból 18 esetben kapcsolódott az igéhez datívuszi esetű argumentum (17a), míg a Szeged Korpuszban az összes 29 előfordulásból csupán 11 alkalommal (17b).

(17) a. Akinek ez **bejön,** azzal nagy gond van.
b. Szlovákiának **bejött** az euró.

Úgy véljük, ez a sajátság is összhangban áll korábbi megállapításunkkal, amely szerint a Dívány korpusz termékvélemény-szövegeiben a *bejön* igét átvitt értelemben, a tetszés kifejezésére használják a leggyakrabban.

A Szeged Korpuszban az *átmegy* ige 11 alkalommal állt illatívuszi raggal, ami az összes előfordulás (60) 18,33%-át adja, például

(18) Péter **átment** a szomszédba.

Ugyancsak 11 alkalommal társult hozzá allatívuszi rag, ami összhangban van a szakirodalom és a szótárak megállapításaival, miszerint az ÁT-igekötő jelölhet RAJTA és BENNE helyet is, például

(19) Anna **átment** Julcsihoz.

A Szeged Korpuszban a *bemegy* ige 39 alkalommal állt illatívusszal, ami 39%-át teszi ki az összes előfordulásnak, például

(20) Ha senki nem ér rá, ki **megy be** a boltba?

Elmondható tehát, hogy vizsgálati eredményeink ebben a vonatkozásban összhangban állnak a szakirodalmak és a szótárak definícióival. Egyedüli különbséget a *bejön* és ige pozitív értékelő funkcióban való használatát jelenti, amelyre sem a szakirodalmakban, sem a szótárak definícióiban nem találtunk példát.

6 Összegzés

Dolgozatunkban az *átjön/bejön*, valamint az *átmegy/bemegy* igekötős igepárok argumentumszerkezeteit vizsgáltuk a Szeged [16] és a Dívány korpusz [10] adatai alapján.

A vizsgálat első lépéseként a hazai elméleti nyelvészeti irodalom néhány fontosabb eredményét (vö. [13, 14, 5]) tekintettük át. Az elméleti áttekintés után lekérdeztük a vizsgált igekötős igéket a Dívány és a Szeged Korpuszból, és az adatok alapján feltártuk az igék nyelvhasználati sajátságait a valós szövegekben. Az elemzés keretében mind az előfordulási gyakoriságokat, mind a vonzatokra vonatkozó sajátságokat megvizsgáltuk, valamint összevetettük egymással a két korpusz adatait. Mindemellett a nyelvhasználati adatokat összevetettük a szakirodalmi eredményekkel is.

A vizsgálatok alapján a következő legfontosabb megállapításokat tettük: A Dívány korpuszban a *bejön* jóval gyakrabban fordul elő annál, mint ahányszor a másik vizsgálati korpuszban szerepel, emellett jelentősen nagyobb a frekvenciája az összes többi igekötős igénél is. Úgy véltük, hogy ez a Dívány korpusz sajátságát mutatja. A *bejön* igét ugyanis a termékvélemény-szövegekben valószínűleg gyakran használják átvitt értelemben, tetszés kifejezésére. Az adatok kézi ellenőrzése igazolta a hipotézisünket, ugyanis feltárta, hogy az összes előfordulásból 26 valóban ebben az értelemben szerepelt. Ez az eredmény különösen fontos, hiszen sem a szakirodalom, sem a szótárak szócikkei nem említik az igének ezt a jelentését.

A Szeged Korpusz kvantitatív adatai nem álltak összhangban a Dívány fentebbi adatival. A *bejön* és az *átjön* igékkel szemben ugyanis a *bemegy* és az *átmegy* igék előfordulása volt a gyakoribb. Az MNSZ2 [8] gyakorisági adatai alapján csekély számúnak várt *átjön* ugyanakkor mindkét korpuszban nagyon ritka, legfeljebb csupán 1%-át teszi ki a *jön* ige összes előfordulásának.

Megvizsgáltuk az argumentumok számát is az igekötős és igekötő nélküli igék esetében. Azt tapasztaltuk, hogy a *jön* és a *megy* igék alapvetően több argumentumokkal állnak igekötő nélkül, az igekötők ugyanis mind a két korpuszban és rendre minden ige esetében leredukálták az argumentumok számát. A két korpusz eltérő szöveganyaga azonban vizsgálataink alapján nem befolyásolta az argumentumok számát az egyes igék esetében. A *megy*, *bejön*, *átjön* és *átmegy* igék esetében a legmagasabb argumentumszámok a két korpuszban megegyeztek.

Azt is megnéztük, hogy az általunk választott igekötős igepárok mely vonzatokkal állnak a leggyakrabban. Azokat az igéket, amelyekre valamely korpuszban nagyon

kevés az adat, nem elemeztük. A *bejön* ige kapcsán a datívuszi esetű argumentum volt szembetűnő, amely az összes előfordulás 60%-ában volt megtalálható, míg a Szeged Korpuszban ugyanez csupán 37%-ot tett ki. Gyakori volt azonban az illatívuszi és allatívuszi esetű argumentum is, amely az *-át* igekötő esetében alátámasztotta a szakirodalmak meghatározásait.

A kutatás további tervezett lépései a következők: Egyrészt a jelen vizsgálatot szeretnénk más igepárokra, és lehetőség szerint más korpuszokra is kiterjeszteni. Másrészt, hogy hasonló, elméleti szempontú és korpuszalapú vizsgálatokat szeretnénk végezni az orosz nyelvre vonatkozóan is, majd a kapott eredményeket összevessük a magyar nyelvre vonatkozó eredményeinkkel.

Köszönetnyilvánítás

A jelen kutatást a Precognox Informatikai Kft., az Emberi Erőforrások Minisztériuma Új Nemzeti Kiválóság Programja, az MTA Társadalomtudományi Kutatóközpont „Lendület" RECENS Kutatócsoportja, az EFOP-3.6.1- 16-2016-00008 azonosítójú, EU társfinanszírozású projekt, valamint az European Research Council (ERC) támogatta az Európai Unió „Horizon 2020" nevű, 648693-as számú kutatási és fejlesztési programja keretében. Vincze Veronika kutatásait az Emberi Eröforrások Minisztériuma ÚNKP-17-4 kódszámú Új Nemzeti Kiválóság Programja támogatta.

Bibliográfia

1. A magyar nyelv értelmező szótára. I. Akadémiai Kiadó, Budapest. (1978)
2. Babarczy A., Serény A., Simon E.: Magyar igei vonzatkeretek gépi tanulása. In: Tanács A., Szauter D., Vincze V. (szerk.) VI. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem, Szeged. (2009) 333-342.
3. Čech, R., Pajas, P., & Mačutek, J.: Full valency. Verb valency without distinguishing complements and adjuncts. Journal of Quantitative Linguistics, 17(4), (2010) 291–302.
4. É.Kiss K.: Egy igekötőelmélet vázlata. Magyar Nyelv, 100. (2004) 15-43.
5. Imre A.: Az ÁT viszony jelentésszerkezete a magyar nyelvben. Egyetemi szakdolgozat. Kézirat. Kolozsvár. (1999)
6. Lakoff, G.,Johnson, M.: Metaphors We Live By. Chicago, London, The University of Chicago Press. (1980)
7. Lázár B., Szabó M.K.: Mozgást jelentő igék argumentumszerkezetének korpuszalapú vizsgálata, XII. Alkalmazott Nyelvészeti Doktoranduszkonferencia, (2018) absztrakt benyújtva
8. Oravecz Cs., Váradi T., Sass B.: The Hungarian Gigaword Corpus. In: *Proceedings of LREC 2014,* 2014.
9. Radden, G.: Motion Metaphorized. The Case of Coming and Going, In: Studies in Linguistics III, A Supplement to the Hungarian Journal of English and American Studies, Ed.by Béla Korponay and Péter Pelyvás, Debrecen. (1994)
10. Sass B.: Vonzatkeretek a Magyar Nemzeti Szövegtárban. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005). Szegedi Tudományegyetem, Szeged. (2005) 257-264

11. Szabó M.K., Vincze V.: Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai. In: Tanács A.,Varga V., Vincze V.: (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015). Szegedi Tudományegyetem, Szeged. (2015) 219-226
12. Szabó M.K.,Vincze V.,Hangya V.: Aspektusszintű annotáció és szentimentet módosító elemek egy magyar nyelvű szentimentkorpuszban. In: Tanács A.,Varga V., Vincze V. szerk. XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016). Szegedi Tudományegyetem, Szeged. (2016) 174–182.
13. Szabó M.K., Vincze V., Simkó K., Varga V., Hangya V:. A Hungarian Sentiment Corpus Manually Annotated at Aspect Level. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Portoroz, Szlovénia. (2016). 2873-2878.
14. Szilágyi N.S.: Hogyan teremtsünk világot?, Erdélyi Tankönyvtanács, Kolozsvár (1996)
15. Szili K.: A be igekötő jelentésváltozásai I. Magyar Nyelvőr 129. (2005) 151-164.
16. Szili K.: A be igekötő jelentésváltozásai II. Magyar Nyelvőr 129. (2005) 282-299.
17. Tolcsvai N.G.: Az igekötő + ige szerkezet szemantikája. Nyelvtudományi Közlemények 109. (2013) 187-226.
18. Vincze V., Szauter D., Almási A., Móra Gy., Alexin Z., Csirik J.: *Hungarian Dependency Treebank.* In: Proceedings of the Seventh Conference on International Language Resources and Evaluation. European Language Resources Association, Valletta, Málta (2010) 1855-1862.
19. Vincze V.: Valency Frames in a Hungarian Corpus. Journal of Quantitative Linguistics, 21:2. (2014) 153-176.
20. Forrás: http://www.divany.hu

A szöveg mint skálafüggetlen hálózat

Makrai Márton és Sass Bálint

MTA Nyelvtudományi Intézet
{makrai.marton,sass.balint}@nytud.mta.hu

Kivonat Cikkünkben a szöveget egy egyszerű konstrukció segítségével skálafüggetlen hálózatként ábrázoljuk, és megvizsgáljuk, hogy a hálózatelmélet sztenderd eszközei mit mondanak az ilyen módon reprezentált szövegről.
Kulcsszavak: hálózatok, skálafüggetlen hálózatok, köztiség, Zipf, gyakoriság, szöveg

1. Bevezetés

Azokat a hálózatokat, ahol a befokszámok Zipf-eloszlást követnek, *skálafüggetlen hálózat*oknak nevezzük [1] [2, 703. oldal]. Régóta tudjuk [3], hogy a szöveg szavainak gyakorisági eloszlása Zipf-eloszlású. Tanulmányunkban azt az élsúlyozott irányított gráfot vizsgáljuk, melynek csúcsai a szóalakok, a $\langle w_1, w_2 \rangle$ él súlya pedig a $w_1 w_2$ bigram gyakorisága. Az előbbiek miatt ez a gráf skálafüggetlen.

Ebben a bevezető szakaszban egy játékpéldán szemléltetjük a konstrukciót, majd bemutatjuk a kapcsolódó irodalmat és a HITS linkelemzési algoritmust, amivel a leglátványosabb eredményeket kaptuk. A következő szakaszban mutatjuk be a kísérleti eredményeket.

1.1. Gráf bigramokból

Az a feladat, hogy találjunk egy konstrukciót, melynek segítségével a szövegbeli szógyakoriság éppen a befokszám lesz egy alkalmasan megalkotott gráfban. Így skálafüggetlen hálózatot kapunk.

Az ötlet nagyon egyszerű: a hálózat csomópontjai a korpusz szavai lesznek, menjünk végig a korpuszon és minden szó esetén rajzoljunk be egy (1 súlyú) nyilat a hálózatba, ami az adott szóból indul ki és őt közvetlenül követő szóhoz vezet. Ha egy szópár többedszerre fordul akkor rajzoljunk be még egy nyilat, vagy ami ezzel ekvivalens, a nyíl súlyához adjunk hozzá 1-et. Az 1. ábrán látható hálózatot kapjuk. Ez egy olyan gráf lesz, melyben a csomópontok szavak, és az $A \to B$ él akkor létezik, ha van AB bigram a korpuszban, és az él súlya AB bigram gyakorisága.

A fenti reprezentáció tehát skálafüggetlen hálózatot eredményez. Ez azért nagyon jó, mert a skálafüggetlen hálózatok elméletének az elmúlt évek során kidolgozott összes eszközét, módszerét, mérőszámát [4] alkalmazni tudjuk rá.

Bízunk benne, hogy ezzel az eszköztárral valami újat tudunk mondani a szövegről, új módon tudjuk megragadni a szöveg bizonyos jellegzetességeit.

1. ábra: „A nagyobb nagyobb a nagynál vagy a nagy nagyobb a nagyobbnál.” példamondat ábrázolása. A dupla nyilat ábrázolhatjuk egy 2-es súllyal bíró szimpla nyíllal is.

1.2. Kapcsolódó irodalom

A mienkhez leghasonlóbb kutatás alighanem a TextRank [5]. Mihalcea és Tarau az irányított, 2-széles ablakkal végzett kutatásaikról azt írják, hogy rosszabb eredményeket hoztak, mint a az irányítatlan eset. Azt a meglepő következtetést vonják le, hogy a szövegnek nincs természetes iránya. Ha nyelvi adatból készült skálafüggetlen gráfról beszélünk, nem kerülhető meg [6] sem, ők azonban szemantikus hálókat vizsgálnak, míg mi magából a korpuszból vonunk le első sorban szintaktikai tanulságokat.

A mi konstrukciónk lényegesen eltér attól a bevett megközelítéstől, mely esetében akkor húzunk be egy (irányítatlan!) élt két szó között, ha egy közös trigramban megtalálhatók, más szóval, ha az egyik a másiktól (jobbra vagy balra) 1 vagy 2 szó távolságra van [7]. Ennél a modellnél a szöveg természetes balról jobbra rákövetkezése nincs reprezentálva, szemben a mi modellünkkel, ahol viszont lényegi elem. Emiatt a most bemutatott modell várhatóan kevésbé szemantikai, inkább szintaktikai jellemzőket tud majd megragadni. A kutatás feltáró alapkutatás jellegéből adódóan az eredmények esetleges majdani alkalmazása nem témája jelen cikknek.

A következő alszakaszban bemutatandó HITS algoritmus, amivel a leglátványosabb eredményeket kaptuk. nagyjából egyidős a skálafüggetlen gráfok ma népszerű fogalmával. Az utóbbi évtizedekben természetesen számos kutatás vizsgált skálafüggetlen gráfokat a HITS segítségével [8].

1.3. HITS

A HITS (Hyperlinkindukált témakeresés, *Hyperlink-Induced Topic Search* vagy hubok és tekintélyek, *hubs and authorities*) egy linkelemzési algoritmus [9], körülbelül egyidős a PageRankkel [10], csak persze sokkal kevésbé elterjedt. Az az alapötlete, hogy a fontos internetes oldalak kétfélék. A *hubok*, mint az index.hu

vagy a vajdasag.lap.hu, nagy linkgyűjteményként működnek: a rajtuk magukon megjelenő információnak nincs tekintélye, viszont más, hiteles oldalakra, a *tekintélyekre* irányítják a felhasználókat. A hubok és a tekintélyek definíciója kölcsönös: jó hub egy olyan oldal, amely sok tekintélyes oldalra mutat, nagy tekintélyük pedig azoknak az oldalaknak van ebben a modellben, melyekre számos jó hub mutat.

A hub- és tekintélyérték számítása iteríatíve történik. Kezdetben a számokat tetszés szerint választjuk (például minden oldalnak ugyanazt), majd minden iterációban egy oldal mértékadósága a rá mutató oldalak hubértékének összege lesz, a hubérték pedig a lap által mutatott oldalak tekintélyének összege. Az iterációk között a hub- illetve tekintélyértékek négyzetösszegét normalizáljuk. Ezzel az algoritmussal kaptuk a leginkább szembeötlő eredményünket.

2. Eredmények

Vizsgálatainkat[1] az MNSZ2 [11] véletlenül választott 1000 illetve 10000 mondatán végeztük. Az elemzéseinkhez a NetworkX python csomagot használtuk [12].

2.1. Erős összefüggőség

Az első nagyon egyszerű kérdés, hogy erősen összefüggő-e a gráf, azaz minden szóból elérhető-e irányított úton az összes többi szó. Lényegében erősen összefüggő lesz a gráf, esetleg a korpusz elején és végén lévő hapax szavakból álló „farok” fordulhat elő, ami megbontja az erős összefüggőséget (ld. az 1. ábrán a *nagyobbnál* szót).

2.2. Kisvilág tulajdonság

A kisvilág-szerkezet szemléltetésére idézzük Karinthy 1929-es *Láncszemek* című novelláját:

> Tessék egy akármilyen meghatározható egyént kijelölni a Föld másfél milliárd lakója közül, bármelyik pontján a Földnek – [a társaság egyik tagja fogadást ajánlott], hogy legföljebb öt más egyénen keresztül, kik közül az egyik neki személyes ismerőse, kapcsolatot tud létesíteni az illetővel, csupa közvetlen – ismeretség – alapon

A meglepően kis távolság, melyet úgy formalizálhatunk, hogy az L átlagos távolság csak logaritmikusan növekszik a csúcsok számában, nem a skálafüggetlen gráfok sajátja: Erdős–Rényi-féle véletlen gráfoknál is fennáll [13], ha a élek beválasztását kontrolláló p elég nagy ahhoz, hogy az egész hálózat összekapcsolódjon. (Vegyük észre a definícióban, hogy önmagában egy gráf kisvilág tulajdonságáról nem beszélhetünk, csak egy olyan gráfsorozat esetében, ahol a csúcsszám

[1] `https://github.com/makrai/TextBetweenness.git`

végtelenhez tart.) A kisvilág-szerkezet jellemzésében az alacsony C klaszterezési együtthatót is meg szokták követelni [14], de ennek az irányított gráfokra való általánosítása nem tűnik triviálisnak, ezért tanumányunkban a legrövidebb utak hosszából számított statisztikákra szorítkozunk. A mi 7 K-csúcsú, 13 K-élű gráfunkban az átlagos távolság $L = 4.89$. Az aszimptotikus tulajdonságot a poszteren elemezzük, amit a projekt repójában talál meg az olvasó.

2.3. Skálafüggetlenség

A skálafüggetlen gráfok definiáló tulajdonsága, hogy a fokszámok hatványeloszlást követnek [1]. Ezt nálunk Zipf törvénye miatt biztosítja az, hogy (az első és az utolsó szó kivételével) a be- és a kiélek súlyösszege egyaránt megegyezik a szónak a korpuszban való gyakoriságával. Az elméletileg garantált tulajdonságot statisztikailag is ellenőriztük. A fokszámeloszlást exponenciális eloszlással összehasonlítva 133-as likelyhood-ratiót és $5.91 \cdot 10^{-6}$-os szignifikanciaszintet kaptunk. A Zipf-együttható 2.20-nak adódott, ami összhangban van azzal, hogy az angolban 1.25 körülre teszik [15], a magyar pedig gazdagabb morfológiájú, így a Zipf-együttható is magasabbnak várható.

2.4. Távolságok

Egy másik gráfmérték is hasznosnak tűnik az együtt-előfordulási gráf elemzésében, a csúcsok különcsége (*eccentricity*), vagyis az adott csúcsból az összes többi csúcsba vezető legrövidebb utak hosszának maximuma. Míg a többi vizsgálatot tízezer mondatos mintán végeztük, az ebben a pontban említetteket csak ezer mondatoson, mert az összes (rendezett) csúcspárra ki kell hozzá számolni a legrövidebb út hosszát, aminek nagy az időkomplexitása. A *sugár* (a legkisebb különcség) nálunk 9, az *átmérő* (a maximális különcség) 19. Szemléletesen *közép*nek *(center)* hívják azokat a pontokat, amelyeknek a különcsége megegyezik a sugárral. Esetünkben egy ilyen csúcs van, a sok funkcióban használatos vessző (,) token.

2.5. Closeness centrality

A közelségi központiság (*closeness centrality*) mértéke (mely az adott csomóponttól az összes többi csomópontba vezető legrövidebb utak hosszának átlagaként adódik) egy érdekes jelenséget mutat (2. ábra). A szavak egységes eloszlásban helyezkednek el. Az eloszlásból néhány olyan elem lóg ki, amely „nem illeszkedik a magyar szövegbe”: ilyen az egyenlőségjel és egy HTML entity (`|`), illetve két angol szó (a *the* és az *of*), melyek előfordulnak a korpuszban. Ezeknek a tokeneknek tehát kisebb a közelségi központiság értékük annál, mint amit gyakoriságuk alapján várnánk. A kilógó elemek pontos karakterizálásához további vizsgálat szükséges.

2.6. HITS

A bigramgyakorliságokból épített gráfban egy szó hubértéke az őt követő szavak tekintélyének összege, egy szó tekintélyét pedig az őt megelőző szavak együttes hubértéke adja. A 3. ábrán első körben azt látjuk, hogy a gyakoribb (nagyobb fokszámú) szavaknak nagyobb a tekintélye – ez nem meglepő. Érdekesebb, hogy a kötőszavak és az igék jól elkülönülnek: az egész szókincsre jellemző bal lent – jobb fent átlótól inkább följebb vannak a kötőszavak, és ettől az átlótól inkább lejjebb vannak az igék. Azaz a kötőszavak tekintélye nagyobb annál, ami a gyakoriságuk alapján várható, az igéké pedig kisebb. (Tehát nem egyszerűen arról van szó, hogy a kötőszavak gyakorisága kiemelkedően magas, hiszen az azonos gyakoriságú igék és kötőszavak is jól elválnak egymástól.)

Az eredményre a következő intuitív magyarázatot javasoljuk: a kötőszavak nagy tekintélye azt jelzi, hogy az őket megelőző tokenek együttes hubértéke nagy. Mivel a szavak hubértéke az őket követő szavak tekintélyének összege, az ördögi körből úgy léphetünk ki, ha a kötőszavak előtti tokenek után más tekintélyes szavak is megjelennek. A gráf tehát tükrözi azt, hogy a kötőszavak olyan helyen állnak, ahol a balkontextus alapján sok más szó is állhat. A tipikus példa ilyen balkontextusra a már a center kapcsán is említett vessző token. Az igék esetében épp ellenkező a helyzet: az igék előtt megjelenő szavak (igemódosítók és bővitményi frázisok utolsó szavai) jobban determinálják, hogy igének kell következnie, mint más szavak balkontextusa az adott szót.

3. Összefoglalás

Az irányított skálafüggetlen hálózattá alakított szöveget érintő első vizsgálatainkban feltérképeztük, hogy az egyes hálózatelméleti eszközök mit mondanak erről a hálózatról, milyen értékek a jellemzőek. Legérdekesebb kezdeti eredményünk az, hogy úgy tűnik, hogy a HITS algoritmus képes egymástól elválasztani bizonyos szócsoportokat, és ezek a csoportok összefüggésben vannak a szófajokkal.

Köszönetnyilvánítás

Sass Bálint kutatásait az MTA Bolyai János Kutatási Ösztöndíja támogatta (ügyszám: BO/00064/17/1; időtartam: 2017-2020).

Hivatkozások

1. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science **286**(5439) (1999) 509–512
2. Kovács, L., Orosz, K., Pollner, P.: Magyar szóasszociációk hálózata. Magyar Tudomány **173**(6) (2012) 699–705
3. Zipf, G.K.: The Psycho-Biology of Language; an Introduction to Dynamic Philology. Houghton Mifflin, Boston (1935)

4. Barabási, A.L.: Scale-free networks: A decade and beyond. Science **325** (2009) 412–413
5. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing. (2004)
6. Steyvers, M., Tenenbaum, J.B.: The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive science **29**(1) (2005) 41–78
7. i Cancho, R.F., Solé, R.V.: The small world of human language. Proceedings of The Royal Society of London. Series B, Biological Sciences **268** (2001) 2261–2266
8. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web, ACM (2007) 221–230
9. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46** (1999) 604–632
10. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999)
11. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014), Reykjavík (2014)
12. Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA USA (2008) 11–15
13. Erdos, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5**(1) (1960) 17–60
14. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. nature **393**(6684) (1998) 440–442
15. Kornai, A.: Mathematical Linguistics. Advanced Information and Knowledge Processing. Springer (2008)

2. ábra: A szavak eloszlása *closeness centrality* szerint.

3. ábra: A HITS algoritmus eredménye. A nagyobb piros ponttal jelölt kötőszavak balra fent (magasabb authority), a nagyobb zöld ponttal jelölt igék jobbra lent (alacsonyabb authority) helyezkednek el a fokszám (gyakoriság) vs authority grafikonon.

Az érzelmi epizódok narratív konstrukciója és az érzelmi intelligencia érzelemszabályozás komponense közötti kapcsolat

Pólya Tibor[1], Puskás Éva[2]

[1]Magyar Tudományos Akadémia, Természettudományi Kutatóközpont,
Kognitív Idegtudományi és Pszichológiai Intézet, Pf.: 286
1519 Budapest, Magyarország
{polya.tibor@ttk.mta.hu}
[2]Károli Gáspár Református Egyetem, Pszichológiai Intézet,
1037 Budapest, Magyarország
{puskase6@gmail.com}

Tanulmányunkban az érzelmi intelligencia szövegelemzés alapú mérési eljárásnak kidolgozását célul kitűző kutatásunk első eredményeit mutatjuk be. Az érzelmi intelligencia az érzelmi vonatkozású információ feldolgozásában való jártasságot jelenti. Feltételezésünk az, hogy az érzelmi intelligencia szintje azonosítható az érzelmi epizódokról beszámoló történetek konstrukciójában. Feltételezésünket empirikus vizsgálatban ellenőriztük. Vizsgálati személyeinktől érzelmi epizódok elbeszélését kértük, történeteiket a Narratív Kategoriális Tartalomelemzővel elemeztük. Az érzelmi intelligencia mérésére az Érzelemszabályozás Szituációs Tesztet alkalmaztuk. Eredményeink azt mutatják, hogy a történetkonstrukció automatikus elemzése biztosítja a kritérium és konvergens validitás teljesülését az érzelmi intelligencia érzelemszabályozás komponensének mérése során.

1 Bevezetés

A pszichológia területén az automatikus szövegelemzési eljárásokat elsősorban feltáró vizsgálatokban alkalmazzák. A feltáró vizsgálatok elsődleges célja, hogy kimutassák a kapcsolat meglétét egy pszichológiai konstruktum vagy jellemző és a szöveg automatikus eljárásokkal elemzett jellemzője vagy jellemzői között. A feltáró vizsgálatok jellemzően nem foglalkoznak azzal a kérdéssel, hogy a szöveg jellemzőinek elemzése alapján mérni lehet-e az adott pszichológiai konstruktumot vagy jellemzőt. A mérés negligálásának egyik oka nagy valószínűséggel az, hogy a pszichológiai szövegelemzést domináló paradigma, a pszichológiai tartalomelemzés rendszerint nem fordít elég figyelmet a vizsgált pszichológiai konstruktumok validitásának tisztázására. Azaz a pszichológiai tartalomelemzés módszere rendszerint nem alkalmaz olyan mérő eljárásokat – például kérdőíveket –, amelyek révén empirikusan igazolható lenne, hogy az elemzett szövegjellemzők valóban kapcsolatba hozhatók a feltételezett pszichológiai konstruktummal vagy jellemzővel. A mérés lehetőségének ki nem aknázása általános gond a pszichológiai tartalomelemzésben, de

azokban az esetekben, amikor szövegelemzés automatikus eljárásokkal történik még hangsúlyosabban jelentkezik ez a hiányosság. Ennek magyarázata az, hogy a pszichológiai tartalomelemzés területén az egyes szerzők által használt kódrendszerek még abban az esetben is mutatnak kisebb-nagyobb mértékű különbséget, amikor ugyanannak a pszichológiai konstruktumnak a vizsgálatát célozzák meg. Ezért minden egyes kódrendszert validitását külön-külön kellene elvégezni. A számítógépes szövegelemző eljárásokban ugyanakkor kevesebb a variáció. Amennyiben kidolgozásra kerül egy automatikus szövegelemzési eljárás, azt rendszerint változatlan formában használják, így ezen elemzési eljárások validálását is könnyebb lenne elvégezni, és ezzel jelentős lépést lehetne tenni a szövegelemzésre építő mérési eljárások kidolgozása felé a pszichológia területén.

2 Pszichológiai konstruktumok mérése szövegelemzés alapján

A pszichológia területén ugyanakkor vannak olyan széles körben használt szövegelemzésre építő eljárások, amelyeket mérési, sőt sok esetben diagnosztikus céllal használnak. Ilyen például a Rorschach-próba [1] és a Tematikus Appercepciós Teszt [2]. Mindkét tesztben bizonytalan és határozatlan ingeranyagot kap a személy – a Rorschach-próba esetében „tintafoltokat", a Tematikus Appercepciós Teszt esetében fényképszerű jeleneteket, amelyekről szabad szöveget kell produkálnia a személynek. A Rorschach-próba esetében arra a kérdésre kell válaszolni a személynek, hogy mit lát a képeken, a Tematikus Appercepciós Teszt esetében pedig egy olyan történetet kell elmesélnie, amelybe beleilleszthető a képen látott jelenet. A szövegek kódolásához mindkét teszt esetében részletes útmutatók állnak rendelkezésre, és a kódolás eredménye alapján feltérképezhető a válaszadó személyiségének struktúrája.

A pszichológia területén alkalmazott számítógépes szövegelemzési eljárások között azonban csak néhány olyan elemzési eljárás került kidolgozásra, amely fontosnak tartja a validálás lépését és ennek révén pszichológiai mérési eljárásként is használható. A legrészletesebben kidolgozott ilyen mérési eljárás a Gottschalk és Gleser által létrehozott „Psychiatric Diagnosis and Content Analysis" nevű kódrendszer [3]. Ez a pszichológiai tartalomelemző eljárás számos pszichológiai jellemző mentén jellemezni a személy pszichológiai állapotát, mint például a szorongás, az ellenségesség és a társas elidegenedés. A tartalomelemzési eljárást a szerzők előbb kézi kódolási eljárásként dolgozták ki majd az ezredforduló környékén automatizálták elemzési eljárásukat. A magyar nyelvre kifejlesztett automatikus elemzési eljárások közül a Narratív Kategoriális Tartalomelemző [4,5] az az eljárás, amelynek kidolgozása során kiemelt cél volt az, hogy felhasználható legyen pszichológiai állapotok mérésére. Ez az automatikus tartalomelemző eljárás a történetek kompozíciós kategóriái mentén elemzi a történeteket, és a történetet elbeszélő személy személyes vagy szociális identitás állapotának jellemzésére használható.

Mindkét automatikus tartalomelemző eljárás kidolgozása során számos empirikus vizsgálatban ellenőrizték az elemzési eljárások validitását és ezzel megteremtették annak alapját, hogy mérési eljárásként is felhasználhatók legyenek [6,7]. Bár a

Narratív Kategoriális Tartalomelemzővel végzett korábbi kutatások elsősorban az identitás állapot fogalmához kapcsolódnak, ez nem zárja ki annak lehetőségét, hogy más pszichológiai konstruktumok vizsgálatára is felhasználható legyen ez az elemzési eljárás. A 2017-es MSZNY konferencián bemutatott elemzés szerint a Narratív Kategoriális Tartalomelemző a döntés utáni helyzetben keletkező kognitív disszonancia állapotának kimutatására is alkalmazható [8]. Az itt bemutatásra kerülő vizsgálatban egy újabb pszichológiai konstruktumhoz, az érzelmi intelligencia fogalmához kapcsolódóan alkalmazzuk ezt az elemzési eljárást.

3 Érzelmi intelligencia és történetkonstrukció

Az érzelmi intelligencia konstruktuma az érzelmi vonatkozású információ feldolgozásában való jártasságot jelenti. A pszichológiai kutatások kétféleképpen határozzák meg ezt a fogalmat. Az egyik elképzelés szerint az érzelmi intelligencia azon képességek együtteseként határozható meg, amelyek részt vesznek az érzelmi vonatkozású információ feldolgozásában [9]. Az érzelmi intelligenciát képességként meghatározó legjelentősebb megközelítés négy csoportba sorolja a képességeket. Ezek a képesség csoportok a következők: az érzelmek percepciója, integrációja, megértése és szabályozása. A négy képességcsoport mindegyike számos képességet foglal magába. Az érzelmek szabályozása például a következő képességeket öleli fel. Az a képesség, hogy nyitottak legyünk az érzésekre, kellemesekre és kellemetlenekre egyaránt. Az érzelmek monitorozásának és mérlegelésének képessége. Az a képesség, hogy létrehozzunk, fenntartsunk egy érzelmi állapotot, vagy éppen eltávolodjunk tőle, attól függően, hogy az állapotot mennyire ítéljük informatívnak vagy hasznosnak. Az a képesség, hogy kezeljük mások érzéseit, illetve a saját érzéseinket. A másik elképzelés szerint az érzelmi intelligencia személyiségjellemzőként határozható meg [10].

A két, egymással éles vitában álló megközelítés közötti eltérés abban is jól megragadható, hogy hogyan gondolkodnak az érzelmi intelligencia méréséről. Az érzelmi intelligenciát személyiségjellemzőként meghatározó kutatók szerint az érzelmi intelligenciát önbeszámolós kérdőívekkel tudjuk mérni [11]. A kérdőívek kitöltése során a személyeknek olyan kérdésekre kell válaszolniuk, amelyek megválaszolása révén felmérhetjük, hogy bizonyos tulajdonságok, viselkedések milyen mértékben jellemzőek rájuk. Az érzelmi intelligenciát képességként meghatározó kutatók szerint viszont nem használhatunk kérdőíveket az érzelmi intelligencia mérésére, ehelyett – a klasszikus intelligencia mérésével megegyezően –, úgy kell eljárnunk, hogy olyan feladatot adunk a személynek, amelynek a megoldása során felhasználja az érzelmi intelligencia fogalmába tartozó képességeit. Ebben az esetben a feladatot megoldásának színvonala alapján tudjuk megmérni a személy érzelmi intelligenciájának szintjét [12]. Számos kritériumnak kell megfelelnie annak a feladatnak, amely alkalmas arra, hogy a megoldás színvonalából következtetni tudjunk a személy érzelmi intelligenciájának szintjére [13]. A kritériumok közül kettőt emelek ki. Egyrészt biztosnak kell lennünk abban, hogy a feladat megoldásához igénybe kell venni az érzelmi intelligencia fogalmába sorolt képességeket. Másrészt a feladat

megoldásának színvonalát objektív kritériumok alapján kell tudni értékelni. Nem könnyű egyszerre megfelelni ennek a két kritériumnak. A képesség alapú érzelmi intelligencia mérésére legelterjedtebben használt Mayer Salovey Caruso Érzelmi Intelligencia Teszt [12] például két olyan feladattal méri az érzelmi intelligencia érzelemszabályozás komponensét, amelyben a kitöltőnek arról kell döntést hoznia, hogy egy érzelmeket kiváltó helyzetben mely cselekvést tekinti hatékonynak. Az újabban kidolgozott szituációs tesztek [14] ehhez nagyon hasonlóan érzelmileg és a személyek közötti kapcsolatokat tekintve is nehéz szituációk sorozatát írják és a kitöltőnek az a feladata, hogy négy lehetőségből válassza ki azt a reakciót, ami véleménye szerint a leghatékonyabb választ jelenti az adott helyzetben. Az egyik tétel például a következő helyzetet írja le: Mindig János foglalkozott egy bizonyos ügyféllel, de most egy nagyon összetett feladat kapcsán a főnök ezt a munkát egy kollegára bízta. János csodálkozik, hogy a főnöke azt hiszi, ő nem tudja kezelni ezt a fontos munkát. Mi lenne János részéről a leghatékonyabb reakció? (a) Hinni abban, hogy jól teljesít és, hogy a következő ilyen komplex munkát ő kapja. (b) Jól dolgozni, hogy a jövőben ő kapja ezeket a komplex feladatokat. (c) Megkérdezni a főnökét, hogy miért a kollégának adta ezt a munkát. (d) Nem aggódni emiatt.

A történetkonstrukció és érzelmi intelligencia fogalma közötti szoros kapcsolat feltételezése elméleti megfontolásokon alapul. Egyrészt mivel a történetek rendszerint egy probléma keletkezését és a probléma megoldására tett kísérleteket foglalják magukba, könnyen belátható, hogy a történetek tartalmának jelentős részét érzelmek adják, hiszen a probléma keletkezése, azok megoldása vagy éppen a megoldás meghiúsulása érzelmeket indukál a történet szereplőiben [15,16,17]. A történetek tartalmát adó érzelmekben való eligazodás azt is feltételezi, hogy a történet elbeszélője a társas világra vonatkozó gazdag tudással rendelkezzen [18]. Másrészt a történetek elbeszélőjének valamilyen érzelmi viszonyban kell lennie a történetbe foglalt érzelmi élményekkel kapcsolatban [19,20]. E nélkül az érzelmi viszony nélkül a történetek meglehetősen unalmasak lennének.

3.1 Hipotézis

Általános hipotézisünk az, hogy szoros kapcsolatot találunk a történetkonstrukció és az elbeszélő személy képességként meghatározott érzelmi intelligencia szintje között. Ezen belül az itt bemutatásra kerülő vizsgálatban ennek a hipotézisnek azt a részét teszteljük empirikusan, amely a történetkonstrukció és a képességként meghatározott érzelmi intelligencia érzelemszabályozás komponense között áll fenn. Vizsgálatunk kiterjed a kritérium validitás ellenőrzésére: azaz a képesség alapú érzelmi intelligencia érzelemszabályozás komponensét kérdőíves eljárással [13] mérjük, és ennek kapcsolatát keressük a történetkonstrukció jellemzőivel. Emellett megvizsgáljuk azt is, hogy a történetkonstrukció elemzése esetén teljesül-e a konvergens validitás feltétele is. A konvergens validitás ellenőrzése során azt várjuk el, hogy a történet konstrukciója kapcsolatban legyen olyan pszichológiai konstruktumokkal, amelyekről tudjuk, hogy összefüggésben állnak a képességként meghatározott érzelmi intelligenciával. Két ilyen konstruktum szerepel vizsgálatunkban, a klasszikus intelligencia és a depresszió. Mindkét konstruktumról tudjuk, hogy kapcsolatban van a

képességként meghatározott érzelmi intelligencia szintjével [21], és ez alapján azt várjuk, hogy kapcsolatban lesz a történet konstrukciójával is. Mindezek alapján elvárásunk az, hogy a történetkonstrukció a legszorosabb kapcsolatot az érzelmi intelligencia érzelemszabályozás komponensével fogja mutatni, és ennél a kapcsolatnál gyengébb összefüggéseket fogunk találni a klasszikus intelligencia és a depresszió szintjével. Emellett azt is várjuk, hogy a történetkonstrukció különböző összetevői kapcsolódnak az érzelmi intelligencia érzelemszabályozás komponenséhez egyfelől és a klasszikus intelligenciához illetve a depresszió szintjéhez másfelől. További elvárásunk az, hogy a történetkonstrukció és az érzelemszabályozás komponens közötti kapcsolatot nem befolyásolják olyan további tényezők, amelyekről azt gondoljuk, hogy függetlenek a képességként meghatározott érzelmi intelligenciától, így például a személyiség általános jellemzőitől és a verbális képesség szintjétől.

4 Vizsgálat

4.1 Vizsgálati személyek

A vizsgálatban 52 személy vett részt. Két személyt kizártunk a vizsgálatból mivel nem töltötték ki a kérdőíveket. Így végül 50 fő került a vizsgálati mintába. Közülük 11 férfi és 39 nő volt. A legfiatalabb résztvevő 18, a legidősebb 30 éves volt. A vizsgálati személyek átlagéletkora pedig 24,5 év volt (SD=3,3).

4.2 Kérdőívek

A képesség alapú érzelmi intelligencia érzelemszabályozás komponensének mérése az Érzelemszabályozás Szituációs Tesztet [13,14] alkalmaztuk. A klasszikus intelligenciát a Raven Standard Progresszív Mátrixok teszttel [22], a depressziót a CES-D kérdőívvel [23], a személyiség jellemzőit a „Big Five" kérdőívvel [24], és végül a verbális képességet az Általános Képesség Teszt [25] verbális próbájával mértük.

4.3 Szövegkorpusz

A vizsgálati személyeket négy érzelmi epizód felidézésére és elmesélésére kértük. Az érzelmi epizódok a következők voltak: emlékezetes szilveszter, stresszes egyetemi vizsga, megismerkedésük egy jó barátjukkal, valamint a legutóbbi alkalom, amikor konfliktusba kerültek valakivel. A történetek elmesélése után került sor a tesztek illetve kérdőívek felvételére.

Az elmesélt történeteket diktafonnal rögzítettük, majd lejegyeztük azokat. Az így kapott 200 történet teljes terjedelme 29 653 szó volt. A történetek átlagos hossza pedig 148,3 szó volt (SD=69,9). A történetek elemzését a Narratív Kategoriális

Tartalomelemzővel [4,5] végeztük el, amely a következő kompozíciós kategóriák mentén elemzi a történeteket: szereplő (egyes szám vagy többes szám első személyű névmás vagy személyrag), cselekvés (aktív vagy passzív jelentésű igével kifejezve), pszichológiai perspektíva (mentális és érzelmi állapotra vonatkozó szavak és kifejezések), értékelés (értékelést kifejező szavak vagy kifejezések). Az érzelmi állapotok és értékelések esetén az elemzés elkülöníti a pozitív és negatív valenciát is. A legutolsó komponens a téridői perspektíva volt, amely három formát különböztet meg a megnyilatkozás deiktikus centruma és a megnyilatkozások tartalmát adó események téridői elhelyezése alapján.

4.4 Eredmények

A Narratív Kategoriális Tartalomelemzővel kapott eredményeket relatív gyakorisági adatokká alakítottuk át. A szereplőkre, a cselekvésre, a pszichológiai perspektívára, az értékelésre és a tagadásra vonatkozó találatok számát a történet szószámban mért terjedelmével osztottuk el. A téridői perspektíva esetén, mivel az elemzési egység a tagmondat, a találatok számát a történet tagmondatban mért terjedelmével osztottuk el. A kompozíciós kategóriák relatív gyakoriságának átlagát és standard devianciáját az 1. Táblázat mutatja be.

1. Táblázat: A kompozíciós kategóriák relatív gyakoriságának átlaga és standard devianciája

Kompozíciós kategóriák	**Átlag %**	**Standard deviancia**
Szereplő		
E1	,68	,32
T1	,15	,19
Cselekvés		
Aktív ige	,17	,17
Passzív ige	,14	,16
Pszichológiai perspektíva	,89	,50
Gondolkodás	,30	,23
Érzelem	,59	,48
Érzelem negatív	,31	,05
Érzelem pozitív	,28	,12
Értékelés	,49	,25
Értékelés negatív	,15	,14
Értékelés pozitív	,33	,18
Tagadás	1,11	,45
Téridői perspektíva		
Visszatekintő forma	23,5	13,3
Átélő forma	69,6	14,5
Metanarratív forma	6,9	5,7

Hipotézisünk teszteléséhez első lépésben korrelációs elemzést végeztünk. Mivel a kompozíciós kategóriák relatív gyakorisága nem normális eloszlású volt, a Spearman-féle korrelációs eljárást alkalmaztuk. A kompozíciós kategóriák relatív gyakorisága és az érzelmi intelligencia érzelemszabályozás komponense, a klasszikus intelligencia és a depresszió szintje között talált korrelációs együtthatókat a 2. Táblázat foglalja össze. Mivel minden vizsgálati személytől 4-4 történetünk volt, az eredmények értelmezésekor Bonferroni-féle korrekciót hajtottunk végre és így csak a legfeljebb p = 0.01 szignifikancia szintű korrelációs együtthatókat vettük figyelembe.

A korrelációs elemzés eredménye azt mutatja, hogy az érzelmi intelligencia érzelem szabályozás komponense a történet téridői perspektívájával van a legszorosabb kapcsolatban. A visszatekintő forma relatív gyakoriságával kapott korrelációs együttható értéke r=.29. Ugyanez a mutató az átélő forma esetében r=-.24, a metanarratív forma esetében pedig r=-.23. Az érzelmi intelligencia érzelemszabályozás komponense szoros kapcsolatban van a történet pszichológiai perspektívájával is. A pszichológia perspektíva relatív gyakoriságával kapott együttható értéke r=.28. Ezen belül a gondolkodás (r=.20) és a pozitív érzelmi kategóriák (r=.17) relatív gyakoriságával mutat összefüggést az érzelemszabályozás komponens szintje.

2. Táblázat A kompozíciós kategóriák és az érzelmi intelligencia (ÉI) érzelemszabályozás komponense, a klasszikus intelligencia és a depresszió szintje közötti korrelációs együtthatók (p < .01 szinten szignifikáns eredmények bolddal jelölve)

Kompozíciós kategóriák	**ÉI: Érzelem-szabályozás komponens**	**Klasszikus intelligencia**	**Depresszió**
Szereplő			
E1	-.01	**.19**	-.03
T1	.03	**.21**	-.12
Cselekvés			
Aktív ige	-.04	.11	.06
Passzív ige	.01	-.08	**.24**
Pszichológiai perspektíva	**.20**	**.28**	.00
Gondolkodás	**.18**	.11	**-.22**
Érzelem	.04	**.19**	.01
Érzelem negatív	.06	.10	.13
Érzelem pozitív	**.17**	**.20**	.00
Értékelés	-.13	**-.19**	-.12
Értékelés negatív	-.06	**-.17**	**.20**
Értékelés pozitív	.00	-.08	.08
Tagadás	.01	.10	.04
Téridői perspektíva			
Visszatekintő forma	**.29**	**.17**	.14
Átélő forma	**-.24**	-.14	**-.18**
Metanarratív forma	**-.23**	-.13	.02

A klasszikus intelligencia és a kompozíciós kategóriák közötti kapcsolatok részben átfedésben vannak az érzelmi intelligencia érzelemszabályozás komponensével kapott összefüggésekkel, de vannak olyan kapcsolatok, amelyek csak a klasszikus intelligencia szintjével jelentkeznek. A klasszikus intelligencia szintje is kapcsolatban van a pszichológiai perspektíva relatív gyakoriságával (r=.28) és a pozitív érzelmek (r=.20) relatív gyakoriságával is, de emellett az érzelem kategóriák (r=.19) előfordulásának gyakoriságával is. A téridői perspektíva esetében csak a visszatekintő forma relatív gyakoriságával kaptunk értelmezhető szintű korrelációs együtthatót (r=. 17). Mindezek mellett a klasszikus intelligencia szintje a történet két szereplő kategóriájával is kapcsolatban van. Az egyes szám első személyű szereplő kategória esetén a korrelációs együttható értéke r=.19, a többes szám első személyű szereplő kategória esetén pedig r=.21. Továbbá, a klasszikus intelligencia szintje kapcsolatban van az értékelés kategória előfordulásával is. Az értékelés kategória relatív gyakoriságával kapott korrelációs együttható értéke r=-.19, ezen belül a negatív értékelések relatív gyakoriságával kapott korrelációs együttható értéke r=-.17.

A depresszió szintje szintén összefüggésben van a történet kompozíciós kategóriáinak előfordulásával. A kapott összefüggések itt is részben megegyeznek, részben pedig különböznek az érzelemi intelligencia érzelemszabályozás komponensével kapott összefüggésekkel. Megegyező kapcsolatok a téridői perspektíva átélő formájának relatív gyakoriságával kapott összefüggés (r=-.18) és a pszichológiai perspektíva gondolkodás komponensével kapott összefüggés (r=-.22). Utóbbi esetben azonban eltérő az összefüggés előjele. Különböző kapcsolatok pedig a cselekvés kategóriáján belül a passzív igék előfordulási gyakoriságával kapott összefüggés (r=.24) és a negatív értékelések előfordulási gyakoriságával kapott összefüggés (r=.20).

A korrelációs elemzés során megnéztük azt is, hogy az érzelmi intelligencia érzelemszabályozási komponense és a kompozíciós kategóriák közötti összefüggésre hatással vannak-e a személyiségjellemzők és a verbális képesség szintje. A függetlenség vizsgálatához parciális korrelációs eljárást használtunk. A személyiségjellemzők illetve a verbális képesség kiparciálása nem volt hatással a téridői perspektíva kategóriájával kapott összefüggésekre, ugyanakkor a pszichológiai perspektíva kategóriájával korábban kapott összefüggések eltűntek. Az érzelemszabályozási komponens parciális korrelációs együtthatói a következőképpen alakultak. A visszatekintő perspektíva forma relatív gyakoriságával r=.21 (p=.001), az átélő forma relatív gyakoriságával r=-.14 (p=.030), és végül a metanarratív forma relatív gyakoriságával r=-.20 (p=.003). A pszichológiai perspektíva relatív gyakoriságával r=.04 ($p < .10$), a gondolkodás relatív gyakoriságával r=.10 (p=.086), és végül a pozitív érzelem relatív gyakoriságával r=-.02 ($p < .10$).

A statisztikai elemzés második lépéseként regresszióanalízist végeztünk, hogy megállapítsuk, a történetek téridői perspektívája hogyan predikálja az érzelmi intelligencia érzelemszabályozás komponensének szintjét. A regresszióanalízisbe a téridői perspektíva három formájának relatív gyakoriságát foglaltuk bele prediktor változóként. A téridői perspektíva három formájának relatív gyakoriságát egy-egy vizsgálati személy 4 történetében átlagoltunk. A változók kezelésére „stepwise" beléptetési módszerrel alkalmaztunk. Az érzelmi intelligencia érzelemszabályozás komponensének szintje volt a regresszióanalízis függő változója. A regresszióanalízis

eredménye szerint csak a visszatekintő perspektíva forma előfordulási gyakorisága predikálja szignifikánsan az érzelemszabályozás komponens szintjét (Adjusted R^2=.104, $F(1,49)$=6.708, p=.013), a másik két perspektíva forma relatív gyakoriságának predikciós ereje nem volt szignifikáns. Minél több visszatekintő perspektíva fordul elő a történetben, annál magasabb az érzelemszabályozási komponens szintje (β=.35, t=2.59, p=.013).

4.5 Diszkusszió

A vizsgálatunk eredményei megerősítik azt a hipotézisünket, hogy az érzelmi epizódokról beszámoló történetek konstrukciója összefüggésben van az érzelmi intelligencia érzelemszabályozás komponensével. A kritérium validitás szempontját megerősítően azt találtuk, hogy az érzelemszabályozás szintje a legerőteljesebben a történetek téridői perspektívájával van összefüggésben. Az érzelemszabályozás szintje pozitívan korrelál a visszatekintő perspektíva forma használatával, és negatívan korrelál az átélő és metanarratív formák használatával. Az eredmények a történetkonstrukció elemzésének konvergens validitását is megerősítik, hiszen azt találtuk, hogy a történetkonstrukció azokkal a konstruktumokkal is kapcsolatban van, amelyek kapcsolódnak a képességként meghatározott érzelmi intelligenciához. Két ilyen konstruktumot vizsgáltunk a klasszikus intelligenciát és a depresszió szintjét. Mindkét konstruktum esetében jellemzően gyengébb kapcsolatokat találtunk és azt is láttuk, hogy van eltérés abban, hogy a történetkonstrukció mely komponenseivel mutat összefüggést az érzelmi intelligencia érzelemszabályozás komponense egyfelől és a klasszikus intelligencia és a depresszió szintje másfelől. Végül azt is igazolják eredményeink, hogy a képességként meghatározott érzelmi intelligenciától független konstruktumok nem befolyásolják a téridői perspektíva és az érzelemszabályozás komponens között kapott összefüggéseket.

Eredményeink összességükben azt mutatják, hogy sikerült azonosítani a történetkonstrukció azon részét, ami egyértelműen mutatja az érzelmi intelligencia érzelemszabályozási komponensének szintjét. A téridői perspektíva szerepére vonatkozó eredményeink összhangban vannak a korábbi vizsgálati eredményekkel [15,26,27,28], amelyek szintén azt mutatják, hogy az érzelmi élményekre való visszatekintés az érzelmi élmények hatékony szabályozását jelzi. Ezzel szemben az érzelmi élményekbe való bevonódás – amelyet az átélő és a metanarratív perspektíva formák is kifejezhetnek – az érzelemszabályozás kevésbé hatékony működését jelzik.

Vizsgálatunkban empirikusan igazoltuk a történetkonstrukció automatikus nyelvi elemzésére építő módszer kritérium és konvergens validitását. ezekből az eredményekből az a következtetés vonható le, hogy a történetkonstrukció automatikus nyelvi elemzésének módszerét felhasználhatjuk pszichológiai konstruktumok mérésére. Egy megbízhatóan alkalmazható mérési eljárás kidolgozásához azonban még további vizsgálatok szükségesek. Egyrészt a korrelációs együtthatók értéke bár szignifikáns, de viszonylag gyenge kapcsolatot mutatnak. Ezért a téridői perspektíva részletesebb elemzésére van szükség ahhoz, hogy azonosítani tudjuk, hogy a téridői perspektíva használatának melyek azok a jellemzői, amelyek erősebbé tehetik ezt a kapcsolatot. Másrészt jelen vizsgálatunk a képességként meghatározott érzelmi

intelligencia egyetlen komponensét vizsgálta, a további vizsgálatoknak ki kell terjednie az itt nem vizsgált komponensekre is.

A történetkonstrukció kompozíciós kategóriáinak automatikus elemzésére építő mérési eljárás kidolgozása számos olyan előnnyel rendelkezik, amely kedvezően befolyásolhatja egy ilyen mérési eljárás széleskörű felhasználását. Egyrészt a történetek elmésélésére építő mérési eljárásnak magas az ökológiai validitása. Ez azt jelenti, hogy a történetek révén olyan helyzetben mérhetjük a személyek érzelmi intelligenciáját, amilyen helyzetben azt a személyek elterjedten használják hétköznapjaikban is. Vagyis nem kell a személyeket egy, a természetestől sok szempontból különböző helyzetbe hozni ahhoz, hogy megmérhessük érzelmi intelligenciájuk szintjét. Másrészt a történetek automatikus elemzésére építő mérés objektív eredményeket ad és ráadásul gyorsan el is végezhető. Mindezek alapján azt gondoljuk, hogy az itt bemutatott kezdeti lépések után érdemes tovább folytatni az érzelmi intelligencia automatikus szövegelemzésre építő mérési eljárásának kidolgozását.

Köszönetnyilvánítás

A tanulmány elkészítését az NKFIH K 124206 számú kutatási projektje támogatta.

Bibliográfia

1. Mérei F. A Rorscach-próba. Budapest, Medicina. (2002)
2. Murray, H. A. Tematikus Appercepciós Teszt, Budapest, OS Hungary. (1995)
3. Gottschalk, L. A. & Gleser, G. C. The measurement of psychological states throught the content analysis of verbal behavior. Berkeley, University of California Press. (1969)
4. Ehmann B., Csertő I., Ferenczhalmy R., Fülöp É., Hargitai R., Kővágó P., Pólya T., Szalai K., Vincze O., László J. Narratív kategoriális tartalomelemzés: a NARRCAT. In Tanács A., Varga V., & Vincze V. (Szerk.), X. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2014. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport. 136–147. (2014)
5. László J., Csertő I., Ferenczhalmy R., Fülöp É., Hargitai R., Péley B., ... Ehmann B. Narrative language as expression of individual and group identity: The Narrative Categorical Content Analysis. Sage Open. April-June, 1-12. (2013)
6. László J. A történetek tudománya. Bevezetés a narratív pszichológiába. Budapest: Új Mandátum. (2005)
7. László J. Történelem történetek. Bevezetés a nartatív szociálpszichológiába. Budapest, Akadémiai Kiadó. (2012)
8. Pólya T. A kognitív disszonancia narratív markereinek azonosítása termékleírásokban. In: Vincze Veronika (szerk.) XIII. Magyar Számítógépes Nyelvészeti Konferencia. 219-227. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport. (2017)
9. Mayer, J. D. & Salovey, P. What is emotional intelligence? In Salovey, P. & Sluyter, D. (Eds). Emotional development and emotional intelligence. (pp. 3-31.) New York: Basic Books. (1997)

10. Petrides, K. V. Ability and trait emotional intelligence. In Chamorro-Premuzic, T., Furnham, A., & von Stumm, S. (Eds.), The Blackwell-Wiley Handbook of Individual Differences. (pp. 656-678.). New York: Wiley. (2011)
11. Petrides, K. V., & Furnham, A. Trait emotional intelligence: Psychometric investigation with reference to established trait taxonomies. European Journal of Personality, 15(6), 425-448. (2001)
12. Mayer, J. D., Salovey, P., & Caruso, D. R. Emotional intelligence test. CD ROM version. Neddham, Virtual Knowledge. (1997)
13. Nagy H., Magyaródi T., & Séllei B. A képességalapú érzelmi intelligencia: új paradigmák a tesztfejlesztésben és pontozásban. Hazai tapasztalatok az érzelemmegértés és érzelemszabályozás szituációs tesztekkel. Magyar Pszichológiai Szemle, 70(4), 827-846. (2015)
14. MacCann, C., & Roberts, R. D. New paradigms for assessing emotional intelligence: theory and data. Emotion, 8(4), 540. (2008)
15. Bodor P. Az érzelmek társas konstruktivista értelmezése és az érzelmek aktív versus passzív paradoxona. In Boros G. & Pólya T. (Szerk.). Szenvedély, szerelem, narrációk. Filozófiai és pszichológiai tanulmányok. (pp. 41-60). Budapest: Eötvös Kiadó. (2014)
16. Pérez, J. C., Petrides, K. V., & Furnham, A. Measuring trait emotional intelligence. Emotional intelligence: An international handbook, 181-201. (2005)
17. Schank, R. C. Tell me a story. A new look at real and artificial memory. New York: Macmillan Publishing Company. (1990)
18. Schank, R. C. & Abelson, R. P. Knowledge and memory: The real story. In R. S. Wyer, Jr., (Ed.). Knowledge and memory: The real story. Advances in social cognition. Vol. 8. (pp. 1-85.). Hillsdale: Lawrence Erlbaum Associates Publishers. (1995)
19. Labov, W. Language in the inner city. Oxford: Blackwell. (1972)
20. Labov, W., & Waletzky, J. Narrative analysis: Oral version of personal experience. In: Helm, June (ed.): Essays on the Verbal and Visual Arts. American Ethnological Society, Seattle, 12-44. (1967)
21. Nagy H. A Salovey-Mayer-féle érzelmi intelligencia modell érvényességének elemzése. Magyar Pszichológiai Szemle, 67(1), 105–124. (2012)
22. Raven, J. C. Raven's progressive matrices. Oxford: Oxford Psychologists Press. (1998)
23. Szeifert, L. (2010) Depresszió és életminőség krónikus veseelégtelenségben szenvedő betegek körében, Doktori értékezés, Semmelweis Egyetem
24. Rózsa, S, Kő, N, Oláh, A. Rekonstruálható-e a Big Five magyar mintán: A Carpara-féle „Big Five Kérdőív" (BFQ) felnőtt változatának hazai adaptációja és nemzetközi összehasonlító elemzése. Magyar Pszichológiai Szemle, 26(1), 57-76.Rumelhart, D. E. (1988).Megjegyzések egy történetsémáról. In Kanyó Z. & Síklaki I. (Szerk.). Tanulmányok az irodalomtudomány köréből. (330-348). Budapest: Tankönyvkiadó. (2006)
25. Smith, P., & Whetton, C. GAT: General Ability Tests manuale. Budapest, O.S. Hungary. (1996)
26. Berán E. A narratív perspektíva szabályozó szerepe a terápiás diskurzusban. Doktori Disszertáció. Eötvös Loránd Tudományegyetem Pedagógiai és Pszichológia Kar, Budapest. (2009)
27. Pólya T. Az érzelmi intelligencia narratív indikátorai. Psychologia Hungarica Caroliensis, 3(2), 23-38. (2015)
28. Pólya T., Kis, B., Naszódi, M. & László, J. Narrative perspective and the emotion regulation of a narrating person. Empirical Text and Culture Research, 7(3), 50-61. (2007)

A magabiztosság-krízis index alkalmazása angol nyelvű megnyilatkozásoknál és rossz minőségű felvételeknél

Puskás László[1], Pólya Tibor[2]

[1]Pécsi Tudományegyetem Bölcsészettudományi Kara, Pszichológia Doktori Iskola
laszlopuskas@gmail.com
[2]Magyar Tudományos Akadémia, Természettudományi Kutatóközpont,
Kognitív Idegtudományi és Pszichológiai Intézet
1117 Budapest, Magyar tudósok körútja 2.
polya.tibor@ttk.mta.hu

Kivonat: Tanulmányunkban korábbi kutatásaink eredményeit kívánjuk megvizsgálni angol nyelvű megnyilatkozásokon és rossz minőségű felvételeken. 2011-ben került bemutatásra a Magyar Számítógépes Nyelvészeti Konferencián az az új narratív pszichológiai eljárás, amelyben összekapcsoltuk a narratív pszichológiai tartalomelemzést és a vokális mintázatok pszichológiai „tartalomelemzését" [8]. Ennek az eljárásnak a részeként mutattuk be a magabiztosság-krízis indexet, amely a megnyilatkozás nyelvi, tartalmi elemeit és az elhangzottak fonetikai struktúráját vizsgálva von le következtetéseket a közlő lelkiállapotára vonatkozóan. Azt a feltételezésünket igyekeztünk adatokkal is alátámasztva igazolni, hogy a krízishelyzet nyelvi-fonetikai mintázata jól körülhatárolható, és ezen jegyek alapján a közlő lelkiállapotára vonatkozóan pszichológiailag értékelhető megállapítások tehetők. Vizsgálatunk nyelvi anyagát akkor Shakespeare Lear királya első és utolsó monológjának magyar nyelvű változata alkotta. 2014-ben a Magyar Számítógépes Nyelvészeti Konferencián mutattuk be a magabiztosság-krízis skála spontán megnyilatkozásokon történő alkalmazását [7], 2016-ban pedig olasz nyelven vizsgáltuk Lear megnyilatkozásait [11]. Eddigi kutatásainkhoz képest jelentős áttörést jelent, hogy kidolgoztunk egy becslési eljárást, amely lehetővé teszi, hogy rosszabb minőségű felvételeket is bevonjunk az elemzési körbe. Korábban csak stúdióminőségű felvételek elemzésére volt lehetőség, ami a kidolgozott eljárás alkalmazását nagyon szűk keretek között tette lehetővé, és ezzel az eljárás széleskörű alkalmazásának lehetőségét is jelentősen akadályozta. A kidolgozott becslési eljárással azonban most már szinte bármilyen – jól érthető, de rossz minőségű – hanganyag vizsgálat alá vonható.

1 Bevezetés

Identitásunk és pszichológiailag érvényes valóságunk számos elemét a történetek révén alkotjuk meg. A történetek vallanak az elbeszélők megküzdési képességeiről és várható viselkedési adaptációjáról. A tudományos narratív pszichológia az elbeszélést komplex pszichológiai tartalmak hordozójának tekinti, melynek tanulmányozása révén eredményesen vizsgálható az emberi társas alkalmazkodás. Azt a szoros kapcsolatot

hangsúlyozza, amely a pszichológiai folyamatok, az elbeszélés és az identitás között van [5].

A narratív pszichológia írott szövegek elemzésével foglalkozik, még azokban az esetekben is, amikor a kutatások hanganyagát rögzítették. A rögzített hanganyagok kizárólag azt a célt szolgálták, hogy lejegyezzék a szöveget. Mivel a fonetikai vizsgálat nem volt célja ezeknek a kutatásoknak, így az is elég volt, ha a hanganyag lejegyezhető minőségű, így ezeken a felvételeken korábban valószínűleg egyébként sem lehetett volna ilyen vizsgálatokat lefolytatni. A tudományos narratív pszichológia elemzési kereteinek bővítése fonetikai elemek vizsgálatával még nem tekint vissza hosszú múltra, de az eddig lefolytatott vizsgálatok eredményei megerősítik a keretek bővítésének szükségességét, és újabb távlatokat nyithatnak. [7, 8, 9, 10, 11]

2 A Lear monológok pszichológiai vizsgálatának előzményei

Shekespeare Lear királya már számos irodalomtörténész és pszichológus képzeletét megmozgatta. Pennebaker és Ireland [6] a magabiztosság és a krízis jegyeit vizsgálták Lear első és utolsó monológjában, és megállapították, hogy mind a krízis, mind pedig a magabiztosság nyelvi jegyei jól körülhatárolhatók. Vizsgálatukban Lear mellett egy létező személy, Giuliani New York-i polgármester krízis előtti és utáni megnyilatkozásait is megvizsgálták, és azt tapasztalták, hogy közel négyszáz év távlatából is, a magabiztosság és a krízis jegyei beazonosíthatók, mind Lear, mind pedig Giuliani esetében.

Jelenlegi és korábbi kutatásainkban Lear király első és utolsó monológjában mi is a magabiztosság és a krízis jegyeit vizsgáltuk a Pennebaker és Ireland [6], valamint a László János és munkatársai [4] által kidolgozott módszer segítségével, kiegészítve azt a krízis fonetikai jegyeinek vizsgálatával, a színész modellálta helyzetben. Létrehoztuk a magabiztosság-krízis indexet, amelynek összetevői a nyelvi tartalmi elemeket és a fonetikai jegyeket egyaránt tartalmazzák. [7, 8, 10, 11]

Az első eredmények bemutatására 2011-ben került sor, a Magyar Számítógépes Nyelvészeti Konferencián [8]. Azóta több tanulmányunk is megjelent a témában [7, 10, 11]. Megtörtént az eljárás során alkalmazott index finomhangolása, spontán megnyilatkozások vizsgálata, valamint idegen nyelvű hanganyagokat vizsgálva bizonyítottuk, hogy a magabiztosság és a krízis mintázata nyelvfüggetlen.

Ez a tanulmány a legújabb kutatási eredményeinkről számol be. Egyrészt megerősíti, hogy az eljárás nyelvtől függetlenül is lefolytatható, két angol nyelvű Lear-előadás vizsgálatával, másrészt bemutatunk egy becslési eljárást, amellyel rossz minőségű megnyilatkozások is vizsgálhatók. Ez jelentős áttörés lehet a vizsgálati anyagok felhasználhatósága szempontjából, hiszen az eddigi vizsgálatok stúdió minőségű hanganyagokat igényeltek, ezzel meglehetősen leszűkítve a vizsgálati lehetőségeket. Az index alkotóelemein is finomítottunk. A sokéves tapasztalatok és eredményeink alapján kijelenthető, hogy a magabiztosság-krízis index nyelvfüggetlenül alkalmazható, és eredményes eszköz arra, hogy a közlő lelki egyensúlyára vonatkozóan megállapításokat tehessünk.

3 A vizsgálat

3.1 A vizsgálati anyag

Vizsgálatunk nyelvi anyagát egyrészt Lear király 1998-ban készült angol nyelvű televíziós adaptációjából választottuk ki, amelyben Lear király szerepét Ian Holm játszotta, Richard Eyre rendezésében [13]; másrészt Lear király 1983-ban készült, szintén angol nyelvű adaptációjából, amelynek főszerepét Laurence Olivier játssza, Michael Elliot rendezésében [12]. A korábbi vizsgálathoz hasonlóan Lear első és utolsó monológját elemeztük. A szöveg nyelvezetének eltéréseit is figyelembe véve, nem törekedtünk teljes megfelelésre sem az angol, sem pedig a más nyelven elemzett változatok között. Mivel teljesen azonos krízist mutat be valamennyi változat, feltételeztük – ahogy a magabiztosság-krízis indexszel lefolytatott vizsgálat eredményeinél is –, hogy a krízisjegyek a teljes szövegbeli egyezéstől függetlenül megjelennek. Természetesen nem azt vártuk, hogy az index értékei azonosak lesznek a korábbi vizsgálatban kapott értékekkel, hanem – ahogy az olasz és a magyar nyelvű változat összehasonlításánál is – azt, hogy az értékek a skála „megfelelő" pólusán helyezkednek majd el.

A Ian Holm-féle változat hanganyaga csak nagyon rossz minőségben állt rendelkezésünkre. Komoly fejtörést okozott, hogy mihez tudunk kezdeni a vizsgálati anyaggal, hiszen korábban gyakorlatilag stúdió minőségű felvételekkel dolgoztunk, és még az esetlegesen meglévő zajok, hanghibák szűrésének lehetőségére is felkészültünk módszertanilag. Amikor a kérdés felmerült, nemcsak azzal kellett szembenéznünk, hogy a meglévő hanganyagot meg tudjuk-e vizsgálni, hanem azzal is, hogy a hasonlóan rossz minőségű felvételeket be tudjuk-e vonni a vizsgálati körbe, valamint azzal, hogy a jó és a rossz minőségű felvételekből kapott eredményeket össze tudjuk hasonlítani.

Azzal, hogy korábban stúdió minőségű hanganyagokat használtunk fel gyakorlatilag nemcsak jól elemezhető vizsgálati anyagokkal dolgoztunk, hanem az amúgy is gyéren hozzáférhető felvételek számát egyben minimálisra is csökkentettük. Ennek legfőbb oka az volt, hogy a fonetikai paraméterek vizsgálatához jó minőségű felvételekre van szükség, melyek hiányában nem végezhetők el a vizsgálatok. A kérdés tehát az volt, hogy gyenge minőségű felvételeken le tudjuk-e folytatni a korábbi vizsgálatainkat, és ha igen, akkor ezt milyen módszerrel tudjuk kivitelezni.

A Laurence Olivier-féle hanganyag kiváló minőségben állt rendelkezésre, viszont ennél is felmerült egy kérdés, amely tisztázásra szorult. A második monológban halk aláfestő zene szól, ami a felvétel érthetősége, kezelhetősége szempontjából nem jelent gondot, viszont a korábban használt módszertanunk alapján, nem használhattuk volna fel a hanganyagot.

3.2 Módszer

3.2.1 A fonetikai paraméterek vizsgálata

Az érzelmi állapotok fonetikai paraméterekre gyakorolt hatásának tanulmányozásakor a témában korábban lefolytatott vizsgálatainkat használtuk fel [7, 8, 10, 11], azzal a különbséggel, hogy a magabiztosság-krízis indexnél felhasznált fonetikai elemeket pontosítottuk. A magabiztosság-krízis index részletes leírását a következő fejezetben ismertetjük. A monológok vizsgálatához a Praat [1] fonetikai programot használtuk fel, amelyet az Amszterdami Egyetemen fejlesztettek ki.

3.2.2 A magabiztosság-dominancia és a krízishelyzet skálázása, a magabiztosság-krízis index

A krízis, Caplan meghatározása szerint, olyan lelkiállapot, amely külső események hatására alakul ki, amikor az egyének olyan problémákkal találják magukat szemben, amelyek mindennél fontosabbá válnak számukra, és amelyeket sem elkerülni, sem pedig a szokásos eszközökkel megoldani nem tudnak [2]. A krízis meghatározásából következik, hogy egy olyan összetett jelenségcsoportot határoz meg, amely mind intenzitásában, mind pedig sokszínűségében nagy változékonyságot mutat.

A krízishelyzet vizsgálatánál, annak időbeli elhúzódását is figyelembe kell vennünk, illetve a feldolgozás időtartamát. A krízishelyzet és hatásainak feldolgozása, illetve azok kezelhetővé válása, jelentősen csökkentheti a krízishelyzetre utaló jeleket a közlő elbeszélésében, amiből legfeljebb a krízis feldolgozottságának mértékére következtethetünk [7].

Az új vizsgálatok lefolytatásánál változtattunk a fonetikai vizsgálatok paraméterein, és a finomhangolt indexszel újraszámoltuk a korábbi kutatásainkban kapott értékeket is. Az index finomhangolásának lehetőségével már korábbi tanulmányunkban is foglalkoztunk [11]. Az indexet most ezeknek a lehetőségeknek a tükrében módosítottuk. A rövid beszédszakaszok hosszát korábban kettő másodpercben határoztuk meg, amelyet most 1,6 másodpercre csökkentettünk. Korábbi tanulmányunkban még lehetőségként számoltunk ezzel [11].

Vizsgálataink során folyamatosan tanulmányoztuk, hogy a magabiztosság-krízis index összetevői hogyan viselkednek, illetve a paraméterek határértékeinek módosításának szükségességét. Elsősorban a magabiztos és a krízissel terhelt lelkiállapot közötti markáns eltérésekre érdekeltek. Az olasz Lear király elemzését bemutató 2016-os cikkünkben foglalkoztunk először a rövid beszédszakaszok határértékének megváltoztatásával, valamint a hangerő-intervallum paraméter elhagyásával [11]. Tanulmányunkban nemcsak felvetettük a rövid beszédszakaszok hosszának 2-ről 1,6 másodpercre csökkentését, de az 1,6 másodperces határértékekkel is kiszámoltuk az index értékét, és a kapott eredmények alapján felmerült, hogy a későbbiekben ismét megvizsgáljuk a csökkentés lehetőségét.

Az 1,6 másodperces értékre azért esett a választás, mert a rendelkezésre álló eredmények alapján a vizsgált lelkiállapotok közötti különbség ezzel a határértékkel markánsabbá tehető. Amikor először megvizsgáltuk a csökkentés lehetőségét, azt tapasztaltuk, hogy az 1,5, illetve az 1,4 másodperces határérték is, minimálisan ugyan, de már csökkentené a magabiztosságnál és a krízisnél mért index eredménye közötti

különbséget az 1,6 másodperchez képest. Ez a határérték ugyanakkor a jelenlegi kutatásunknál is bevált. Minimális csökkenés az index két értéke között csak a magyar Lear királynál mutatkozott, mivel az első monológ indexének értéke nem változott, míg a második monológé elhanyagolható mértékben csökkent, de ez az eredmény használhatóságát nem befolyásolta. Az esetszámok növekedésével továbbra is meg kívánjuk figyelni a kapott eredményeket, és az adatokat folyamatosan elemezzük.

A másik változtatás az volt, hogy az index összetevői közül a hangerő-intervallumot elhagytuk. Az index kiszámításához öt arányszámot használtunk fel, amelyhez a korábbi vizsgálataink szolgáltak alapul [7, 8, 10, 11], és ezek értékét egymással összeadtuk:

1. Rövid beszédszakaszok: az 1,6 másodpercet meg nem haladó beszédszakaszok száma osztva a vizsgált szöveg szószámával.
2. Magas hangerő: a hangerőcsúcsokat tartalmazó beszédszakaszok száma osztva a vizsgált szöveg szószámával. (Ebbe a kategóriába tartozik minden nyolcvan dB-t meghaladó beszédszakasz, de a megnyilatkozótól függően ennek mértéke a beszélőhöz mérten csökkenthető.)
3 Szelf-referencia: a szelf-referenciára vonatkozó szavak száma osztva a vizsgált szöveg szószámával.
4. Tagadás: a tagadásra vonatkozó szavak száma osztva a vizsgált szöveg szószámával.
5. Mi-referencia (negatív korrekciós index): a mi-referenciára vonatkozó szavak száma osztva a vizsgált szöveg szószámával, negatív előjellel.

3.2.3 Rossz minőségű felvételek elemzési lehetőségei

Az érzelmi állapotok nyelvi, tartalmi jegyekre, illetve fonetikai paraméterekre gyakorolt hatásához korábbi kutatásainkat vettük alapul [7, 8, 10, 11], azzal a különbséggel, hogy kidolgoztunk egy olyan eljárást a fonetikai paraméterek vizsgálatára, amely gyenge minőségű felvételeknél is alkalmazható.

Ha a felvétel olyan minőségű, hogy az elhangzott megnyilatkozás tartalmának döntő hányada érthető, lejegyezhető, akkor ez a nyelvi, tartalmi elemek vizsgálatát nem érinti, hiszen ezek az elemek a rossz minőségű felvétel alapján is elemezhetők. Ebben az esetben tovább léphetünk a megnyilatkozás fonetikai struktúrájának elemzésére. A Ian Holm-féle Lear király rendelkezésre álló hanganyaga alapján az elhangzott megnyilatkozás szövege jól felismerhető, lejegyezhető volt.

A „fonetikai” paramétereknél azonban meg kellett vizsgálnunk, hogy mi az, amit a rendelkezésre álló adatok alapján meg tudunk állapítani, és mi az, amivel nem tudunk foglalkozni. A temporális elemeket, így a beszédszakaszok hosszát is gyengébb minőségű felvételek alapján is meg lehet állapítani. Erre a vizsgált megnyilatkozás alkalmas volt.

A hangerő vizsgálatára vonatkozó műszeres mérési eredmények azonban teljesen használhatatlanok voltak. Hiába lehetett jól kivenni a közlő beszédéből, az elhangzott megnyilatkozást meghallgatva, hogy üvölt, miközben ennek műszeres detektálása gyakorlatilag lehetetlen volt. Így azt a módszert alkalmaztuk, hogy az elhangzott beszédszakasz sajátosságai alapján jelöltük meg, ahol üvöltést, illetve emelt hangerőt hallottunk, vagyis becslést alkalmaztunk a magas hangerő megállapításához. Tisztában vagyunk azzal, hogy ez a módszer nem ad pontos eredményt, de ezt egy közelítő

becsléssel pótolhatjuk. A korábban vizsgált hangerő-intervallumok ezzel a módszerrel nem becsülhetők kellő pontossággal, ennek a paraméternek a vizsgálatát elhagytuk, és a korábbi vizsgálatainkban is újraszámoltuk a magabiztosság-krízis indexet, ennek a paraméternek a használata nélkül. A hangerő-intervallumok vizsgálatának elhagyási lehetősége, figyelembe véve az idővel gyarapodó tapasztalatokat, már a korábbi kutatásunk alkalmával is felmerült [11]. Az index értékeiben nem következett be lényeges változás, ugyanakkor az azonos módszerrel végzett vizsgálatokat is könnyedén összehasonlíthattuk.

A Laurence Olivier-féle második monológnál nem alkalmaztunk becslést, mert a halk aláfestő zene a beszéd hangerejét sehol sem közelítette meg, illetve a magas hangerő értékétől messze elmaradt, így a kapott értékeket felhasználtuk.

A monológok akusztikai változásait a Praat fonetikai programmal vizsgáltuk meg [1].

4 A magabiztosság-krízis index segítségével nyert eredmények

A magabiztosság-krízis indexszel kapott eredmények a két angol nyelvű változatnál is [12, 13] azt mutatják, hogy Lear első monológját a magabiztosság, kiegyensúlyozottság jellemzi, míg utolsó monológját erőteljes krízishelyzet. Az angol nyelvű előadások eredményeit összevetettük a magyar [14], és az olasz nyelvű [15], előadások eredményeivel [8, 10, 11], melyet az 1. táblázat mutat.

Mivel a rossz minőségű felvételek vizsgálatára kidolgozott eljárásunk eredményeinél az index egyik összetevője becslésen alapul, így közelítő értéket kapunk rá, ami viszont az eredmény használhatóságát nem befolyásolja. Azt feltételezzük, hogy a kapott érték valamivel alatta marad annak, mintha a felvétel jó minőségű lenne, mivel a magas hangerőt csak akkor jelöltük meg, ha az elhangzott beszédszakasz sajátosságai alapján jó eséllyel valószínűsíteni tudtuk ennek magas értékét, míg ha ez kérdéses vagy bizonytalan volt a jelölés elmaradt. Így lehetnek olyan beszédszakaszok, amelyeknél a ténylegesen magasabb hangerőt becslés alapján nem tudjuk megállapítani. A jövőben az eljárás pontossága olyan módon is tesztelhető lenne, hogy azonos felvételek rontott, illetve jó minőségű változatából kapott eredményeket hasonlítunk össze egymással.

A táblázatból kitűnik, hogy a magabiztosság-krízis index értékei valamennyi monológnál a skála „megfelelő” pólusa felé térnek ki. A Ian Holm-féle második monológ értéke a legalacsonyabb az utolsó monológok közül, de még ebben az esetben is a várt eredményt kapjuk. Ráadásul ez az a felvétel, amelynél az index értéke becslésen alapul – és feltételezzük, hogy alulbecsült értéken –, amiből az is következik, hogy a becslési eljárásunk is jól használható. A másik szembeötlő eltérés, hogy a nyelvi tartalmi elemek részindexei és a fonetikai jegyek részindexei markánsan eltérnek egymástól, de összességében is nagy változatosságot mutatnak a részindexek, miközben az összesített eredmények minden esetben az elvárásainknak megfelelően alakulnak, vagyis a részindexek együttesen határozzák meg a magabiztosság, illetve a krízis mintázatát.

A fonetikai paraméterek vizsgálatánál a kiválasztott beszédszakaszok szószámát osztottuk el a vizsgált szöveg szószámával. Ha egy beszédszakasz több vizsgált fonetikai paraméternek is megfelelt, akkor valamennyi fonetikai paramétert külön számítottuk be, mintha annyi megjelölt szó lenne az adott beszédszakaszban, ahány az általunk vizsgált fonetikai paraméternek megfelel, függetlenül attól, hogy hány szóból állt a beszédszakasz. Erre azért volt szükség, mert ha több kiugró értéket tartalmaz egy beszédszakasz, akkor intenzívebb a megnyilatkozó lelkiállapota. [7, 8, 10, 11]

1. táblázat: A magabiztosság-krízis index értéke az öt felhasznált mérőszám alapján

Elemzett monológok	Rövid beszéd-szakaszok	Magas hangerő	Szelf-referencia	Tagadás	Mi-referencia	Összesen
Lear 1. – angol, Holm	0,0282	0,0282	0,0141	0,0070	-0,1268	-0,0493
Lear 2. – angol, Holm	0,1408	0,0433	0,0650	0,0433	0,0000	0,2924
Lear 1. – angol, Olivier	0,1250	0,0703	0,0156	0,0078	-0,1250	0,0938
Lear 2. – angol, Olivier	0,2129	0,0684	0,0608	0,0456	0,0000	0,3878
Lear 1. – olasz	0,1647	0,0059	0,0059	0,0059	-0,1176	0,0647
Lear 2. – olasz	0,2574	0,0924	0,0858	0,0726	0,0000	0,5083
Lear 1. – magyar	0,0540	0,0270	0,0000	0,0135	-0,2162	-0,1217
Lear 2. – magyar	0,2632	0,2533	0,1200	0,0133	0,0000	0,6498

A fonetikai paraméterek vizsgálatánál a kiválasztott beszédszakaszok szószámát osztottuk el a vizsgált szöveg szószámával. Ha egy beszédszakasz több vizsgált fonetikai paraméternek is megfelelt, akkor valamennyi fonetikai paramétert külön számítottuk be, mintha annyi megjelölt szó lenne az adott beszédszakaszban, ahány az általunk vizsgált fonetikai paraméternek megfelel, függetlenül attól, hogy hány szóból állt a beszédszakasz. Erre azért volt szükség, mert ha több kiugró értéket tartalmaz egy beszédszakasz, akkor intenzívebb a megnyilatkozó lelkiállapota. [7, 8, 10, 11]

A beszédszakaszok hosszáról azt feltételezzük, hogy következtethetünk belőlük a beszélő gondolatainak összetettségére, illetve arra, hogy az adott helyzetre milyen korábbi konstruált sémával rendelkezik, valamint a váratlan, fájdalomteli helyzetekre, amelyekre kevésbé vannak jól begyakorolt sémáink, és ezeket kevésbé tudjuk kontrollálni. Rövid beszédszakaszok magabiztos megnyilatkozásokban is előfordulnak, de feltételezésünk szerint kisebb arányban. [7, 8, 10, 11]

A hangerőcsúcsokat tartalmazó beszédszakaszok a korábbi vizsgálatok tapasztalatai alapján fontos szerepet töltenek be a krízis meghatározásában [7, 8, 10, 11].

A szelf-referencia és a tagadás előfordulási gyakoriságát nemcsak Pennebaker és Ireland [6] használták fel vizsgáltukban, de László és munkatársai [4] is, akik ezek relatív gyakoriságát tanulmányozták a szövegben. Az énre való túlzott utalás a befelé fordulás jele, míg a 'mi'-re történő utalás a mások irányába való nyitást fejezi ki. Patológiás esetben a magas én-referencia összefüggést mutat a depresszióval, a szuicid tendenciákkal. A tagadást pszichodinamikai szempontból az egészséges emberi környezethez és morális mércékhez való alkalmazkodásra, illetve a világ értéktelenítésére, a destrukcióra és öndestrukcióra való hajlamra vonatkozóan vizsgálták [3]. Krízishelyzetben problémás az alkalmazkodás a megváltozott környezethez, a tagadás fokozottan fordulhat elő az elbeszélésben.

A mi-referencia a kiegyensúlyozott megnyilatkozásoknál gyakrabban fordul elő, így a többi vizsgálta paraméterrel szemben, ellentétes hatást fejt ki, ezért ezt negatív korrekciós mérőszámként használtuk fel az indexnél [7, 8, 10, 11].

5 Összegzés

Vizsgálatunkban, korábbi kutatásainkhoz kapcsolódva egy olyan új narratív pszichológiai szemlélet meghonosítására tettünk kísérletet, amely a tudományos narratív pszichológiai tartalomelemzést és a fonetikai jegyek vizsgálatát kapcsolja össze, egy *összetett tudományos narratív pszichológiai eljárás* keretében. Az eddig lefolytatott vizsgálatok eredményei, akár színészi játékról, akár spontán megnyilatkozásokról legyen szó, azt mutatják, hogy a krízis mérésére kidolgozott eljárásunk, melyet a magabiztosság-krízis indexszel számszerűsítünk, eredményesen alkalmazható a magabiztos, illetve a krízissel terhelt lelkiállapot mérésére. [7, 8, 10, 11]

A vizsgált négy különböző Lear király sok dologban eltér egymástól, és mindezen eltérések ellenére a magabiztosság-krízis indexnél mégis a várt értékeket kapjuk. A szerep megformálása, a színészi játék, de még a színészek karaktere is teljesen eltér egymástól. Az egyetlen közös dolog a játékukban, hogy valamennyien Lear királyt alakítják. Pont ezek az eltérések erősítik meg feltételezéseink helyességét. Mivel színházi előadásokról van szó, felvetődik a kérdés, hogy nem azért kapjuk-e meg a kívánt értékeket, mert a különböző előadások hatással voltak egymásra. A legkorábbi előadás az olasz, amely 1960-ban készült, ami egészen biztosan nem volt hatással az 1978-ban készült magyar változatra. Ha figyelembe vesszük, hogy akkor még magyar művészek csak korlátozottan jutottak el a vasfüggönyön túlra, valamint a nyelvi

akadályokat, a két előadás között eltelt időt, és azt, hogy abban az időben ilyen felvételeket nem volt egyszerű beszerezni, videómegosztó csatornák pedig nemcsak hogy nem működtek, de még internet sem volt, akkor gyakorlatilag ez jó eséllyel kizárható. Az 1983-as és az 1998-as angol előadás sem hathatott a magyar előadásra, mivel azok később keletkeztek, és az is nehezen elképzelhető, hogy a magyar, vagy akár az olasz előadás hatott volna rájuk. Az egymásra hatás legfeljebb a két angol nyelvű előadásnál merülhetne fel, de Lear megformálása annyira markánsan eltér egymástól a két felvételnél, hogy ilyen hatásról ebben a két esetben sem beszélhetünk.

A lefolytatott vizsgálat megerősítette, hogy a magabiztos és a krízissel terhelt megnyilatkozásoknak meghatározott mintázata van, amely nyelvi tartalmi elemekkel és fonetikai jegyekkel együttesen írható le, és ez a mintázat nyelvfüggetlen. Vizsgálatunk során az eljárás lefolytatásához használt magabiztosság-krízis indexet finomhangoltuk, valamint kidolgoztunk egy olyan becslési eljárást, amellyel a rosszabb minőségű hanganyagokat is tanulmányozni tudjuk, ezzel nagymértékben kibővítve a vizsgálható hanganyagok körét. A kutatások további iránya lehet a vizsgálat kiterjesztése más nyelvekre, illetve az esetszámok növekedésével az eljárás esetleges finomítása a tapasztalatok tükrében, a kidolgozott becslési eljárás pontosságának vizsgálata ugyanazon megnyilatkozások jó minőségű és „rontott" hanganyagainak az összehasonlítása révén, valamint más jelenségek vizsgálata az *összetett tudományos narratív pszichológiai eljárás* lefolytatásával.

Hivatkozások

1. Boersma, P., Weenink, D.: Praat: Doing phonetics by computer [computer program]. Forrás:http://www.praat.org/ (2013)
2. Caplan, G. (1964). Principles of preventive psychiatry. New York, Basic Books.
3. Hargitai, R. Naszódi, M., Kis, B., Nagy, L., Bóna, A., László, J.: A depresszív dinamika nyelvi markerei az én-elbeszélésekben. A LAS VERTIKUM tagadás és szelfreferencia modulja. Pszichológia, 2 (2005) 181-199.
4. László, J.: The science of stories.: An introduction to narrative psychology. London; New York: Routledge. (2008)
5. László, J.: Történelemtörténetek – Bevezetés a narratív pszichológiába. Budapest, Akadémiai Kiadó. (2012)
6. Pennebaker, J. W., Ireland, M.: Analyzing Words to Understanding. In: Jan Auracher, William van Peer (Eds.): New Beginnings to Literary Studies. Cambridge Scholar Publishing. (2008) 24-48.
7. Puskás, L.: A magabiztosság-krízis skála gyakorlati alkalmazása. Forrás: Takács A., Varga V., Vincze V. (szerkesztették): X. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Tanszékcsoport. (JATE Press), Szeged (2014) 155-163.
8. Puskás L.: Paralingvisztikai jegyek a narratív pszichológiai tartalomelemzésben: a magabiztosság-krízis skála. Forrás: Takács A., Vincze V. (szerk.): VIII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Tanszékcsoport. (JATE Press) Szeged (2011) 231-239.
9. Puskás, L., Karsai, B.: A New Method in Narrative Psychology. In: Cognition and Interpretation. Pécs Studies in Psychology. Edited by Beatrix Lábadi. PTE BTK Pszichológiai Intézet. (2008)

10. Puskás, L., László, J., Fülöp, É.: Lear király lelkiállapot-változása első és utolsó monológjának szövegbeli és akusztikai jegyei alapján. Pszichológia, 2 (2012)
11. Puskás, L., Pólya, T.: A magabiztosság-krízis skála alkalmazása idegen nyelvű megnyilatkozásoknál. In: Takács A., Varga V., Vincze V. (szerk.): XII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Tanszékcsoport. Szeged (2016) 305-312.
12. Shakespeare, W.: King Lear. Rendezte: Michael Elliot. Granada Television. (Tévéjáték, 158 perc.) (1983)
13. Shakespeare, W.: King Lear. Rendezte: Richard Eyre. BBC. (Tévéjáték, 150 perc.) (1998)
14. Shakespeare, W.: Lear király. Vörösmarty Mihály fordítását Mészöly Dezső dolgozta át. Rendező: Vámos László. Magyar Televízió, 1978. (Tévéjáték, 156 perc.) (1978)
15. Shakespeare, W.: Re Lear. Gino Chiarini fordítását Sandro Bolchi dolgozta át és rendezte. (Tévéjáték, 185 perc.) (1960)

Igei többszavas kifejezések felismerése nyelvfüggetlen módszerekkel

Simkó Katalin Ilona[1,2], Kovács Viktória[2], Vincze Veronika[2,3]

[1]Szegedi Tudományegyetem, Informatikai Intézet,
Szeged, Árpád tér 2.
simko@hung.u-szeged.hu
[2]Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék,
Szeged, Egyetem u. 2.
viki921015@hotmail.com
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Kivonat Cikkünk a PARSEME Shared Task-ra beadott igei többszavas kifejezéseket felismerő rendszerünk kiegészített változatát mutatja be. Redszerünk szófaji egyértelműsítésen és szintaktikai elemzésen keresztül ismer fel egy- és többtokenes igei többszavas kifejezéseket a szövegekben. A rendszer nyelvfüggetlenül működik, a Shared Taskon 18-ból 9 nyelven versenyzett. Az új változat mind a 18 nyelvre nyújt elemzést. Cikkünkben bemutatjuk a rendszert és hibaelemzést is adunk.
Kulcsszavak: többszavas kifejezés, MWE, nyelvfüggetlen

1. Bevezetés

A többszavas kifejezések (multiword expression, MWE) gyakori elemei a természetes nyelveknek. Olyan kifejezések, amelyek több lexémából állnak, de a jelentésük nem számítható ki az összetevőik jelentéséből. Ide tartoznak például a gyakori hasonlatok (*ritka, mint a fehér holló*), a kollokációk (*kerek szám*), vagy az idiómák (*feldobta a talpát*). Cikkünkben igei többszavas kifejezésekkel foglalkozunk, olyan többszavas kifejezésekkel, amelyeknek egyik tagja ige, illetve a kifejezés összessége igei értelmet hordoz. Ilyenek például az igekötő-ige konstrukciók (*beszáll*) vagy a könnyűigés szerkezetek (*döntést hoz*).

Cikkünkben az igei többszavas kifejezések automatikus felismerésére létrehozott rendszerünket mutatjuk be, amelynek első verzióját a PARSEME Workshopra [1] készítettük. Jelen cikkünkben ezt a rendszert és ennek új, kibővített változatát is bemutatjuk. Rendszerünk szófaji egyértelműsítésen és függőségi elemzésen alapul és képes egy- és többtokenes igei többszavas kifejezések felismerésére. A rendszer nyelvfüggetlen: az első verziót a Shared Task 9 nyelvére adtuk be, az új, kibővített rendszerrel már a Shared Tasknak mind a 18 nyelvére adunk eredményeket.

A cikkben először az eredeti rendszer működését írjuk le a Shared Taskon elért eredményeinkkel és azok hibaelemzésével. Ezután a bővítésről szólunk, valamint az új rendszerrel mind a 18 nyelvre adunk eredményeket. Végül összehasonlítjuk az eredeti és az új megközelítésünk által adott eredményeket magyar adatokon.

2. Az eredeti rendszer

Az eredeti rendszerünket a 2017-es MWE workshop [1] igei többszavas kifejezések felismerésére kiírt Shared Taskjára hoztuk létre.

A shared task célja igei többszavas kifejezések felismerése sok nyelven. Összesen 18 nyelvre készült annotáció, az annotációs utasítások univerzális és nyelvspecifikus jegyek figyelembevételével készültek.

Az utasítások öt különböző igei MWE kategóriát különítenek el: idiómákat (ID), könnyűigés szerkezeteket (LVC), ige-partikula konstrukciókat (VPC), inherensen reflexív igéket (IRefIV), valamint egy „egyéb" kategóriát (OTH). A számítógépes nyelvészetben ezek felismerése nehéz feladat, mivel gyakran nem projektív szerkezetekben állnak, jelentésük pedig nem kompozicionális, a kategóriák heterogének és a struktúrák szintaktikai variabilitása nagy.

Csapatunk a magyar adatbázis elkészítésében, annotálásában is részt vett, így rendszerünk nagyrészt az ott szerzett tapasztalatokra épül.

2.1. A rendszer leírása

Rendszerünk az MWE-k és a szintaktikai elemzés kapcsolatára épül, amit több forrás is igazol [2,3,4,5,6,7,8], valamint a PARSEME csoport [1] munkájának is az egyik vezérelve.

A rendszerünk alapját, a függőségi elemzés MWE-felismerésre való használatát, korábban már [9] is bemutatta, valamint [10] cikkünkben is foglalkoztunk a függőségi címkekészlet manipulálásának hatásával az elemzés további felhasználására. Mivel a shared task több nyelve is a morfológiailag gazdag, szabad szórendű nyelvek közé tartozik, így az MWE tagjai is nagyobb szintaktikai flexibilitást élveznek: nem feltétlenül egymás mellett fordulnak elő, távolabb kerülhetnek egymástól a mondatban, így ez a megközelítés jobb stratégiának tűnik, mint a szekvenciajelölés és más, hasonló megközelítések.

A fent említett cikkben [9] bemutatott rendszer az MWE típusának megfelelő függőségi élekkel dolgozik, így például az ige-tárgy könnyűigés szerkezetekben a tárgy OBJ-LVC szintaktikai-szemantikai annotációt kap.

Ezzel ellentétben jelenlegi rendszerünk csak az MWE típusát tünteti fel a függőségi élen és egytokenes MWE-kre is alkalmazható. Mivel több nyelvben is előfordultak egytokenes MWE-k – a többtokenesek mellett, amelyeket a függőségi elemzés szintjén kezelünk –, így a szófaji egyértelműsítésre is kiterjesztettük a megközelítést. Gyakori egytokenes MWE-k például a német és magyar VPC-k: amikor a partikula közvetlenül megelőzi az igét, a német és a magyar helyesírási

[1] http://typo.uni-konstanz.de/parseme/

szabályok alapján egy szóba írjuk őket, ám ennek ellenére a szerkezet két lexémából áll és jelentése nem kompozicionális (például német *auffallen* (rá+esik, „kitűnik") vagy magyar *felad* „abbahagy" értelemben). Ezzel ellentétben az idiómák szinte mindig többtokenes MWE-k (például *előre iszik a medve bőrére*).

Az MWE-k specifikus morfológiai, szintaktikai és szemantikai tulajdonságokkal rendelkeznek. Megközelítésünkben a többtokenes MWE-ket a szintaxis szintjén kezeljük - az Universal Dependencies nyelvtan [11] `mwe` relációjához hasonlóan -, az egytokeneseket pedig a morfológia szintjén.

Rendszerünk négy lépésben dolgozik: a fő MWE-felismerő lépések a szöveg szófaji egyértelműsítésének és függőségi elemzésének szintjén zajlanak. Rendszerünk a shared task szervezői által biztosított szófaji egyértelműsítésen és függőségi szintaktikai elemzésen, valamint az MWE annotált szövegeken alapszik. A rendszer ezen inputok megléte mellett nyelvfüggetlen elemzésre képes.

Első lépésben a fent említett inputokból létrehozzuk a tanító adathalmazt. A tanító adatok MWE-annotációját a morfológiai és a függőségi annotációba illesztjük az egy- és a többtokenes MWE-re külön-külön. Az egytokenes MWE-k szófaji címkéjét az MWE típusukkal helyettesítjük, míg többtokenes MWE-k esetén a függőségi fa élein cseréljük le a címkéket: a fában lejjebb elhelyezkedő MWE tagok függőségi címkéjét cseréljük az MWE típusokra.

Az 1-3. ábrák az egytokenes MWE-k szófaji címke cseréjét és a többtokenes MWE-k függőségi címke cseréjét mutatják be VPC és LVC magyar példákra.

bekezdés	NOUN	SubPOS=c|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none
bekezdés	VPC	SubPOS=c|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none
határozathozatal	NOUN	SubPOS=c|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none
határozathozatal	LVC	SubPOS=c|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none

1. ábra: VPC és LVC példák átcímkézése MWE szófaji címkékkel a *bekezdés* és *határozathozatal* példákban.

A többtokenes MWE-k esetén megközelítésünk arra a hipotézisre alapoz, hogy az MWE alsóbb tagjai közvetlenül más MWE tagokhoz kapcsolódnak a függőségi fában. A fa struktúráját nem változtatjuk meg, csak az MWE-k alsóbb tagjait kapcsoló, meglévő élek címkéit. Például a 2 példában található *ellát* VPC tagjai: *lát* és *el* esetén az alsóbb tag, *el* **PREVERB** címkével a felsőbb MWE taghoz, *lát*hoz kapcsolódik; az átalakítással ezt a **PREVERB** címkét írjuk át az MWE-t jelző **VPC** címkéjére.

Több mint kéttokenes MWE-k esetén az átalakítás az összes alsóbb MWE tag címkéjét átírja az MWE-specifikus címkére. Például 4 mondat esetén az *az első követ veti* idióma legfelső tagja a függőségi fában a *vetette* ige. Az összes több MWE tag címkéje **ID** MWE címkére változik.

Rendszerünk második lépése a szintaktikai elemző tanítása: munkánkban a Bohnet elemzőt használtuk [12] szófaji egyértelműsítésre és függőségi elemzésre is. Az egytokenes MWE-khez a Bohnet elemző szófaji egyértelműsítő modulját tanítottuk az MWE átalakított korpuszon, a többtokeneshez pedig a függőségi

2. ábra: VPC többtokenes MWE címke hozzáadása a függőségi fához a *Péter fontos feladatokat lát el* mondatban.

3. ábra: LVC többtokenes MWE címke hozzáadása a függőségi fához a *Péter fontos döntést hoz* mondatban.

4. ábra: ID többtokenes MWE címke hozzáadása a függőségi fához a *Péter vetette rá az első követ* mondatban.

elemzőt. Az elemző az MWE szófaji és függőségi címkéket ugyanúgy kezeli, mint az általános szófaji és függőségi címkéket.

Minden nyelv esetén ugyanezeket a lépéseket végeztük el és szófaji egyértelműsítő és függőségi elemző modelleket hoztunk létre, amelyek MWE felismerésére is képesek. Néhány nyelv esetén egyes mondatokat el kellett távolítanunk a tanító adatból, amelyek (néhány esetben 500 token feletti) hosszuk miatt hibákat okoztak a tanítás során.

Harmadik lépésben leelemeztettük a szófaji egyértelműsítő és függőségi elemző modelleket az egyes nyelvek teszt korpuszain. Az így kapott elemzés MWE szófaji címkéket és függőségi címkéket is tartalmazott az adott nyelv általános címkéi mellett.

Negyedik és egyben utolsó lépésben kinyertük az MWE szófaji és függőségi címkéket a kimenetből. Az MWE szófajú

szavak egytokenes, a szófaji címkében jelzett típusú MWE annotációt kaptak. Az MWE függőségi címkék esetén az MWE címkékkel összekötött szavak együtt alkotnak a címke által jelzett típusú, többtokenes MWE-t.

5. ábra: Rendszerünk által végrehajtott lépések.

2.2. Eredmények

A shared taskon a rendszerünket csak azokra a nyelvekre nyújtottuk be, amelyekre függőségi elemzést és szófaji egyértelműsítést is biztosítottak. A szófaji egyértelműsítésre az egyes nyelvekben gyakori, egytokenes MWE-khez volt szükségünk, míg a függőségi elemzést a többtokenes MWE-khez használtuk fel. Azon nyelvek esetén, ahol csak szófaji címkézés volt elérhető, megpróbálkoztunk részleges eredményeket (csak az egytokenes MWE-k elemzését) létrehozni, de ezeken a nyelveken nem vagy csak nagyon kevés egytokenes MWE volt jelen, így ez nem volt elegendő a tanításhoz.

Eredményeink a kilenc nyelvre a 1. táblázatban találhatóak. Rendszerünket német (DE), görög (EL), spanyol (ES), francia (FR), magyar (HU), olasz (IT), lengyel (PL), portugál (PT) és svéd (SV) nyelvekre nyújtottuk be.

Az F-értékben mért eredmények nagy váltakozást mutatnak nyelvenként és rendszerenként is. Német, magyar és svéd nyelvekre a legjobb vagy közel a legjobb rendszert hoztuk létre. A többi nyelv esetén (a lengyel és a portugál kivé-

1. táblázat. Legjobb, legrosszabb és az USzeged rendszerek eredményei a nyelvekre tokenszintű F-érték alapján rangsorolva.

	Rendszer	P-MWE	R-MWE	F-MWE	P-token	R-token	F-token
DE	BEST, USZEGED	0,5154	0,3340	0,4053	0,6592	0,3468	0,4545
	LAST	0,3652	0,1300	0,1917	0,6716	0,1793	0,2830
EL	BEST	0,3612	0,4500	0,4007	0,4635	0,4742	0,4688
	USZEGED	0,3084	0,3300	0,3188	0,4451	0,3757	0,4075
	LAST	0,4286	0,2520	0,3174	0,5616	0,2953	0,3871
ES	BEST	0,6122	0,5400	0,5739	0,6574	0,5252	0,5839
	USZEGED	0,2575	0,5000	0,3399	0,3635	0,5629	0,4418
	LAST	0,6447	0,1960	0,3006	0,7233	0,1967	0,3093
FR	BEST	0,6147	0,4340	0,5088	0,8088	0,4964	0,6152
	USZEGED	0,0639	0,0520	0,0573	0,5218	0,2482	0,3364
	LAST	0,8056	0,0580	0,1082	0,8194	0,0532	0,1000
HU	BEST, USZEGED	0,7936	0,6934	0,7401	0,8057	0,6317	0,7081
	LAST	0,8029	0,5471	0,6508	0,8208	0,5015	0,6226
IT	BEST	0,5354	0,3180	0,3990	0,6134	0,3378	0,4357
	USZEGED	0,1503	0,1560	0,1531	0,4054	0,3064	0,3490
	LAST	0,6125	0,0980	0,1690	0,6837	0,1053	0,1824
PL	BEST	0,7798	0,6020	0,6795	0,8742	0,6228	0,7274
	LAST, USZEGED	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
PT	BEST	0,7543	0,6080	0,6733	0,8005	0,6370	0,7094
	LAST, USZEGED	0,0129	0,0080	0,0099	0,6837	0,1987	0,3079
SV	BEST	0,4860	0,2203	0,3032	0,5253	0,2249	0,3149
	USZEGED	0,2482	0,2966	0,2703	0,2961	0,3294	0,3119
	LAST	0,5758	0,1610	0,2517	0,6538	0,1677	0,2669

telével) a középmezőnyben végeztünk. A nyelvek közötti eltérést nagyban befolyásolja az adott nyelvben gyakori MWE-típusok és rendszerünk működésének összefüggései.

2.3. Hibaelemzés

A tesztkorpusz kézi annotációjának nyilvánosságra hozása után megvizsgáltuk a rendszerünk által vétett hibákat.

Hibaelemzésünk megmutatta, hogy rendszerünk magasan a legjobban a VPC típusú igei MWE-ken teljesít: a VPC-k közel 60%-át helyesen találja meg, míg a többi kategória esetén ez az arány körülbelül 40%. Az ige-partikula konstrukciók esetén jellemzően szintaktikai kapcsolat is fennáll az MWE tagok között, ami segíti a rendszerünk működését.

A német, magyar és svéd nyelvek tartalmaztak legnagyobb arányban VPC típusú MWE-ket a shared taskban, ami szintén összefüggésben áll azzal, hogy az elemzőnk miért ezeken a nyelveken működött legjobban. Ezzel szemben az újlatin és egyéb nyelvek alig tartalmaznak VPC-ket. Ezek alapján úgy tűnik, a rendszerünk által egy adott nyelven elért eredményt leginkább az befolyásolja, hogy az adott nyelv milyen MWE típusokat használ gyakran.

Francia és olasz nyelveken az IRefIV típus felismerése okozott nehézségeket. Általánosságban pedig elmondható, hogy a hosszabb idiómák és LVC-k, valamint prepozíciókat tartalmazó többszavas kifejezések problémásak voltak. A rendszer alapfelépítéséből fakadó hiba, hogy azokban az esetekben, amikor az MWE tagjait nem köti össze szintaktikai reláció, nem sikerült helyesen felismernünk az MWE-t; például németben a *sein* kopula gyakran nem kapcsolódik közvetlenül az MWE többi tagjához a szintaktikai elemzésben.

3. A kibővített rendszer

Rendszerünk bővítésének elsődleges célja a shared taskban szófaji egyértelműsítés és függőségi adat nélkül kiadott tanító adatbázisú nyelvekre is elemzést adni. Ezt úgy értük el, hogy a rendszert egy előfeldolgozó lépéssel bővítettük, amelyben ezt az annotációt hozzáadtuk a tanító adatbázishoz.

6. ábra: Kibővített rendszerünk által végrehajtott lépések.

A kibővített rendszer előfeldolgozó lépéséhez a Universal Dependencies Project [13] tanító adatbázisait használtuk fel az egyes nyelvekre. Az adatbázisokon szintén a Bohnet elemző szófaji egyértelműsítő és függőségi elemző moduljait tanítottuk. Ezután a már bemutatott lépéseket követtük: a leelemzett MWE tanítóadatbázisok szófaji és függőségi címkéit az MWE-specifikusakra cseréltük, majd ezeken a tanító adatbázisokon újra a Bohnet elemzőt tanítottuk, majd a teszten lefuttatva a modelleket, kinyertük az MWE-specifikus szófaji és függőségi címkéket.

3.1. Eredmények

Kibővített rendszerünk eredményei a 2. táblázatban láthatóak a shared task minden nyelvére. A már az eredeti rendszer által is elemzett nyelvek a táblázat felső felében találhatóak, az elemzett tanító adatbázisú nyelvek pedig a táblázat alsó felében. A „felső" nyelvek eredményei valamelyest eltérnek az 1. táblázatban láthatóaktól, ezek a változások egyéb, kisebb javítások eredményei, illetve a statisztikai elemző működésének köszönhetők.

2. táblázat. A kibővített rendszer eredményei: a már az eredeti rendszer által is elemzett nyelvek vannak először feltüntetve.

	P-MWE	R-MWE	F-MWE	P-token	R-token	F-token
DE	0,5927	0,326	0,4206	0,689	0,3034	0,4213
EL	0,3701	0,302	0,3326	0,4914	0,3265	0,3923
ES	0,3006	0,098	0,1478	0,4587	0,1283	0,2005
FR	0,2786	0,258	0,2679	0,4007	0,3312	0,3626
HU	0,6218	0,5371	0,5763	0,5876	0,5311	0,5579
IT	0,0748	0,038	0,0504	0,3028	0,117	0,1688
PL	0,3264	0,5	0,3949	0,3815	0,5425	0,448
PT	0,264	0,32	0,2893	0,3308	0,3668	0,3479
SV	0,2602	0,2161	0,2361	0,3394	0,2564	0,2921
BG	0,5326	0,4313	0,4766	0,7779	0,4925	0,6032
CS	0,4495	0,5793	0,5062	0,576	0,6446	0,6084
FA	0,6958	0,462	0,5553	0,8578	0,5314	0,6563
HE	0,4118	0,084	0,1395	0,5522	0,0863	0,1493
LT	0,3333	0,07	0,1157	0,4048	0,0697	0,1189
MT	0	0	0	0,0465	0,0031	0,0059
RO	0,4629	0,674	0,5489	0,5301	0,7168	0,6095
SL	0,5949	0,188	0,2857	0,6646	0,1877	0,2927
TR	0,3934	0,3792	0,3862	0,4207	0,3949	0,4074

Az, hogy az első lépésben etalon vagy elemzett szófaji és függőségi adatokat használunk, nem befolyásolja szignifikánsan az eredményeket (ahogy azt a 4 bekezdésben is tárgyaljuk), a máltai (MT) kivételével. Máltai nyelvre jelenleg nincs elérhető Universal Dependencies treebank, így keresztanítást végeztünk az elemző tanításához. A máltai nyelv legközelebbi rokona az az arab és az olasz,

viszont latin karaktereket használ, így az olasz Universal Dependencies tanító adatbázis használata mellett döntöttünk szófaji egyértelműsítőnk és függőségi elemzőnk tanításához. Sajnos ez nagyon rossz eredményekhez vezetett: egyetlen teljes MWE-t sem sikerült helyesen felismernünk a tesztkorpuszban. A többi elemzett tanítóadatot használó nyelvnél – amelyek esetén az elemzés tanító adata is az adott nyelvről származott – a végső eredmények nagyon közeliek az etalon adatot használókhoz.

A 3. és 4. táblázatok a különböző MWE típusokra mutatják az F-értékben mért eredményeinket; az áthúzással jelölt cellák azt mutatják, hogy a típus nem szerepelt az adott nyelvben.

Összességében a rendszer legjobban az inherensen reflexív igéken működik, legrosszabbul pedig az „egyéb” kategórián, mivel itt a legnagyobb a variabilitás és az MWE-k hossza. A könnyűigés szerkezetek és ige-partikula konstrukciók a különböző nyelvek között nagy különbségeket mutatnak.

3. táblázat. MWE-szintű F-értékek a különböző MWE típusokra.

	mind	VPC	LVC	ID	IReflV	OTH
BG	0,4766	-	0,2044	0,1879	0,607	-
CS	0,5062	0	0,2696	0,0251	0,6171	0
DE	0,4206	0,6316	0	0,2039	0,087	-
EL	0,3326	0,56	0,3668	0,0952	-	0,08
ES	0,1478	0	0,2346	0,0567	0,1712	0
FA	0,5553	-	-	-	-	0,5553
FR	0,2679	-	0,1909	0,1718	0,5314	0
HE	0,1395	0,0935	0,242	0	-	0,1376
HU	0,5763	0,6873	0,2947	-	-	-
IT	0,0504	0,2857	0,0174	0,0316	0	0
LT	0,1157	-	0,2353	0	-	-
MT	0	-	0	0	-	0
PL	0,3949	-	0,1162	0	0,626	-
PT	0,2893	-	0,3086	0,1026	0,3727	0
RO	0,5489	-	0,7059	0,2051	0,5492	-
SL	0,2857	0	0,1429	0,019	0,4571	0
SV	0,2361	0,3134	0,1538	0	0,0606	0
TR	0,3862	-	0,3977	0,3271	-	0,3434
Átlag	0,3163	0,2422	0,2273	0,0893	0,1266	0,4079

4. Etalon vagy elemzett?

Ebben az utolsó szekcióban egy rövid kísérletet írunk le, amellyel az új rendszerünkben bevezetett, elemzett és az etalon szófaji és függőségi tanítóadat által okozott különbségeket vizsgáltuk meg. A magyar adatokon hasonlítottuk össze az etalon szófaji és függőségi tanítóadatot használó megközelítésünk MWE-azonosításban elért eredményeit az elemzettel. Az elemzett adatot ugyanazzal a

4. táblázat. Token-szintű F-értékek a különböző MWE típusokra.

	mind	VPC	LVC	ID	IReflV	OTH
BG	0,6032	-	0,3263	0,243	0,7405	-
CS	0,6084	0	0,3178	0,1787	0,7244	0
DE	0,4213	0,6344	0,0392	0,3003	0,1455	-
EL	0,3923	0,56	0,4065	0,1997	-	0,0404
ES	0,2005	0	0,2552	0,1538	0,1709	0
FA	0,6563	-	-	-	-	0,6563
FR	0,3626	-	0,2184	0,3292	0,5362	0
HE	0,1493	0,1382	0,2321	0,1062	-	0,1083
HU	0,5579	0,6821	0,3279	-	-	-
IT	0,1688	0,2857	0,096	0,1659	0	0
LT	0,1189	-	0,233	0,0328	-	-
MT	0,0059	-	0,0071	0,0033	-	0
PL	0,448	-	0,1415	0,1	0,6727	-
PT	0,3479	-	0,3348	0,2454	0,3766	0
RO	0,6095	-	0,7303	0,4888	0,5575	-
SL	0,2927	0,017	0,1587	0,0936	0,4629	0
SV	0,2921	0,3284	0,15	0,0619	0,0606	0
TR	0,4074	-	0,4132	0,3443	-	0,3636
Átlag	0,3676	0,253	0,2572	0,1797	0,1351	0,4548

módszerrel hoztuk létre, mint azon nyelvek esetén, ahol nem volt elérhető etalon adat: a magyar Universal Dependencies treebanken tanítottuk a Bohnet elemző szófaji egyértelműsítő és függőségi elemző moduljait, majd ezeket a modelleket futtattuk az MWE tanító adatbázison.

A 5. táblázat a kísérlet eredményeit mutatja; a HU-ETALON oszlop eredményei az előző táblázatban lévő magyar számokkal megegyeznek. Az eredmények alapján az etalon és elemzett tanítóadatbázis használatával nagyon hasonló eredmények érhetőek el az igei többszavas kifejezések felismerésében. Érdekes módon, MWE- és tokenszintű kiértékelés mellett is magasabb a pontosság, de valamivel alacsonyabb fedés érhető el az elemzett adat felhasználásával, mint az etalonnal.

MWE-típusok szempontjából az LVC-k okozzák a legnagyobb különbséget a két megközelítés között. A VPC típusban nagyon sok az egytokenes MWE a magyarban, így ezt a kategóriát kevésbé befolyásolják az etalon és elemzett függőségi fák között előforduló különbségek.

Összességében elmondható, hogy a kibővített rendszerünk által alkalmazott mindkét megközelítéssel nagyjából azonos eredményeket érhetünk el igei többszavas kifejezések felismerésében.

5. Összegzés

Cikkünkben igei többszavas kifejezések felismerésére készített rendszerünket mutattuk be. A rendszer szófaji egyértelműsítést és függőségi elemzést használ az igei MWE-k felismerésére több nyelven. Rendszerünk nyelvfüggetlen: igei MWE-

5. táblázat. Etalon és elemzett eredmények magyarra.

	HU-ETALON	HU-ELEMZETT
P-MWE	0,6218	0,7335
R-MWE	0,5371	0,491
F-MWE	0,5763	0,5882
P-token	0,5876	0,7383
R-token	0,5311	0,4423
F-token	0,5579	0,5532
F-VPC-MWE	0,6873	0,679
F-LVC-MWE	0,2947	0,2623
F-VPC-token	0,6821	0,6831
F-LVC-token	0,3279	0,2768

kre, szófaji címkékre és függőségi relációkra annotált tanító adatbázis, vagy igei MWE-kre annotált tanító adatbázis és adott nyelvi szófaji egyértelműsítéssel és függőségi treebank felhasználásával bármilyen nyelvre alkalmazható.

Cikkünk megmutatja, hogy lehetséges nyelvfüggetlen MWE-felismerő rendszert építeni limitált nyelvspecifikus adat esetén is elfogadható eredményekkel.

A jövőben érdekes lehet megvizsgálni, hogy más végalkalmazások is integrálhatóak-e a függőségi elemzésbe és ha igen, hogyan.

Köszönetnyilvánítás

A kutatást az EFOP-3.6.1- 16-2016-00008 azonosítójú, EU társfinanszírozású projekt támogatta. Vincze Veronika kutatásait az Emberi Erőforrások Minisztériuma ÚNKP-17-4 kódszámú Új Nemzeti Kiválóság Programja támogatta.

Hivatkozások

1. Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., Doucet, A.: The parseme shared task on automatic identification of verbal multiword expressions, Association for Computational Linguistics (2017) 31–47
2. Constant, M., Nivre, J.: A transition-based system for joint lexical and syntactic analysis. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Association for Computational Linguistics (2016) 161–171
3. Nasr, A., Ramisch, C., Deulofeu, J., Valli, A.: Joint dependency parsing and multiword expression tokenization. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, Association for Computational Linguistics (2015) 1116–1126
4. Candito, M., Constant, M.: Strategies for contiguous multiword expression analysis and dependency parsing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, Association for Computational Linguistics (2014) 743–753

5. Green, S., de Marneffe, M.C., Bauer, J., Manning, C.D.: Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK., Association for Computational Linguistics (2011) 725–735
6. Waszczuk, J., Savary, A., Parmentier, Y.: Promoting multiword expressions in A* TAG parsing. In: COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan. (2016) 429–439
7. Wehrli, E., Seretan, V., Nerima, L.: Sentence analysis and collocation identification. In: Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), Beijing, China, Association for Computational Linguistics (2010) 27–35
8. Green, S., de Marneffe, M.C., Manning, C.D.: Parsing Models for Identifying Multiword Expressions. Computational Linguistics **39**(1) (2013) 195–227
9. Vincze, V., Zsibrita, J., Nagy T., I.: Dependency Parsing for Identifying Hungarian Light Verb Constructions. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, Asian Federation of Natural Language Processing (2013) 207–215
10. Simkó, K.I., Kovács, V., Vincze, V.: Szintaktikai címkekészletek hatása az elemzés eredményességére. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2017) 316–322
11. Nivre, J.: Towards a Universal Grammar for Natural Language Processing. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Springer (2015) 3–16
12. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). (2010) 89–97
13. Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D.: Universal dependencies v1: A multilingual treebank collection. In Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)

Jogosultság- és kötelezettségfogalmak a KRESZ-ben

Syi[1], Markovich Réka[234], Hamp Gábor[1]

[1]BME Szociológia és Kommunikáció Tanszék
i@syi.hu, hampg@eik.bme.hu
[2] LSE Centre for Philosophy of Natural and Social Science
[3]ELTE Logika Tanszék
[4]BME Üzleti Jogi Tanszék
markovich@phil.elte.hu

Kivonat Kutatásunk célja a közlekedés rendjét szabályzó jogszabály formális leírása, az ehhez szükséges fogalmakat leírni képes szakontológia felállítása. Ebben a tanulmányban azt mutatjuk meg, hogy a szabályok klasszikus szemantikai elemzéséhez használt deontikus modalitások mellett milyen szerepe lehet egy másik megközelítésnek: a jogosultságok és kötelezettségek korrelatív párokba rendezett rendszere jogi ontológiába való beemelésének. A fogalmakat és relációikat Wesley Newcomb Hohfeld erre vonatkozó rendszere szerint kezeljük, de bemutatjuk a klasszikus deontikus modalitáshoz való viszonyukat; a KRESZ-ből vett példákkal megmutatjuk, hogy hogyan különíthetők el egymástól az egyes jogosultság- és kötelezettségfogalmak és milyen jogelméleti megfontolások milyen jogi terminológiai megoldásokkal érhetők tetten, és hogy mikor melyik rendszertanra érdemes támaszkodnunk.[1]
Kulcsszavak: KRESZ, ontológia, deontikus modalitások, jogosultságok

1. Bevezetés

A KRESZ-szabályokat akkor rögzítették,[2] amikor még szó sem volt az önvezető autók megjelenéséről, a közlekedési járművek vezetőit, a közlekedés szereplőit embereknek gondolták. Az önvezető autók prototípusainak megjelenésével a jelenben változik ez a helyzet. Az autonóm járműveket gépek (szoftverek) vezérlik, és izgalmas kérdés, milyen hasonlóságok és különbségek vannak az önvezető autók vezérlési rendszerei és emberek közlekedését szabályozó jogszabályok között. Erre a szempontra figyelve az ontológiai építkezés során igyekszünk olyan fogalmakat használni, amelyekkel képesek lehetünk mind az ember által irányított,

[1] Jelen tanulmány az Emberi Erőforrások Minisztériuma ÚNKP-17-3 kódszámú Új Nemzeti Kiválóság Programjának, valamint az OTKA K-116191 számú pályázatának támogatásával készült.

[2] 1/1975. (II. 5.) KPM-BM együttes rendelet a közúti közlekedés szabályairól (KRESZ)

mind az emberi vezető nélküli járművek szabályozását értelmezni. Jelen tanulmányban még csak a jogrendszer, és így a jogszabályok központi kategóriájának számító jogosultság- és kötelezettség-fogalmakat elemezzük, a – gépi vagy emberi – cselekvések vezérlését végző szabályrendszerre itt még jogrendszerként hivatkozunk. A jogosultság- és kötelezettség-fogalmak tárgyalásához szükség van a cselekvés, a deontikus modalitás, a norma, valamint a normatív pozíciók fogalmára. Ezek rövid bemutatása után bontjuk ki a jogszabályokban használható jogosultság- és kötelezettség-fogalmakat, amelyeket a KRESZ-ből vett példákkal szemléltetünk. A tanulmány második részében a közlekedés jogi szabályozásának pár további fontos kérdését elemezzük.

2. Cselekvés

A KRESZ, mint minden más normatív igényű szabályozás, az emberek cselekvéseire vonatkozik. Azzal a céllal hozunk létre normákat, hogy valamilyen cselekvési tartományban előre jósolható viselkedésekre számíthassunk. Kétféle cselekvést kell elkülöníteni egymástól: a tevést és a tartózkodást. Az előzőre példa lehet a haladni, kanyarodni, gyorsítani, lassítani, az utóbbira az állni vagyis nem haladni, elengedni valakit előttünk, nem zavarni másokat a közlekedésben. A tartózkodás nem pusztán valaminek a nem megtételét jelenti, tehát nem egyszerűen a tevés tagadása, szükséges még a nem-tevés szándékoltsága is. Számos dolgot nem teszünk meg, de ezek közül csak az számít tartózkodásnak, amelyet szándékoltan nem teszünk meg (a *szándékoltság* mozzanata természetesen a tevéshez is ugyanúgy hozzátartozik). A szándékoltság tényét sok esetben nagyon nehéz feltárni, „bizonyítani", de ezt a problémát itt most nem elemezzük (erről bővebben [1]).

3. Deontikus modalitás

Mivel a KRESZ is normarendszer, ezért a leírásához szükség van a normatív rendszerek központi kategóriájának, a deontikus modalitásnak az értelmezésére. Hat deontikus modalitást különíthetünk el:

- megengedett (permitted),
- tilos (prohibited),
- kötelező (obligatory),
- mellőzhető (omissible),
- opcionális (optional),
- preskriptív (prescriptive).

Ezen modalitások közül elég egyet (a megengedést) kiemelni definiálatlan alapfogalomként, és a többi már definiálható belőle (vö. [2]).

- tilos ≡ nem megengedett megtenni
- kötelező ≡ nem megengedett tartózkodni ≡ nem megengedett nem megtenni
- mellőzhető ≡ megengedett tartózkodni ≡ megengedett nem megtenni

- opcionális ≡ megengedett megtenni és megengedett nem megtenni
- preskriptív ≡ nem megengedett megtenni vagy nem megengedett nem megtenni

A hat modalitás közül az opcionális és a preskriptív „gyűjtőkategóriák”: a jogszabályokban ilyennel nem találkozunk, de a logikai viszonyok feltérképezésében jelentősége van a modalitások közé történő beemelésüknek. Mind az opcionalitásból, mind a kötelezőségből következik egy cselekmény megengedettsége; ugyanígy következik az opcionalitásból is és a tilosságból a mellőzhetőség; míg a kötelezőségből és a tilosságból is következik a preskriptív normatív státusz. A deontikus hatszög 3 kontradiktorikus „párt” tartalmaz: a kötelező és a megengedett, a preskriptív és az opcionális, illetve a kötelező és a mellőzhető kizárják egymást. De van ezen kívül még két fajta viszony: a kötelező, a tilos és az opcionális egymással kontrárius viszonyban vannak, azaz nem lehetnek egyszerre igazak, de lehetnek egyszerre hamisak; míg a megengedettség, a mellőzhetőség és preskriptív státuszokat szubkontrárius viszony „köti össze”: lehet belőlük egyszerre kettő igaz, de hamis nem.

4. Norma

A norma fogalmának elemzéséhez, a norma komponenseinek kibontásához Georg von Wright elméletét használhatjuk ([3]). Von Wright szerint a norma fogalma hat komponensre bontható:

- címzett (subject),
- kibocsátó (authority),
- karakter(character),
- tartalom (content),
- feltétel (condition),
- körülmény (occasion).

A norma *címzettje* az a cselekvő, akinek igazodnia kell a norma előírásaihoz, a *kibocsátó* az, aki a normát deklarálja, a norma *karaktere* a cselekvéshez kapcsolat deontikus modalitást jelenti, norma *tartalma* maga a szabályozni kívánt cselekvés, a *feltétel* a norma alkalmazásának feltételeül tételezett másik cselekvésre mutat(hat) rá, a *körülmény* pedig a norma alkalmazási feltételeit írja le. A von Wright-i modellt érdemes egy további komponenssel kiegészíteni, és felvenni a norma modelljébe a *szankció* jelenségét is:

- szankció.

A norma egy jövőre mutató elvárás, amelynek teljesülését a szankcionálás társadalmi eszközével biztosíthatjuk. Érdemes még azt megjegyezni, hogy a pontos modell értelmezéséhez szükségünk lehet a kibocsátás mint cselekvés, illetve a szankcionáló mint ágens terminusok felvételére. Ezek ontológiai elkötelezettség szempontjából nem tekinthetők új fogalomnak, hiszen a szankció fogalma

egyrészt a szankcionálás cselekedetére, másrészt a szankcionálást végző ágensre is utal, míg a kibocsátó ágens fogalma feltételezi a kibocsátás cselekedetét. A standard deontikus logika személytelen: nem kezel ágenseket. Az analitikus jogelméletnek azonban meghatározó, de az ágens-bázisú modellekkel és multi-ágens rendszerekkel való kézenfekvő összekapcsolhatósága miatt a mesterséges intelligencia és jog kutatásokban is megalapozónak tekintett Hohfeld rendszere a különböző jogosultság- és kötelezettségtípusokról. A tárgybeli hivatkozott relevanciáját és a deontikus logikai jelentőségét az adja, hogy ágenseket, egészen pontosan ágenspárokat kezel: adott jogosultságok és kötelezettségek korrelatív (vagy komplementer[3]) párjainak ágenseit. Tekintettel arra, hogy a közlekedés is kifejezetten multi-ágens rendszerként fogható fel, fontos tanulságokkal szolgálhat, ha ezen – az egyes jogosultságviszonyok két oldalán levő – normatív pozíciókat is beemeljük a fogalmi elemzésbe.

5. Jogviszonyok, normatív pozíciók

Wesley Newcomb Hohfeld elméletéhez mind a jog hagyományos elemzését végző jogelmélet, mind a jog formális reprezentálásával foglalkozó logika sokszor visszanyúl az elmúlt száz évben (többek között H.L.A. Hart [4], Stig Kanger [5],[6],[7], L. Lindahl [8], D. Makinson [9], Herrestad és Krogh [10]), és foglalkoznak ma is (pl. Sergot [11], van Eijck [12], Markovich [13], [14]). Hohfeld 1913-ban és 1917-ben publikált (két részes) tanulmányának – *Alapvető jogi fogalmak a bírói érvelésben* [15] – kiindulópontja, hogy a 'jog' ('right', azaz 'jogosultság', nem a jog mint rendszer) szó túl sokat és pongyolán használt – több különböző jogosultságviszonyra, amelyek így összemosódnak, holott igen különböző tartalmú normatív pozíciókról van szó. Ezért felvázolt egy rendszert, amelyben a jogosulság-fogalma alá rendelte az *igény(jog)*, a *privilégium* (szabadság), a *felhatalmazottság* és az *immunitás* fogalmait; míg a kötelezettség alá a *kötelesség*, az *igényhiány*, a *beavatkozásnak kitettség* és a *beavatkozásképtelenség* fogalmait. Hohfeld kategóriái páronként korrelatív vagy komplementer, illetve ellentétes viszonyban állnak egymással, mely viszonyokat mutatja az alábbi ábra.

[3] Az eredeti hohfeldi szóhasználtaban és a fordításában is 'korrelatív' szerepel, de [2] pl. 'komplementer' viszonyként hivatkozik rá.

1. ábra: Hohfeld jogosultság- és kötelezettségfogalmai.

A korrelatív viszony lényege, hogy egy ágens igényjoga mindig együttjár egy másik ágens kötelezettségével, ahogyan valaki privilégiuma valaki más igényhiányával és így tovább. Logikailag ez ekvivalenciát jelent. Az ellentét pedig negációt: pontosan akkor igaz, hogy egy ágensnek privilégiuma van egy cselekvésre, ha nem áll fenn a másik ágensnek az adott cselekvéstől való tartózkodásra vonatkozó igényjoga. Az ábrán látható, hogy a kategóriák eme viszonyaik mentén két csoportba rendeződnek. Nagyon hasonlónak tűnik a két csoport, de valójában komoly különbségek vannak köztük. Ha a modalitás tetszőleges, egyszerű, valós (fizikai) cselekvésekre vonatkozik, akkor elsőrendű fogalomról beszélhetünk – ezek vannak a bal oldali négyszögben, míg ha a modalitás a jogi pozíciók (szabályok) létrehozására, megváltoztatására, megszüntetésére irányul, akkor másodrendű fogalomról van szó – ezek vannak a jobb oldalon. A hohfeldi fogalmak jelentésének szemléltetésére a KRESZ területéről a következő példák hozhatók.

6. A hohfeldi pozíciók megjelenés e a KRESZ-ben

6.1. Igényjog és kötelezettség

A hohfeldi fogalmakat a páronkénti ekvivalencia következtében – értelemszerűen – párban érdemes kezelni: bármely pozícióra vonatkozó nyelvi utalást találunk, azzal annak korrelatív párját is „megtaláltuk”. Az igényjog-kötelezettség páros legszemléletesebb megjelenése a KRESZ-ben az elsőbbségadás szabályaiban van, hiszen itt kézenfekvő, hogy két ágens egymással szembeni normatív pozícióiról van szó:

> t) „Egyenrangú utak kereszteződése” (88. ábra); a tábla azt jelzi, hogy az úton olyan útkereszteződés van, amelyben a jobbról érkező járműnek van elsőbbsége; [16. § (1)][4]

Amikor ilyen helyzetbe kerülnek a közlekedők, akkor a jobbkéz-szabály jogosultjának igényjoga az, hogy a másik fél adjon neki elsőbbséget, vagyis a másik

[4] Ez az idézett rész és a továbbiak a hivatkozott jogszabályból (1/1975. (II. 5.) KPM-BM együttes rendelet a közúti közlekedés szabályairól (KRESZ)) származnak.

kötelessége a jogosult „előre engedése". Az igényjog másik - talán kevésbé kézenfekvő - példájaként hivatkozhatunk a fizetős utak igénybevételének szabályára.

> (4) Az (1) bekezdés a) pontjában említett jelzőtábla alatt elhelyezett kiegészítő jelzőtábla (8/a. ábra) azt jelzi, hogy az út úthasználati díj megfizetésével vehető igénybe. [11. §]

Az igényjog közvetett módon érhető itt tetten. Amíg az ingyenes úton szabadon lehet közlekedni, nincs senkinek igénye a közlekedéstől való tartózkodásra - és ez szép példája a privilégium-igényhiány párra, addig a fizetős úton a közlekedési útvonalat fenntartó szereplő azzal az igénnyel lép fel a közlekedőkkel szemben, hogy fizessék meg a megfelelő díjat az út használatáért, vagyis az ott közlekeződknek van egy ilyen fizetési kötelezettségük az út fenntartójával szemben. Fontos látnunk, hogy amikor Hohfeld kötelezettségről beszél, nem specifikálja a cselekvést tevésre és tartózkodásra, így a deontikus logika kötelező és tilos modalitásainak megjelenései mind besorolhatók a hohfeldi kötelezettség fogalom alá. Az általános, az adott közlekedési szituációt nem két ágens viszonyában leíró szabályoknál kevésbé szembetűnő, hogy kezelhetők a hohfeldi fogalmakkal, pedig arról van csak szó, hogy a közlekedésben részt vevő valamennyi ágensnek igényjoga van a közlekedésben résztvevő valamennyi másik ágens irányába a kötelező cselekmények végrehajtására és a tilos cselekményektől való tartózkodásra.

Tilos A tilalom és a kötelezőség modalitások nyelvi megjelenésének formáiról a magyar (jogi) nyelvben már készültek kutatások (lásd pl. [16], [17], [18]), a KRESZ-ben a következő példákat találjuk: A *tiltott* modális operátort nyelvi szinten a 'tilos' különböző alakjai jelölhetik.

> (5) A hatósági jelzés(eke)t [rendszámtábla(áka)t]... megváltoztatni, letakarni vagy jogosulatlanul eltávolítani tilos. [5. §]

Mivel logikailag a tilos modalitás a megengedett tagadását jelenti, ezt nyelvi szinten is kifejezhetjük hasonló módon.

> (2) A gépkocsi vezetője menet közben kézben tartott mobil rádiótelefont nem használhat. ... [3. §]

A tiltást jelölheti még a 'nem lehet' vagy a 'nem megengedett' kifejezés is.

> (9) A magyar hatósági jelzéssel ellátott gépjárművön – ide nem értve a DT és a CK betűjelű rendszámtáblával közlekedő gépjárművet - nem lehet olyan jelzést - így különösen CD jelzésű matricát - feltüntetni, amely arra utal, hogy a gépjármű a Magyarországon működő diplomáciai és konzuli képviseletek, nemzetközi szervezetek vagy ezek nemzetközi szerződés alapján mentességet élvező tagjának, illetve a mentességet élvező családtagjának használatában van. [5. §]

Megint más nyelvi kifejezést (a 'továbbhaladás tilalmát jelzi') láthatunk a tiltásra a következő idézetben.

a) a rendőr mindkét karjának oldalirányú kinyújtása...
- a kinyújtott karra merőleges irányból érkezők részére a továbbhaladás tilalmát jelzi: a gyalogos az úttestre nem léphet; a járművel az útkereszteződésnél - a kijelölt gyalogosátkelőhely, illetőleg a megállás helyét jelző útburkolati jel, ezek hiányában az útkereszteződés előtt - meg kell állni; [6. § (1)]

A fenti passzusban van még egy új elem: a továbbhaladás tilalmát egy másik deontikus operátor (a kötelezés) segítségével fejezi ki a mondat: 'meg kell állni'. Ezt azért tehetjük meg, mert a deontikus modalitások kifejezhetők egymással. Itt úgy, hogy az aktív cselekvést, tevést jelentő 'áthaladás' predikátuma helyett egy tartózkodás kifejező predikátumot alkalmazunk (a 'megállás'-t). Az ekvivalencia pedig így állítható fel: tilos továbbhaladni = kötelező megállni. A tiltás ki lehet fejezni megengedő modalitással is akkor, ha más szavakkal írunk elő korlátozást.

23. § (1) Az utas csak álló járműre szállhat fel és csak álló járműről szállhat le.

A normamondat azt akarja kifejezni, hogy csak az álló járműre vagy álló járműről való fel- vagy leszállás megengedett. A mondatban szereplő 'csak' szó beillesztésével a felszínen a megengedés jelző modalitás átfordul az ellentétébe, és tiltás lesz belőle.

Kötelező A kötelezés modalitás leggyakrabban a 'kötelező' jelző vagy a 'kell' segédige illesztésével oldja meg a jogalkotó.

(3) Az (1) bekezdésben említett engedélyt a vezetőnek vezetés közben magánál kell tartania. [4. §]

(4) A jármű hatósági engedélyét - ha jogszabály eltérően nem rendelkezik - a vezetőnek vezetés közben magánál kell tartania.
(5) A hatósági jelzés(eke)t [rendszámtábla(áka)t] a járművön a felszerelésére kijelölt helyen, jól olvasható állapotban kell tartani, azt megváltoztatni, letakarni vagy jogosulatlanul eltávolítani tilos. [5. §]

Természetesen más nyelvi alakban is kötelezhetünk valamire (a következő esetben az 'ad utasítást' kifejezés segítségével).

f) ha a rendőr kinyújtott karjával maga felé int, ezzel a sebesség növelésére, ha pedig kinyújtott karját le- és felfelé mozgatja, a sebesség csökkentésére ad utasítást. [6. §]

Lehetséges azonban modális operátor direkt használata nélkül is kötelezést kifejezni.

(2) Ha a rendőr a forgalmat nem útkereszteződésben irányítja,
a) a járművek megállítására
aa) karjának vagy fehér szegélyű piros tárcsának függőleges feltartásával, piros vagy fehér fény körbeforgatásával, illetőleg síppal ad jelzést, [6. §]

A fenti mondatban mindenféle modális operátor alkalmazása nélkül, kijelentő módban van kifejezve az utasítás. A kötelezés és az utasítás közti szemantikai kapcsolatot „kihasználva” megint csak más nyelvi formában jelenik meg a kötelezés az alábbi példában (utasításnak megfelelően 'kell eljárnia'). Mivel itt nem egy konkrét cselekvést kell szabályozni, hanem egy általánosított, sok mindenre kiterjedő cselekvéshalmazra kell kötelezni, ezért „”kell” az általánosított formát használni.

> b) a megkülönböztető fényjelzést használó gépkocsiból vagy motorkerékpárról, illetőleg rendőrségi helikopterről hangszórón keresztül adott, a követendő közlekedési magatartásra vonatkozó utasításának megfelelő módon kell az úton közlekedő jármű vezetőjének eljárnia. [6. § (2)]

A következő mondat minősége a predikátum alapján nyelvileg megengedőnek tűnhet, pedig kötelezőséget akartak kifejezni vele. A nyelvi szinten megengedést jelölő 'haladhatnak' kifejezést a 'mindig' terminus alakítja át kötelező minősítéssé. A 'mindig' helyett a csak' vagy a 'kizárólag' terminus is ugyanezt a hatást éri el.

> (3) Ha a gyalogosok az úttesten közlekednek, az úttest szélén, egysorban
> a) lakott területen lehetőleg a menetirány szerint a bal oldalon, lakott területen kívül mindig a bal oldalon, a járműforgalommal szemben,
> b) a segédmotoros kerékpárt vagy kerékpárt toló, továbbá a kerekes székkel közlekedő mozgáskorlátozott gyalogosok mindig a menetirány szerint a jobb oldalon haladhatnak. [21. §]

A 'mindig/csak/kizárólag jobb oldalon haladhatnak' értelme az, hogy kötelezően a jobb oldalon kell haladni. A kötelezést a megengedés modalitás és egy feltétel (úthasználati díj fizetése) segítségével is ki lehet fejezni.

> (4) Az (1) bekezdés a) pontjában említett jelzőtábla alatt elhelyezett kiegészítő jelzőtábla (8/a. ábra) azt jelzi, hogy az út úthasználati díj megfizetésével vehető igénybe. Az (1) bekezdés a) és b) pontjában említett jelzőtábla alatt elhelyezett kiegészítő jelzőtábla (8/b. ábra) az úthasználati díjfizetési kötelezettség végét jelzi. [11. §]

Az alábbi idézet két előírása mutatja, hogy ugyanaz a szabályozási igény többféle nyelvi formában is kifejezhető (két egymást követő bekezdésben).

> (6) A külföldre távozó magyar rendszámú gépjárművet - a „H” betűt feltüntető - államjelzéssel kell ellátni.
> (7) Magyarország területén csak olyan külföldi rendszámú gépjármű közlekedhet, amelyet a rendszámot kiadó állam jelzésével elláttak. [5. §]

A kötelezést is ki lehet fejezni bizonyos szavak alkalmazásával akkor, amikor a felszíni szerkezet megengedő modalitást fejezne ki.

> (3) Ha a gyalogosok az úttesten közlekednek, az úttest szélén, egysorban
> a) lakott területen lehetőleg a menetirány szerint a bal oldalon, lakott területen kívül mindig a bal oldalon, a járműforgalommal szemben,
> b) a segédmotoros kerékpárt vagy kerékpárt toló, továbbá a kerekes székkel közlekedő mozgáskorlátozott gyalogosok mindig a menetirány szerint a jobb oldalon haladhatnak. [21. §]

A 'haladhatnak' kifejezés önmagában még megengedésre utalna, de a „mindig a baloldalon" kifejezés beszúrása a szabályozást átváltja kötelezéssé.

6.2. Privilégium (szabadság) és igényhiány

A szabadság komplementere az igényhiány, ami annyit tesz, hogy két cselekvő közötti viszonyban az egyik cselekvése attól szabad, hogy a másiknak nincs arra vonatkozó igényjoga, hogy az első ágens ettől a cselekvéstől tartózkodjon. A szabadságnak ez az elemi fogalma (tehát nem a védett szabadságok, ún. szabadságjogok) mindösszesen a vonatkozó kötelezettség hiányát jelenti, ami megfelel a deontikus modalitások megengedettségének. S mint ilyen, a magyar (jogi) nyelvben leggyakrabban az igei predikátum hatóképzővel módosított alakjában érhető tetten:

> 4. § (1) Járművet az vezethet, aki a) a jármű vezetésére jogszabályban meghatározott, érvényes engedéllyel rendelkezik, és a jármű vezetésétől eltiltva nincs; b) a jármű biztonságos vezetésére képes állapotban van, továbbá c) a vezetési képességre hátrányosan ható szer befolyása alatt nem áll, és szervezetében nincs szeszes ital fogyasztásából származó alkohol.

Természetesen más módon is jelezhetjük, hogy valakinek szabadsága van a cselekvésre.

> a) a rendőr mindkét karjának oldalirányú kinyújtása
> - a kinyújtott karral párhuzamos irányból érkezők részére szabad utat jelez: a gyalogos az úttesten áthaladhat; a járművel - ha jelzőtábla nem tiltja - az erre vonatkozó szabályok megtartásával egyenesen, valamint jobbra vagy balra bekanyarodva tovább szabad haladni; [6.§ (1)]

Az idézetben a szabad cselekvés négyféle formában is megjelenik: 'szabad utat jelez', 'áthaladhat', 'nem tiltja', 'szabad haladni'. További példa az engedélyezettség konkrét nyelvi megjelenése:

> b) a villogó zöld fény az áthaladásra engedélyezett időtartam végét jelzi: a gyalogos az úttestre nem léphet, az úttesten levő gyalogos pedig köteles az áthaladást mielőbb befejezni; [8. § (2)]

A fenti paragrafus ugyan tiltást fejez ki, közvetett módon szerepel egy szabadságra vonatkozó mozzanat is, amit az 'áthaladásra engedélyezett' kifejezés jelez. Ezekben a példákban a norma címzettje szabadon dönthet arról, végrehajtja-e a cselekvést vagy sem, és a cselekvés végrehajtását egy másik cselekvő se nem akadályozhatja meg, se nem írhatja elő.

6.3. Felhatalmazottság és bevatkozásnak-kitettség

A felhatalmazottság azt jelenti, hogy az adott szereplő megváltoztathatja mások normatív pozícióit. A legparadigmatikusabb, felhatalmazottsággal rendelkező szereplő ebben a térben (is) a jogalkotó. Mivel a KRESZ-t itt adottnak vesszük, nem foglalkozunk a változtatásaival, ezért itt a jogalkotó szerepére nincs mód (és nincs szükség) reflektálni. A jogalkotó a közlekedési helyzetekben nem vesz részt, a KRESZ egészéhez képest külsődleges a viszonya. Vannak azonban a KRESZ világán belül komoly felhatalmazottsággal rendelkező szereplők. A legfontosabb ilyen ágens a rendőr. A rendőr közlekedési szerepe sok esetben az, hogy a felhatalmazottságára támaszkodva megváltoztatja a közlekedők jogi pozícióit – azaz a közlekedők kitettek a rendőr beavatkozásával szemben: az kötelezettséget róhat rájuk, igényjoggal ruházhatja fel őket. A rendőr felhatalmazottságát nyelvileg direkt módon nem könnyű kimutatni a jogszabályban, de implicit módon természetesen tetten érhető.

> a) a rendőr mindkét karjának oldalirányú kinyújtása
> - a kinyújtott karral párhuzamos irányból érkezők részére szabad utat jelez: a gyalogos az úttesten áthaladhat; a járművel - ha jelzőtábla nem tiltja - az erre vonatkozó szabályok megtartásával egyenesen, valamint jobbra vagy balra bekanyarodva tovább szabad haladni; [6. § (1)]

A rendőr az, aki a karjelzéseivel megváltoztatja a közlekedők jogait, illetve kötelezettségeit az adott közlekedési színtéren belül. A következő idézetből is a rendőri felhatalmazottságra vonatkozó szabályt olvashatunk ki.

> (2) Ha a baleset halált vagy személyi sérülést okozott, a balesetről a rendőrhatóságot (a legközelebbi rendőrt) haladéktalanul értesíteni kell és a lehetőséghez képest gondoskodni kell a nyomok megőrzéséről; a helyszínt csak az intézkedő rendőr engedélyével szabad elhagyni. [58. §]

Az idézetben kétféle normatív állítást találhatunk. Az első, explicit módon nem megfogalmazott elvárás szerint mindaddig, amíg az intézkedő rendőr nem ad engedélyt, a balesetben érintett vezetőknek a helyszínen kell maradni (a helyszínt nem szabad elhagyni). Az alapértelmezett jogi pozíció tehát a helyszínen maradás kötelessége. A második normatív aktus szerint, ha ezt a normatív pozíciót, vagyis a vezető számára előírt maradási kötelezettséget az intézkedő rendőr – felhatalmazottsága révén – megváltoztatja, és engedélyt ad a távozásra, akkor a vezetők szabadsága érvényesül újra, és elhajthatnak a helyszínről. A felhatalmazottság birtokosa lehet persze más is, ahogy az alábbi példa mutatja.

> (8) Rendkívüli időjárás, természeti körülmények (havazás, jegesedés, árvíz stb.) esetén a közúti járműveknek Magyarország területére történő beléptetése - külön jogszabály alapján - feltételhez köthető vagy korlátozható. [5. §]

A felhatalmazott szereplő itt a szabályzó hatóság, amely rendkívüli helyzetekben dönthet a közlekedési szabályok – ideiglenes – megváltoztatásáról. A KRESZ meglepően sok esetben ad – bizonyos cselekvési területeken – felhatalmazottságot a közlekedési helyzetek adott résztvevőinek. Vegyük a következő példát:

> z/3. A behajtási tilalom alól az út kezelője felmentést adhat. A kiegészítő jelzőtáblán „Kivéve engedéllyel” felirat is elhelyezhető. [14. § (1)]

Ebben a passzusban azt deklarálják, hogy az alapértelmezett jogi pozíciót, a behajtás tilalmát megváltoztathatja az út kezelője, és engedélyezheti az út használatát másoknak. Előfordulhat olyan helyzet is, amikor a jogszabály az érintettek felhatalmazottságra vonatkozóan ír elő – immár harmadrendű – kötelezettségeket, ahogy ez a következő idézetből kiderül:

> (3) Ha a jármű az (1) bekezdésben foglalt rendelkezéseknek nem felel meg;
> a) az üzemben tartó nem engedheti meg, hogy a vezető a járművel elinduljon és az elindulást köteles megakadályozni;
> b) a vezető a járművel nem indulhat el és az elindulásra vonatkozó utasítás végrehajtását köteles megtagadni. [5. §]

A fenti idézet nyilván a fuvarozó cégek munkahelyi viszonyaira vonatkozik. Az alapértelmezett helyzet az, hogy a jármű üzembentartója engedélyezheti (előírhatja) a járművezető számára a jármű használatát. Az idézett normamondat a) pontja az üzembentartó számára ír elő tiltást, amely egyben megvonja az üzembentartó – általában létező – felhatalmazottságát az adott esetre vonatkozóan, míg a b) pont a vezető számára írja elő az üzembentartó felhatalmazottságának elismerésére vonatkozó kötelezettség végrehajtását.

6.4. Immunitás és bevatkozás-képtelenség

Az immunitás, azaz a kitettség hiánya a KRESZ-ben olyan helyen keresendő, ahol az a személy, akinek felhatalmazottsága lenne, nem bír ezzel: vagyis olyan példát érdemes keresni, amikor a rendőr ebbéli lehetőségei korlátozva vannak. Van is ilyen:

> (9) A magyar hatósági jelzéssel ellátott gépjárművön - ide nem értve a DT és a CK betűjelű rendszámtáblával közlekedő gépjárművet - nem lehet olyan jelzést - így különösen CD jelzésű matricát - feltüntetni, amely arra utal, hogy a gépjármű a Magyarországon működő diplomáciai és konzuli képviseletek, nemzetközi szervezetek vagy ezek nemzetközi szerződés alapján mentességet élvező tagjának, illetve a mentességet élvező családtagjának használatában van. [5. §]

A diplomáciai mentesség miatt a diplomatakocsik vezetőit a szabályok megsértése esetén sem vonhatja felelősségre a közúti rendőr, ami a diplomaták immunitását jelzi. A másik – megint csak nem explicit – példa arra, hogy a rendőr immunitást élvez saját jogosultságait illetően az, hogy a közlekedés szereplői közül nem utasíthatja senki sem őt arra, hogy valamilyen jogi pozíciót megváltoztasson (pl. feloldjon egy tilalmat vagy elengedjen egy kötelezettséget).

7. További feladatok

Az ontológia elemeinek kiválasztása és a választás legitimizálása természetesen csak az első lépés egy olyan jogi ontológia felépítésében, amely segít az adott

normatív rendszerek modellezésében. A modalitások számítógépes beazonosításában elért eredmények mellé további kutatást igényel, hogy a hohfeldi viszonyok gép általi felismerhetőségét, az ágensekhez rendelést automatizálható algoritmusokkal segítsük.

Hivatkozások

1. Belnap, N., Perloff, M., Xu, M., eds.: Facing the Future. Agents and Choices in our Indeterminist World. Oxford University Press (2001)
2. Syi: syi.hu/cse. L'Harmattan–Könyvpont (2014)
3. von Wright, G.H.: Norm and Action. Routledge and Kegan Paul, London (1963)
4. Hart, H.: Bentham on legal rights. In Simpson, A., ed.: Oxford Essays in Jurisprudence. Second Series. Clarendon Press, Oxford (1973) 171–201
5. Kanger, S., Kanger, H.: Rights and parliamentarism. Theoria **32** (1966) 85–115
6. Kanger, S.: New foundations for ethical theory. In Hilpinen, R., ed.: Deontic Logic: Introductory and Systematic Readings. Reidel (1971) 36–58
7. Kanger, S.: Law and logic. Theoria **38** (1972) 105–132
8. Lindahl, L.: Position and Change. Springer (1977)
9. Makinson, D.: On the formal representation of rights relations. Journal of Philosophical Logic **15** (1986) 403–425
10. Herrestad, H., Krogh, C.: Obligations directed from bearers to counterparties. In: Proceedings of the International Conference on Artificial Intelligence and Law, ACM (1995) 210–218
11. Sergot, M.: Normative positions. In Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L., eds.: Handbook of Deontic Logic and Normative Systems. College Publications (2013) 353–406
12. van Eijck, J., Ju, F.: Modelling legal relations with dynamic logic. Under submission (2016)
13. Markovich, R.: Agents, Roles and Rights. A Formal Approach to the Hohfeldian Conceptions. In: Proceedings of Advances in Modal Logic. Short Papers. (2016)
14. Markovich, R.: Nem házasságközvetítés, bikondicionálisok. formulák, az állam és az emberi jogok. In Ződi, Z., Hegyi, S., Vinnai, E., eds.: Orbis Iuris. Bíbor Kiadó (2016)
15. Hohfeld, W.N.: Alapvető jogi fogalmak a bírói érvelésben. In Szabó, M., Varga, C., eds.: Jog és nyelv. Kiadó nélkül (2000) 59–96
16. Vinnai, E.: A jogi nyelv nyelvészeti megközelítései. Sectio Juridica et Politica **28** (2010) 145–171
17. Kaczmarek, K.: Az utasítás kifejezőeszközei a jogi szakszövegben a deontikus modalitás tükrében. Filológia **1**(4) (2010) 224–238
18. Markovich, R.: Tilalom és kötelesség—a számítógép szemével. In Szabó, M., ed.: A jog nyelvi dimenziója. Bíbor Kiadó (2015)

Relevance Segmentation of Long Documents

Zsolt Szántó, Alex Sliz-Nagy, István Nagy T., Ádám Csuma-Kovács,
Veronika Vincze, Richárd Farkas

Black Swan Hungary,
Szeged, Tisza Lajos krt. 47.,
szeged@blackswan.com

Abstract: In this paper, we present our methods to identify the most salient topics for a selected domain based on topic modeling. We propose a topic relevance score and segmentation procedure which can split the document into parts referring to various topics. We also offer a solution for visualizing textual spans that are related to a given topic. In this way, it can be easily determined which are the most relevant and most irrelevant segments of a long document (like blog posts or news articles).

1 Introduction

Nowadays, a huge amount of textual data is published every day on the internet, in the form of weblogs, social media posts, posts on official websites etc. However, the large amount of data makes it impossible to process it manually -- for instance, when the user is interested in a topic, the number of documents related to the specific topic might be overwhelming and thus, no human can easily find all the relevant documents. The problem might be even more difficult, considering that a single document can contain several topics itself, some (or all) of which might be relevant to the user. If only a smaller segment of a long document, like a blog post or news article, is relevant to the user, he should not waste time with reading the whole document.

There are various solutions for reducing the human processing time of a long document, for instance, keyphrase extraction (which assigns a number of short phrases to documents, representing the content) [1,4] or document summarization (which offers a few sentence long summary of the whole document) [6].

In this paper, we propose an automatic document segmentation and relevance visualization tool designed for long documents. Our solution is built on the top of LDA topic modelling. The information target of the user is defined through a set of keywords. We first rank the LDA topics by relevance to the input of the user, then segment a long document by assigning a topic for each word smoothed in word sequencing. We also offer a solution for visualizing textual spans that are related to a given topic. In this way, it can be easily detected how many topics occur in a document, which topics are the most salient ones and which are more marginal. We demonstrate our system on Hungarian news articles.

2 Literature review

Topic modeling aims at discovering the abstract topics that occur in a document or set of documents. In other words, it discovers the hidden semantic structures of a text: certain words are expected to occur in connection with certain topics, and their presence strongly indicates that the document is about that given topic. Documents usually consist of several topics, and the most salient topics can easily be identified with topic modeling.

Latent semantic indexing (LSI) is a widely used method to transform the original document vectors to small-dimension vectors [3]. In probabilistic LSI, each word in a document can be seen as a pattern of a mixed model, composed of different topics. Thus, a document can be seen as a mix of different topics.

Another model for representing topics is Latent Dirichlet Allocation (LDA) [2]. It is a generative statistical model allowing observations to be explained by unobserved groups for data similarity. Each document is seen as a mixture of a small number of topics and each word is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model for topic discovery.

Topic segmentation has also been widely investigated. An early attempt to identify topic boundaries in free texts was reported in [7]. A new hierarchical Bayesian model was proposed for unsupervised topic segmentation [5] and LDA is also frequently used for text segmentation [8].

3 Methodology

For our experiments, we downloaded approximately 100,000 documents from different Hungarian news portals (e.g. www.index.hu, www.origo.hu) then tokenized and lemmatised them with magyarlanc [9].

Our method requires two databases, the first one is a set of documents and the second is a manually created set of words that describe a topic. We would like to calculate similarity between this predefined topic and the documents.

3.1 Preprocessing

We used standard preprocessing methods, we removed the stopwords, punctuations, and numbers and we lemmatized [9] the corpora.

3.2 Ranking of the documents

First we ranked the topics based on the similarity between a topic and our predefined topic indicator word list. We ran LDA on the lemmatized dataset and we calculated r_t ranking score to each topic t, with the following equation:

$$r_t = \sum_{w <- indicator\ words} p(w|t)$$

The topics with highest r_t are more similar to our predefined word list.

Based on this topic ranking scores we can get a ranking over the documents. The r_d document ranking score for document *d* is derived from this equation:

$$r_d = \sum_{t <- topics} p(d|t) * r_t$$

Now we have a ranking over the documents where the top of this list is highly related to the predefined topic.

3.3 Topic detection inside the document

Some of the documents contain more than one topic, so we developed a method which can extract these topics inside a document. We used Hidden Markov Model (HMM) to separate different topics in one document. In our HMM the words are the observations and the topics are the hidden states, each observation are generated from a hidden state (like in POS tagging). The HMM requires two parameter matrices as input, the emission probability matrix that describe a distribution of the observations over the hidden states and the transition probability matrix that describes the probability of a transition between two states.

The LDA calculates the distribution of words over a topic, which we can directly use as the emission probabilities in HMM.

For transition probability we only use two values, one for the situation when we keep the previous state, one when we change the state:

$$p(t_i, t_j) = \begin{cases} \alpha, & i = j \\ \dfrac{1-\alpha}{|T|-1}, & otherwise \end{cases}$$

where α is a parameter that determines the size of the topics.

By using the Viterbi algorithm, we can get the topic sequence over the words of document d with the highest probability.

4 Results

In our experiments, we focused on two domains, namely, sports and music. We used Wikipedia-based lists to construct an initial seed list for describing the domains. Here we present our results for ranking topics in connection with the domains and we also report on how documents can be segmented on the basis of topics present in the document.

id	top words					score
39	magyar Hungarian nemzetközi international	olimpiai olympic méter meter	verseny competition nyer win	olimpia Olympics két two	szövetség association hosszú Hosszú	5.2501
9	magyar Hungarian két two	nyer win női female	első first második second	hely place döntő final	csapat team férfi male	4.1828
6	hely place kör lap	első first futam race	autó car tud can	verseny race második second	két two motor motor	4.1565
0	meccs match első first	csapat team játékos player	két two szezon season	pont point nyer win	ben in Nfl NFL	4.0798
47	oldal site twitter Twitter	tud can ember man	facebook facebook szeptember September	fotó photo 24hu 24.hu	címlap title page kép picture	4.0561
2	csapat team mérkőzés match	meccs match gól goal	válogatott national team bajnokság league	játékos player klub club	játszik play pálya pitch	4.0494
28	terület area épít construct	épület building projekt project	terv plan épül build	gép machine beruházás investment	város city Park Park	4.0454
22	manos Manos annyi many	film film látható visible	nap day munka work	szereplő actor néz watch	sztori story Lát See	4.0060
34	magyar Hungarian László László	egyetem university György György	ben in budapesti Budapest	budapest Budapest alapítvány foundation	István István Program Program	4.0018
35	kutató researcher talál find	kutatás research állat animal	föld earth tanulmány study	egyetem universiry tudományos academic	víz water eredmény result	4.0018

Table 1: Salient topics related to sports.

id	top words					score
46	dal song fesztivál festival	szám hit videó video	zene music lemez record	zenekar band című entitled	koncert concert album album	10.929
23	kis small étterem restaurant	hely place kicsi small	például for example egészen totally	szép nice két two	inkább rather név name	8.7910
40	fotó photo tér square	kép picture múzeum musem	épület building János János	magyar Hungarian László László	Hungarian Budapest József József	8.1943
15	ben in három three	kap get idén this year	díj prize vesz buy	első first név name	két two nyer win	8.0201
10	film film mozi movie	című entitled történet story	rendező director rész part	sorozat series jelenet scene	színész actor néző audience	7.9871
6	autó car kör lap	hely place második second	első first futam race	tud can motor motor	két two hamilton Hamilton	7.9504
37	európai European bizottság committee	magyarország Hungary uniós Union	magyar Hungarian unió Union	ország country európa Europe	kormány government orbán Orbán	7.9076
32	manos Manos annyi many	film film látható visible	nap day munka work	szereplő actor néz watch	sztori story Lát See	7.9054
42	gép machine első first	föld earth tud can	hajó ship repülőgép airplane	két two nap day	kutató researcher tudós scientist	7.8989
19	facebook Facebook tud can	oldal site telefon telephone	rendszer system eszköz device	cég company felhasználó user	használ use google Google	7.8964

Table 2: Salient topics related to music.

Tessék, lányok, akik nem szültek, mégis rengeteg gyereket tettek újra nagyon boldoggá
Megmentették a három új-zélandi tehenet, akiket a földrengés vágott el a világtól
Végre egy budapesti építkezés, ahol az épülethálón nem reklám van, hanem a leendő homlokzat
Ha Németh Szilárd cikket olvas, a brazíliai őserdőkben felsír egy fa
Mindent az olvasóért: megkóstoltam a Nutellaburgert, hogy önöknek már ne kelljen
Rogán azt állítja, kávézni ugrott fel régi ismerőséhez. Vajon ezúttal sikerült igazat mondania?
Tessék, lányok, akik nem szültek, mégis rengeteg gyereket tettek újra nagyon boldoggá
Emlékszik még a Balatont hosszában átúszó lányokra?
Nem csak úszásban és példamutatásban jók, de most a Kispingvin Úszó és Vízilabda Előkészítőben úszó gyerekekkel közösen gondoltak egyet és nekiálltak rászoruló gyerekeknek ajándékot gyűjteni. Ez olyan jól sikerült, hogy a Magyar Máltai Szeretetszolgálat Családok Átmeneti Otthonában élő mind a 87 gyermeknek jutott egy hatalmas névre szóló doboz! A dobozokat a Máltai Szeretetszolgálat Jézuskája karácsonykor a megfelelő fa alá fogja tenni.
Ez a világ legnagyobb magányosan álló jármű-monstruma
Németország keleti részében található meg a mai napig ez a hatalmas, 112 métere hosszú, 57 méter magas szerkezet. Melyet a kelet-német ipar egykori dicsőjét mutatja. Ez a hatalmas kotrógép jelenleg a világ legnagyobb magára hagyott járműve. A gép elején látszódnak hatalmas "fogai" vagyis vödrei, amivel szó szerint ette a földet. Ezekkel darabonként 15 köbméter földet tudtak kibányászni egy kör alatt anno.

Fig. 1. Sample text for visualizing topics related to sports.

4.1 Ranking topics for domains

The ten most salient topics related to sports are presented in Table 1. As can be seen, the first 6 topics contain lots of sports words. There seems to be one outlier topic, which is ranked 5th: topic 47 is related to social media (including words like Facebook and Twitter). However, sports events are often advertised and reported in social media, hence the frequency of social media vocabulary can be easily explained in the sports domain as well. Also, topic 28 describes construction works, which again might be connected to sports, for instance, when constructing buildings for sports facilities such as stadiums, sports halls or football pitches.

The ten most salient topics for music are presented in Table 2. The first topic is unambiguously related to music. There is a huge gap between the scores for the first and the second most salient topic, which suggests that the vocabulary of music is utterly distinct from all the other topics. However, topics 15 and 10 might be also loosely related to music, as there are music awards where prizes can be won (topic 15) and films are also accompanied with music (topic 10). Moreover, topic 0 may be of relevance as well: there are several recent reports on sexual harassment from the entertainment industry, so unfortunately a topic on sexual abuse can also be connected to the music domain.

4.2 Document segmentation

Here we illustrate our results on segmenting the documents on the basis of the topics mentioned. For this purpose, we made use of the *Mindeközben* (Meanwhile) column of the news portal *index.hu*, which includes short pieces of news of miscellaneous topics, hence they are supposed to contain multiple topics.

Figure 1 shows a sample from a document where different topics are marked with different colors. Text spans which are related to sport and have a high position in the topic ranking are highlighted with green (and with bold font).

Figure 2 shows another sample from the *Mindeközben* column. Here, green spans denote textual content related to music. As can be seen, the first sentence of the document contains an invitation to a music festival, which is then followed by titles of other short news. Later, the music festival is described in full detail, which was also identified as belonging to the music topic by the algorithm.

Kíváncsi, hogy mulattak a francia udvarban? Menjen a Régizene Fesztiválra, ahol olyat hallhat, amit 274 éve senki

Megmentették a három új-zélandi tehenet, akiket a földrengés vágott el a világtól

Végre egy budapesti építkezés, ahol az épületháló nem reklám van, hanem a leendő homlokzat

Ha Németh Szilárd cikket olvas, a brazíliai őserdőkben felsír egy fa

Mindent az olvasóért: megkóstoltam a Nutellaburgert, hogy önöknek már ne kelljen

Rogán azt állítja, kávézni ugrott fel régi ismerőséhez. Vajon ezúttal sikerült igazat mondania?

Március 2-án kezdődik a Müpában 2. Régizene Fesztivál, s mivel már az előzőért is nagyon odavoltunk , nem mehetünk el szó nélkül a mostani mellett sem. Bár lesznek azért alapdarabok is a műsoron (a Capella Savaria például Bach hat Brandenburgi versenyét adja elő), a központi téma a Magyarországon nem annyira ismert francia barokk lesz.

A legnagyobb durranásnak Mondonville Isbé című operájának koncertszerű előadása ígérkezik vasárnap az Orfeo Zenekar és a Purcell Kórus közreműködésével, Vashegyi György vezényletével.

Ezt az operát ugyanis keletkezése, 1742 óta még soha nem tűzték műsorra!

Akit jobban érdekel a téma, az már egy nappal korábban, szombaton odaballaghat, mert két ingyenes előadás is lesz, az egyiket Vashegyi, a másikat Benoît Dratwicki (a Versailles-i Barokk Zenei Központ művészeti vezetője) tartja a Mondonville-ről és a francia operáról.

A szívemhez legközelebb álló előadás Sebestyén Mártáéké lesz. Ő Andrejszki Judittal, egy valódi provanszál trubadúrral, Miquèu Montanaróval, s annak a fiával Baltazárral lépnek a színpadra. Ahogy ígérik, az "autentikus francia barokk zene mellett a francia népi hangzásvilág és a magyar népzene találkozásának lehetünk fültanúi."

Ilyen az, amikor 20 ezer kacsa egyszerre kel át az úton

Az eset egyébként Kínában történt, és ránézésre óriási dugót okozott, de hát a kacsák szabályosan közlekedtek, így az autósoknak egy szavuk sem lehet.

Benjamin Franklin, az Egyesült Államok Függetlenségi nyilatkozatának egyik aláírója 1790 óta egy philadelphiai temetőben nyugszik, vagyis próbál nyugodni, de a temető látogatói folyton pénzérméket dobálnak a sírjára, évente több tízezer pennyt, mivel Franklin egyszer azt mondta: „Ha félretettél egy pennyt, kerestél egy pennyt" - írja az MTI.

Meg is lett a baj: az időjárás viszontagságainak kitett márványtábla elrepedt, a folytonos baszkurálástól egyre nagyobb lett a repedés, most meg nincs pénz helyrehozni, pedig a sírt kezelő alapítvány még támogatást is kapott rá. De még így is hiányzik tízezer dollár (nagyjából hárommillió forint), amit adománygyűjtéssel szeretnének összehozni. Pedig lehet, hogy csak meg kéne kérni a turistákat, mostantól inkább egydolláros bankókat hajigáljanak, az legalább nem töri össze a sírt, és hamar össze is jönne az összeg, szóval mindenki jól járna, és Franklin is foroghatna tovább a sírjában.

Fánkcsapdával és tökéletes öniróniával figyelmeztet az augustai rendőrség

Ugye ön is ismeri azt a sztereotípiát, hogy az összes amerikai rendőr dagadt és szeret fánkot zabálni? Nos, az augustai rendőrök nemcsak ismerik, de szívesen építenek is rá, ha a lakosságot kell figyelmeztetniük valamire, jelen esetben épp arra, hogy csapdák kihelyezéséhez engedélyre van szükség.

Mindezt egy zseniális Facebook bejegyzésben adták az emberek tudtára: egy rendőr egy doboz ingyen fánk mellett guggol, alá pedig azt írták, hogy bárki is teszi ki ezeket a fánkcsapdákat, most már igazán abbahagyhatná, mert Chase már harmadszor késik el miattuk a héten.

Fig. 2. Sample text for visualizing topics related to music

5 Conclusions

In this paper, we presented our methods to identify the most salient topics for a selected domain in Hungarian news articles, based on topic modeling. We also showed a solution for visualizing textual spans that are related to a given topic, focusing on the sports and music domains. In this way, it can be easily detected how many topics occur in a document, which topics are the most salient ones and which are more marginal with regard to the central topic of the document.

References

1. Berend, G., Farkas, R.: Keyphrase-Driven Document Visualization Tool. In: The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations (2013) 17-20
2. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, No. 5 (2003) 993-1022
3. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science. Vol. 41, No. 6 (1990) 391-407
4. Kim, S.N., Medelyan, O., Kan, M-Y., Baldwin. T.: SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In: *Proceedings of the 5th International Workshop on Semantic Evaluation* (SemEval '10). Association for Computational Linguistics, Stroudsburg, PA, USA (2010) 21-26
5. Lan, D., Buntine, W., Johnson, M: Topic Segmentation with a Structured Topic Model. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia Association for Computational Linguistics (2013) 190-200
6. Qazvinian, V., Radev, D. R., Mohammad, S. M., Dorr, B., Zajic, D., Whidby, M., Moon, T.: Generating extractive summaries of scientific paradigms. *J. Artif. Int. Res., Vol.* 46, No. 1 (2013) 165-201
7. Reynar, J. C.: Statistical models for topic segmentation. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (ACL '99). Association for Computational Linguistics, Stroudsburg, PA, USA (1999) 357-364
8. Riedl, M., Biemann, C.: Text Segmentation with Topic Models. JLCL, Vol. 27, No.1 (2012) 47-69
9. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP (2013) 763-771

A szó elszáll, az írás megmarad? Nyelvtechnológiai eszközök a déli manysi nyelvre

Szilágyi Norbert[1], Horváth Csilla[2], Vincze Veronika[3,4], Nagy Ágoston[2]

[1]Szegedi Tudományegyetem, Finnugor Tanszék
[2]Szegedi Tudományegyetem, Angol-Amerikai Intézet
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[2]Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék
e-mail:{norbertszilagyi91,naj.agi,nagy.agoston.1}@gmail.com
vinczev@inf.u-szeged.hu

Kivonat A cikkben a veszélyeztetett nyelvek digitalizációját célzó törekvésekkel összhangban készített manysi nyelvű, elsősorban a már kihalt déli nyelvjáráshoz készített nyelvtechnológiai eszközöket és forrásokat mutatjuk be. A manysi nyelv demográfiai helyzetének, valamint nyelvjárási felosztásának felvázolása után a 19. és 20. század fordulóján dolgozott két nyelvész kutatásainak eredményeként megjelent szövegekből készített, morfológiailag annotált déli manysi korpuszt ismertetjük. Kitérünk az Univerzális Dependencia projekt által kifejlesztett morfológiai és szintaktikai címkekészlet felhasználásával készített morfológiai és függőségi annotációra (mind az északi, mind a déli manysira), illetve az elkészült digitális déli manysi – magyar szótárra is. Elegendő mennyiségű tanításra használható annotált adat birtokában gépi tanulási módszerek alkalmazását tervezzük modern manysi szövegek automatizált szófaji egyértelműsítésére és függőségi elemzésre.
Kulcsszavak: manysi, uráli, morfológia, korpusz, szintaxis

1. Bevezetés

A manysi nyelvet (Nyugat-Szibériában beszélt uráli nyelv) a kutatók lényegében a szakirodalomban megjelenő első említései óta (pl. [1]) veszélyeztettnek tekintik. A századok során állapota csak fokozottan sérülékenyebbé vált, négy nyelvváltozatából kettő (a déli és a nyugati) kihalt, a harmadik (a keleti nyelvjárás-csoport) kihalásközeli állapotban van, a negyedik (az északi nyelvjárás-csoport) pedig veszélyeztetett. A 20. század utolsó évtizedeiben végbement gyors gazdasági és társadalmi változások eredményeként a még beszélt nyelvjárások esetében drasztikusan felgyorsult a nyelvcsere folyamata, a legutóbbi két összoroszországi népszámlálás adatai alapján a manysi beszélők száma mintegy a kétharmadával csökkent.

Bár a manysi nyelv és kultúra presztízse emelkedik, a manysi nyelv multietnikus, többnyelvű környezetében csak korlátozott szerepet tölt be, nagy hatással van használatára a hagyományos életmód eltűnése és az intenzív urbanizáció. A városi életmód ugyanakkor új nyelvhasználati színterek, és a nyelvelsajátítást,

kulturális örökség megőrzését segítő új eszközök létrehozását is lehetővé teszi. A kulturális örökség még elérhető elemeinek összegyűjtése, megőrzése iránti igény az 1980-as években, a manysi nyelvi revitalizációt és a városi manysi gyerekek örökségnyelvi oktatását célul kitűző kezdeményezések az 1990-es évek végén jelentek meg a városi manysi közösségben. Az internetkapcsolat elterjedésével a digitális környezet is olcsó és könnyen hozzáférhető potenciális nyelvhasználati színtérré vált, mely egyszerre lehet eszköze egy új beszélői közösség megteremtésének és színtere a megelőző és jelenlegi nyelvi dokumentációs projektek adatainak és eredményeinek. Az online adatbázisok és számítógépes nyelvi eszközök növelik egy nyelv beszélőközösségének nagyságát és a nyelv presztízsét, vagyis azt a két jellemzőjét, mely Kornai szerint az online nyelvi vitalitás legfontosabb fokmérője [2].

A természetesnyelv-feldolgozás területén számos projekt célozza veszélyeztetett és kihalt nyelvek dokumentációját. Az Univerzális Dependencia projekt[1] keretében függőségi fákat tartalmazó szintaktikai annotáció készült például ógörög, latin, szanszkrit és óegyházi szláv nyelvre. E törekvések nyomán projektünk a rendelkezésre álló manysi írott források digitalizációjára, ezen túl számítógépes nyelvi eszközök fejlesztésére törekszik, nem kizárólag a még beszélt, hanem a már kihalt dialektusok adatainak felhasználásával is. Manysi nyelvi korpuszt készítettünk terepmunka során gyűjtött és már rendelkezésre álló forrásokból felhasznált szóbeli és írott szövegek segítségével. A cikk beszámol az automatikus szövegfeldolgozást (pl. szófaji egyértelműsítést, szegmentálást) lehetővé tévő morfológiai és szintaktikai elemzésről, elsősorban a modern manysi szövegeket szem előtt tartva. Végül bemutatjuk a digitális szótárat, mely jelen pillanatban déli manysi – magyar nyelvpárra készült el.

2. Manysi nyelvjárások

A manysi (korábbi elnevezése: vogul) nyelvet Nyugat-Szibériában beszélik, a hanti nyelvvel együtt alkotja az uráli nyelvcsalád obi-ugor ágát. A 2010-es oroszországi népszámlálás eredményei alapján 12 262 ember vallotta magát manysinak, míg a nyelvet mindössze 938 ember beszéli. Összehasonlítva ezeket a számokat a 2002-es népszámlálás adataival – 11 432 manysi és 2746 beszélő [3] –, megállapítható a korábban is jellemző tendenciák folytatódása: míg a beszélők száma határozottan csökken, addig a népcsoport mérete lassan, de folyamatosan nő. A manysik többsége (10 977 fő) a Hanti-Manysi Autonóm Körzet – Jugra területén él, fennmaradó hányaduk nagyrészt Oroszország szomszédos közigazgatási egységeiben. A manysi nyelv nyelvjárásokra osztása a 19. század második felében, a kategorizációt lehetővé tévő nyelvi anyag összegyűjtések után vette kezdetét.

A nyugati és oroszországi irodalomban (pl. [4], [5], [6]) egyaránt leggyakrabban használt felosztás négy nyelvjáráscsoportot különböztet meg. Az északi nyelvjáráscsoportba tartozó nyelvváltozatokat a Szoszva és a Ljapin folyók mentén, illetve a Berjozovói körzetben beszélik. Ez a nyelvjáráscsoport (különösen

[1] http://universaldependencies.org/

a szoszvai nyelvváltozat) szolgált az irodalmi manysi nyelv alapjául. Az északi nyelvjárásokat erős orosz, valamint komi és nyenyec hatás jellemzi, emellett az északi hanti dialektussal is kapcsolatban álltak [7].

A nyugati nyelvjáráscsoport a valaha a Lozva folyó középső és alsó folyásánál beszélt nyelvváltozatokból állt, orosz és komi kölcsönhatás jellemezte. A keleti nyelvjárásokat a Konda és a Jukonda folyók mentén beszélték, tatár nyelvi hatások figyelhetők meg rajtuk. A déli nyelvjáráscsoportba tartozó nyelvváltozatokat a Tavda folyó mentén beszélték, itt figyelhető meg a legerősebb tatár nyelvi kölcsönhatás [7]. A déli nyelvjárások a 20. század első felében, a nyugati nyelvváltozatok a 20. század második felében, valószínűleg a '60-as, '70-es években haltak ki, a keleti dialektusnak a kortárs terepmunkaadatok alapján még él néhány beszélője.

3. Manysi nyelvű korpuszok

Statisztikai gépi tanulási módszerrel működő eszközök készítése során az annotált adatbázisok kiemelkedő fontossággal bírnak a természetes nyelvek feldolgozása területén, ezért áttekintjük a manysi nyelvre rendelkezésre álló korpusztípusokat.

Tudomásunk van egy, az egyetlen manysi sajtótermék szövegei felhasználásával épített északi manysi korpuszról [8]. A korpusz 520 000 tokent tartalmaz, XML formátumban. A korpuszt bemutató cikk a digitalizációt nehezítő tényezők között sorolja fel a jelentésmegkülönböztető szerepű hosszú magánhangzók jelölése okozta helyesírási és tipográfiai problémákat. Terepmunkánk során a mindennapi élet legkülönbözőbb területeihez tartozó témákban rögzítettünk északi manysi interjúkat, az adatbázis mintegy tíz órányi spontán beszédet tartalmaz. A felvételek lejegyzése jelenleg is tart, ezután a szövegek az északi manysi nyelvű korpuszt fogják bővíteni.

Kihalt dialektus lévén a déli manysi korpuszt nem kortárs szövegekből vagy interjúkból építettük, hanem két nyelvész, Munkácsi Bernát 1888-ban és 1889-ben, valamint Artturi Kannisto 1903-ban és 1904-ben, vagyis több mint száz évvel ezelőtt gyűjtött szövegmutatványait használtuk fel. Mindketten hosszú, átfogó, a nyelvi adatok mellett a mindennapi életre, folklórra, anyagi kultúrára, hiedelemvilágra is kiterjedő kutatómunkát folytattak, gyűjtésük megjelent anyagának a Vogul Népköltési Gyűjtemény [9] negyedik, illetve a Wogulische Volksdichtung [10] különböző köteteiben publikált déli manysi szövegei, főleg népmesék, énekek, találós kérdések szolgálnak a déli manysi korpusz alapjául. A latin betűs, diakritikus jelekkel gazdagon ellátott lejegyzésű szövegeket digitalizáltuk és elérhetővé tettük a korpusz honlapján[2]. Az adatbázis 2400 mondatból és 11 500 szóalakból áll, ebből 5000 külön lexikális egység. A korpusznak SIL FieldWorks Language Explorerrel (FLEx[3]) kézzel annotált, morfológiailag elemzett verziója is elkészült, az annotáció a magyar fordítással együtt elérhető a honlapon, vagyis az adatok párhuzamos korpuszként is használhatók.

[2] http://norbertszilagyi91.wixsite.com/tawdamansi
[3] http://fieldworks.sil.org/flex/

Mindegyik korpuszunk felhasználhatóságát bővítendő, az adatokat morfológiai és szintaktikai annotációval látjuk el. E célra az Univerzális Dependencia projekt által kifejlesztett morfológiai és szintaktikai címkekészletet [11] alkalmazzuk, különös figyelmet fordítva a rokon nyelvekre (pl. finnre és magyarra) kidolgozott nyelvspecifikus jegyekre. Az annotált adatok az Univerzális Dependencia projekt oldalán lesznek elérhetők. Alább az 1. és a 2. ábrán mutatunk példát egy annotált mondatra.

kōrəm	jū̄-pūńśít	pojít	
NUM	NOUN	ADP	
NumType=Card	Case=Nom		
	Number=Plural		
lo̯-to̯mo̯rkíť	pöŋəp	äsrä̆il	ńaulkä̆tål
NUM	ADJ	NOUN	VERB
NumType=Card	Case=Nom	Case=Nom	Definite=Indef|Mood=Ind
	Degree=Pos	Number=Sing	Number=Sing|Person=3|Tense=Pres

1. ábra: Egy déli manysi mondat morfológiai elemzése („A tizenkét fejű sárkány kergeti a három testvért.”).

2. ábra: Egy déli manysi mondat függőségi elemzése („A tizenkét fejű sárkány kergeti a három testvért.”).

4. Lexikális erőforrások

A 19. és 20. század fordulóján tevékenykedő kutatók manysi szóanyagából összeállított szótárak csak jelentős késéssel jelentek meg [9,10], a szótárak minden, így az azóta kihalt dialektusok szóanyagát is tartalmazzák. Ezen kívül rendelkezésre állnak csak az északi manysi nyelvváltozaton alapuló, némileg modernebb szótárak [12,13] is. Legjobb tudomásunk szerint kizárólag a déli manysi szóanyagot

tartalmazó szótár eddig nem készült, ezért a déli manysi korpusz adatai alapján déli manysi szószedetet hoztunk létre.

A déli manysi szótár a 3. pontban bemutatott korpuszban szereplő minden szót és toldalékot tartalmaz, annak magyar fordításával és morfológiai információkkal együtt. A lista az egyes elemek alapváltozatait, allomorfjait is tartalmazza. Az alábbiakban bemutatunk egy szótári alakot és egy toldalékot az adatbázisból.

äsrä̆il (n)
sárkány, ördög

-ɔ́lt (v>v)
CAUS (műveltető)
ɔlt

Elsőként a szótári alak szerepel, szófaji információkkal együtt ((n – főnév és v>v mint igéből igét képző toldalék), majd a magyar fordítása és a helyesírási változatokat korpuszadatok alapján. Összesen 1333 elemet tartalmaz a szótár, melyek szófaji eloszlása az 1. táblázatban látható.

Szófaj	**Elemszám**
Főnév	481
Ige	338
Határozószó	97
Melléknév	75
Névutó	55
Partikula	38
Toldalék	249

1. táblázat. Szófajok a déli manysi szótárban.

5. Nyelvtechnológiai eszközök

A manysi egy gazdag morfológiájú nyelv, tehát a hozzá készült NLP-eszközök jó morfológiai elemzőt igényelnek. Annak ellenére, hogy ez egy veszélyeztetetett nyelv, néhány morfológiai elemző készült már hozzá. Az északi manysihoz ugyan van morfológiai elemző [14,15], de latin karaktereket használ és csak a Kálmánféle Chrestomathia Vogulica [5] és Wogulische Texte [16] szövegeire, illetve Munkácsi szövegeire van optimalizálva. Ezen felül sajnos nem is nyílt forráskódú.

Az északi manysihoz Vincze és mtsai [17] már említenek egy morfológiai elemzőt, amely nem olyan rég óta érhető el a Giellatekno webhelyről[4]. Ez az elemző már a mai manysi szövegekre van optimalizálva, amelyek a cirill átírást

[4] http://giellatekno.uit.no/cgi/d-mns.eng.html

használják. Ez egy nyílt forráskódú eszköz, amelynek nyelvtani alapjait később ki szeretnénk terjeszteni a déli manysi latin betűs szövegeinek feldolgozásához. Ezen felül, mivel az északi és a déli dialektusok között vannak különbségek (pl. a déli dialektusban a zéró-morfémás jövő idő vs. jelölt jelen idő [18]), a morfológiai paradigmák átírása is szükséges, amelyre csapatunk vállalkozik. Ebben a munkafolyamatban felhasználjuk a korpusz kézzel annotált mondatait is, amelyben a szóalakok szótövekkel és affixumokkal szerepelnek, és ezt beépíthetjük az elemzőbe is.

Mint ahogy korábban említettük, jelenleg a korpusz morfológiai és szintaktikai annotációján dolgozunk. Amint lesz elegendő mennyiségű annotált adatunk, amelyet tanításra használhatunk, tervezzük gépi tanuló módszerek alkalmazását automatizált POS-taggelésre és függőségi elemzésre (parsing), elsődlegesen a modern manysi szövegek automatikus elemzését szem előtt tartva. Az így elkészülő nyílt forráskódú POS-taggerhez és elemzőhöz az elemző iránt érdeklődő személyek ingyenesen hozzáférhetnek.

6. Összefoglalás

A cikkben a veszélyeztetett nyelvek digitalizációját célzó törekvésekkel összhangban készített manysi nyelvű, elsősorban a már kihalt déli nyelvjáráshoz készített nyelvtechnológiai eszközöket és forrásokat mutattuk be. A manysi nyelv demográfiai helyzetének, valamint nyelvjárási felosztásának felvázolása után a 19. és 20. század fordulóján dolgozott két nyelvész kutatásainak eredményeként megjelent szövegekből készített, morfológiailag annotált déli manysi korpuszt ismertettünk, továbbá az északi manysira elkészült korpuszoktól is szót ejtettünk. A korpuszba emelt déli manysi szövegek eredetileg a kutatók gyűjteményes köteteiben jelentek meg évtizedekkel ezelőtt, így már a digitalizálásukkal is elérhetőek váltak úgy nyelvészettel, természetes nyelvek feldolgozásával foglalkozó tudományos körök, mint az őshonos népek kulturális öröksége iránt érdeklődő szélesebb közönség számára is. Ezen felül a szövegben található néprajzi adatok a manysik, vagy általában az oroszországi őshonos kisebbségi népek iránt érdeklődő etnográfusok, folkloristák érdeklődésére is számot tarthatnak.

Kitértünk az Univerzális Dependencia projekt által kifejlesztett morfológiai és szintaktikai címkekészlet felhasználásával készített morfológiai és szintaktikai annotációra, illetve az elkészült digitális déli manysi – magyar szótárra is. Elegendő mennyiségű tanításra használható annotált adat birtokában gépi tanulási módszerek alkalmazását tervezzük (elsősorban modern, északi nyelvjárásban íródott) manysi szövegek automatizált szófaji egyértelműsítésére és függőségi elemzésre.

Reméljük, hogy mind a korpuszunk, mind a készített eszközök segítséget nyújtanak a nyelvük története iránt érdeklődő manysiknak és a veszélyeztetett, kihalásközeli és kihalt nyelvek iránt érdeklődő NLP-kutatók számára egyaránt.

Köszönetnyilvánítás

A kutatás a Számítógépes eszközök a veszélyeztetett finnugor nyelvek nyelvi revitalizációjáért (FinUgRevita) nevű, FNN 107883 azonosítószámú projekt keretében valósult meg, az OTKA támogatásával.

Hivatkozások

1. Munkácsi, B.: Nyelvészeti tanúlmányútam a vogulok földjén [My Linguistic Fieldwork in the Land of Voguls]. Budapesti szemle **17**(60) (1889) 383–408
2. Kornai, A.: Digital language death. PLoS ONE **8**(10) (2013) e77056
3. Sipőcz, K.: www.perepis2002.ru. Finnugor Világ **10**(2) (2005) 23–27
4. Riese, T.: Vogul. Number 158 in Languages of the World/Materials. Lincom Europa, München - New Castle (2001)
5. Kálmán, B.: Chrestomathia Vogulica. Tankönyvkiadó, Budapest (1976)
6. Rombandeeva, E.I., Vakhruševa, M.P.: Mansijskij jazyk. Prosveščenije, Leningrad (1984)
7. Keresztes, L.: Mansi. In Abondolo, D., ed.: The Uralic languages, Routledge (1998) 387–427
8. Horváth, C., Szilágyi, N., Nagy, A., Vincze, V.: Language technology resources and tools for Mansi: an overview. In: Proceedings of the Third International Workshop on Computational Linguistics for Uralic Languages, St. Petersburg, Russia (2017)
9. Munkácsi, B., Kálmán, B.: Wogulisches Wörterbuch. Akadémiai Kiadó, Budapest (1986)
10. Kannisto, A.: Wogulisches Wörterbuch. Kotimaisten Kielten Keskuksen Julkaisuja, Helsinki (2013)
11. Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D.: Universal Dependencies v1: A Multilingual Treebank Collection. In: Proceedings of LREC 2016. (2016)
12. Rombandeeva, E.I.: Russko-mansijskij slovar'. Mirall, Sankt-Peterburg (2005)
13. Rombandeeva, E.I., Kuzakova, E.A.: Slovar' mansijsko-russkij i russko-mansijskij. Prosvešenije, Leningrad (1982)
14. Prószéky, G.: Endangered Uralic Languages and Language Technologies. In: Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria (2011) 1–2
15. Fejes, L., Novák, A.: Obi-ugor morfológiai elemzők és korpuszok. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010), Szegedi Tudományegyetem (2010) 284–291
16. Kálmán, B.: Wogulische Texte mit einem Glossar. Akadémiai Kiadó, Budapest (1976)
17. Vincze, V., Nagy, A., Horváth, C., Szilágyi, N., Kozmács, I., Bogár, E., Fenyvesi, A.: FinUgRevita: Developing Language Technology Tools for Udmurt and Mansi. In: Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages, Tromsø, Norway (2015)
18. Honti, L.: System der paradigmatischen Suffixmorphemen des wogulischen Dialektes an der Tawda. Akadémiai Kiadó, Budapest (1975)

VII. Laptopos bemutatók

Shtylo: stilometriai elemzések webes támogatása

Dobi Jan Sándor[1], Mészáros Tamás[1], Kiss Margit[2]

[1] BME Méréstechnika és Információs Rendszerek Tanszék
H-1117 Budapest, XI. Magyar tudósok körútja 2. I ép. E437.
meszaros@mit.bme.hu
[2] MTA BTK Irodalomtudományi Intézet
1118 Budapest, Ménesi út 11–13.
kiss.margit@btk.mta.hu

Kivonat: A stilometria a számítógépes nyelvészet dinamikusan fejlődő területe. Széles körű felhasználását azonban gátolja az a tény, hogy alkalmazóinak többsége nincs a szükséges informatikai tudás birtokában. Cikkünkben egy olyan rendszert mutatunk be, amely az R nyelven írt stylo programcsomaghoz nyújt egy teljes értékű webes felhasználói felületet, valamint segítséget nyújt a stilometriai kísérletekhez szükséges korpuszok összeállításában és tárolásában is. Az elkészített szoftver működését egyrészt történeti szövegek elemzésén, másrészt plágiumkeresési feladat végrehajtásán mutatjuk be.

1 Bevezetés

A digitalizált szövegkorpuszok létrehozása a hagyományos módszerektől eltérő, új elemzési eljárások kialakítását teszi lehetővé, illetve kívánja meg. Az egyik, napjainkban egyre inkább elterjedő ilyen korpusznyelvészeti módszer a stilometria, amelynek egyik széles körben használt eszköze az R nyelven írt Stylo programcsomag [7, 9]. Az ilyen, alapvetően statisztikai vizsgálatokhoz az R nyelv alkalmazása kézenfekvő megoldás, ám használata olyan speciális szakértelmet kíván, amellyel éppen azok a kutatók kevéssé rendelkeznek, akik a számítógéppel támogatott nyelvi elemzésektől komoly eredményeket várnak. Célunk egy olyan támogató eszköz létrehozása volt, amellyel a felhasználók az R nyelv mélyreható ismerete nélkül is elvégezhetik a manapság leggyakrabban alkalmazott stilometriai elemzéseket. Ehhez kifejlesztettünk egy webalkalmazást, amely egyrészt egyszerűen használható felülettel látja el a Stylo programcsomagot, másrészt kiegészíti korpuszkezelési funkciókkal. Tanulmányunkban az eszköz alkalmazási lehetőségeit két, merőben eltérő adatszerkezetű korpuszon mutatjuk be: egyrészt 18. századi történeti szövegek téma- és stíluskategorizálására alkalmaztuk, másrészt mesterséges nyelvű szöveges dokumentumokon végzett plágiumkeresésre.

1.1 A számítógépes stilometria és problémái

A stilometria, a stílus elemzésének számszerűsített változata, a számítógépes nyelvészet egyik sokat kutatott és széles körben alkalmazott területe. Módszerei alapvetően a statisztikai szövegelemzés területéről kerülnek ki, tipikusan olyan jellemzőket és azokra épülő klaszterezési eljárásokat alkalmaz, amelyek a szövegeket hasonlóságát különféle jellemzőik és azokon értelmezett távolságmértékek szerint ítélik meg. A mértéket úgy szabja meg, hogy a bemeneti szövegeknek a feladat szempontjából releváns jellemzőit ragadja meg.

Egy stilometriai vizsgálat több lépésből áll: a korpusz összeállítása, a kívánt jellemzők, valamint a meghatározásukhoz legmegfelelőbb módszerek kiválasztása, az előfeldolgozás, a jellemzőkinyerés és -feldolgozás, majd az eredmény előállítása. Bár a lépések többsége jól támogatható informatikai eszközökkel, és ily módon lehetőség nyílik nagy méretű korpuszok egyszerű és gyors elemzésére, a gépi stilometriai elemzésnek megvannak a maga akadályai, amelyek megnehezítik az effajta törekvéseket.

Ezek az alábbi csoportokba sorolhatók:

- **Technológiai nehézségek:** Noha a számítógépek kapacitása rohamosan nő, de ugyanez elmondható az elemző módszerek komplexitásáról és a korpuszok méretéről is. A számítógéppel szemben támasztott igényeink általában felemésztik a rendelkezésre álló kapacitást. Az egyre növekvő méretű adathalmazokon futtatott elemzőmódszerek jelentős számítási igénnyel rendelkeznek, ezért van egy felső korlátja annak, hogy egy átlagos kutató mekkora szövegkorpuszokon dolgozhat értelmes időkeretek betartásával.
- **Tudás hiánya:** A korpuszmodell megalkotása és a kívánatos jellemzők meghatározása igen tudásintenzív feladat. A kutatónak tudnia kell, hogy milyen jellegeket kíván kinyerni a szövegekből. Erre vannak ajánlások, de minden kísérlet egyedi. Mi több, manapság elemzőalgoritmusok hada áll rendelkezésre: melyiket válasszuk egy adott szövegkorpusz esetén? Hagyományos statisztikai vagy gépi tanulás alapú módszert válasszunk? Ezekre a kérdésekre nincsen egyértelmű és könnyű válasz.
- **Megfelelő módszerek hiánya:** Noha számos publikált és implementált módszer áll egy kutató rendelkezésére, de a gépi stilometria mégiscsak egy fiatal terület. Egyelőre nem alakultak ki általánosan optimális megoldások, bizonyítottan „jó” eljárások a feladatok megoldására. Néha már a probléma formalizálása is akadályokba ütközik. Ide sorolhatók a nyelvfeldolgozással kapcsolatos problémák is. Egy korpusz részletes nyelvi elemzése minden bizonnyal sokkal gazdagabb eredményeket szolgáltatna, de a természetes nyelvű (különösen a történeti) szövegek informált nyelvi elemzését számos tényező gátolja (nyelvtani változatosság, elírások, nyelvi hibák). Ehelyett a statisztikai és a gépi tanulásos black-box módszerek alkalmazhatók egy-egy kutatási kérdés megválaszolásához.

1.4 Tipikus feladatok

A stilometrián belül vannak tipikusnak mondható kérdések, amelyeket megfogalmaznak egy szövegkorpusz elemzése kapcsán, és alapvetően a szerző stílusára, írásmódjára vonatkoznak. Az alkalmazások alapvetően a szöveg „megmért" stílusjegyeinek felhasználásban különböznek. Az alábbiakban ezekre mutatunk be néhány példát.

Törvényszéki stilometria. A stilometria egy jelentős alkalmazási területe a jogászi munka módszertanába is bekerült, jelentős mértékben a bűnüldözés területén van jelen [2]. Amikor nem állnak rendelkezésre fizikai szövegek, csak digitálisak (ami egyre gyakrabban előfordul) a szöveg íróját nem lehet többé az íráskép alapján beazonosítani. A digitális szövegeknek nincsen írásképük, így a szöveg más tulajdonságai alapján kell beazonosítani, hogy ki írta őket. Tovább nehezíti a feladatot az, hogy esetenként számolni kell a megtévesztés lehetőségével [3]. Ide sorolható a plágiumkeresés is, ugyanis ilyenkor is egy megtévesztési kísérletet kell kiszűrnünk. A mintapéldákról szóló rész bemutat egy, a rendszerünkkel végzett plágiumkeresési tesztet is.

Történelmi szövegek szerzőségi, autentikusságának vizsgálata. A régebbi korok szövegei kapcsán különböző okokra visszavezetve gyakran merülhet fel a valós szerzőség azonosításának a kérdése. Az ilyen kérdések megválaszolása a nyelv- és irodalomtudomány egyik fontos kutatási területe, amely a digitális módszertannak köszönhetően a korábbiaktól hatékonyabb módon képes megválaszolni a tisztázatlan kérdéseket. A stilometriának ez az egyik legtermékenyebb területe [18], amely a modern a korban jelentős eredményeket hozott kezdetben különösen a Shakespeare-kutatásban [4, *19]*; de azóta számtalan más szerzővel kapcsolatban is, ami következésképpen a módszertan folyamatos változását eredményezi [20]. A cikkünkben egy magyar 18. századi szövegeken végzett kísérletet mutatunk be.

Időbeliséganalízis. A harmadik jelentős terület, ahol eredményeket értek el stilometriai eszközökkel: a bizonytalan korú szövegkorpuszok időbeliségének vizsgálata. Lorenzo Valla kísérlete részben szintén ide tartozik, aki elemzését a szöveg latin stílusának korára alapozta. Wincenty Lutosławski elemzése is egy effajta kísérlet: Platón párbeszédeit rendezte időrendi sorba stilisztikai jellemzőik alapján [6]. A platóni dialógusok kérdése azóta számos alkalommal megragadta egy-egy kutató figyelmét, hisz az időbeliség kulcsfontosságú Platón filozófiájának megértéséhez. Napjainkban a szerzőségi vizsgálatokat alkalmazzák nyelvtörténeti periodizációban, de szerzői életmű korszakolásában is [21]. Esettanulmányként mi is végeztünk egy kísérletet, amelyben Mikes Kelemen műveinek kronológiai vizsgálata volt a cél.

Csoportbéli hovatartozás vizsgálata. Egy adott szöveg esetén nem csak a konkrét szerzőre lehetünk kíváncsiak. Számos esetben elegendő, ha el tudjuk dönteni, hogy valamely csoporthoz való tartozása fennáll-e [22]. Bizonyos elemzésekben már ez is értékes információkkal szolgálhat. Kérdés lehet, hogy a szerző férfi-e vagy nő, milyen nemzetiségű, idősebb vagy fiatalabb, milyen iskolázottságú, milyen személyiségtípusba tartozik. Az efféle elemzések egy speciális esete a szerző anyanyelvének meghatározása az idegen nyelvű írása alapján.

1.5 Stilometriai eszközök

A stilometria egyre népszerűbb tudományterület, számos módszerének létezik szoftver-megvalósítása.

A Signature [7] ingyenes szoftver, melyet Peter Millican fejleszt az Oxfordi Egyetemen. A weboldal szerint főleg szerzőségi vizsgálatra alkalmas. Számos esettanulmány található a weboldalon, melyet azonban úgy tűnik nem frissítettek egy ideje. A Java Graphical Authorship Attribution Program, vagy JGAAP [8] egy nyílt forrású, szabad szoftver, melyet a Githubon lehet elérni. Fő fejlesztője Patrick Juola. Például a Robert Galbraith álnéven publikált Kakukkszó című regény valódi szerzőjének (J. K. Rowling) azonosításához a JGAAP szoftvert alkalmazták, az elemzésben Juolanak jelentős szerepe volt. A Stylene az Antwerpeni Egyetem CLiPS kutatócsoportja által működtet stilometriai webszolgáltatás holland nyelvre.

Az általunk is használt stylo, mely a Shtylo alapját képezi, egy R-csomag, melyet Maciej Eder, Jan Rybicki és Mike Kestemont fejlesztett, szabad nyílt forrású szoftver. A következő rész ezt a csomagot mutatja be.

2 A stylo csomag

A csomagot Maciej Eder, Jan Rybicki és Mike Kestemont írta [9]. A csomag szabad szoftver, letölthető a Githubról [10]. 2013-ban készült, megjelenése előtt [11] a teljes stilometriai workflow – feltehetőleg a JGAAP kivételével – nem állt rendelkezésre egy szoftver részeként. A stylo magába foglal számtalan releváns elemzési módszert, továbbá megkönnyíti a teljes stilometriai folyamatot. A csomagba foglalt magas szintű csomagolófüggvények a workflow majdnem minden lépését elvégzik helyettünk, csak a konfigurációjukat kell megadnunk.

2.1 A Stylo főbb funkciói

A csomag öt fő magas szintű csomagolófüggvénnyel rendelkezik, melyek magukba foglalják az egyes lépéseket és egy egész elemzést végigvisznek az elejétől a végéig. Hívható grafikus felülettel, ekkor a paraméterek grafikus vezérlőelemek segítségével állíthatók be, vagy grafikus felület nélkül, ekkor a paraméterek R kódból állíthatók. A felhasználói felületre látható példa az 1. ábrán, amelyen a stylo függvény paraméterek nélküli meghívásának eredményét látjuk. A többi csomagolófüggvény felülete is nagyon hasonló.

1. ábra. *A stylo csomag stylo() függvényének grafikus felülete.*

Az öt függvény:

- **stylo**: A függvény feladata, hogy a stilometria fő irányzatának számító modellalapú, statisztikai módszereken alapuló elemzést hajtsa végre. Számos elemzési eljárás közül választhatunk, mint a klaszteranalízis, a Multidimensional Scaling (MDS), a főkomponens-analízis kovarianciamátrixszal (PCV), a főkomponens-analízis korrelációs mátrixszal (PCR) és a konszenzusfa (BCT). A módszerekben különféle távolságmértékeket alkalmazhatunk (euklideszi távolság, Manhattan-távolság, Burrows-delta [12], Argamon-delta [13], Eder-delta, egyszerű Eder-delta, Canberra-távolság, koszinusztávolság).
- **classify**: A függvény gépitanulás-alapú osztályozó algoritmusok futtatására alkalmas. Az eljárás kétlépcsős: először betanítunk egy osztályozót egy megfelelően strukturált tanítóminta-készlettel, majd a számunkra érdekes szövegeket megkíséreljük az osztályozóval besorolni. A függvény számos eljárást ismer (Burrows távolságmérték, k-Nearest Neighbours, Szupportvektor-gépek, Naiv Bayes, Nearest Shrunken Centroid).
- **rolling.delta, rolling.classify**: A már említett elemzések görgetett változata: céljuk, hogy egy korpusz szövegein belül stilisztikai váltások felismerését tegyék lehetővé. A szövegeket adott méretű átlapolható ablakokra osztjuk, és minden ablakon egy elemzést hajtunk végre. Az így kapott értékeket egy grafikonon ábrázolva hirtelen változásokat lehet felfedezni, melyek számos dolgot jelenthetnek, pl. egy szerző átvette a művet egy másiktól.
- **oppose**: A függvény két szövegkorpusz összehasonlítására való. A függvény eredményeként két szólistát kapunk: az első a tesztelt szerző által kifejezetten

preferált szavakat tartalmazza, míg a második ezzel szemben a szerző által kifejezetten került szavakat.

3 A Shtylo webalkalmazás

A stylo potenciális alkalmazóinak jelentős részétől nem lehet elvárni, hogy ismerje az R programozási nyelvet, és tudjon használni egy R-csomagot. Ez más jellegű szakértelmet igényel. Manapság webrendszerekkel mindenki kapcsolatban áll, így egy megfelelő webes felhasználói felület alkalmazásával kezelhetjük ezt a problémát. Noha a stylo csomagnak létezik egy grafikus és egy kezdetleges webes felülete is (a websty), ezek nehezen használhatók. Számos érv szól amellett, hogy a stylohoz készítsünk egy színvonalas webalkalmazást.

A felhasználói felület azonban nem az egyetlen megoldandó probléma. Érdemes még egy terhet levenni a kutató válláról: a futtatókörnyezet kialakítását (szoftverek telepítését és beállítását). Egy webalkalmazás esetén csupán egy böngészőre van szükség, az R program egy szervergépen fut, ott egyszer kell a szoftvert telepíteni és beállítani.

Egy harmadik szempont is említést érdemel. A stylo elemzései igen számításigényesek is tudnak lenni: nagy korpuszokon sok memóriát és processzoridőt tudnak fogyasztani. Ez jelentősen szűkítheti azon eszközök halmazát, ahol az eszköz használható. Ha egy központi, erős szervergépen fut az alkalmazás, akkor a szolgáltatásokat egy gyenge kliensről is el lehet érni, bárhova magunkkal lehet vinni, és igény szerint lehet elemzéseket futtatni. Ez a megoldás jól illeszkedik az informatikai szolgáltatások felhőalapú (azaz jól skálázható) megvalósításához is.

Az általunk készített Shtylo rendszer az eddigiek mellett egy további jelentős hozzáadott értékkel is bír, ez pedig a korpusz adatbázis-alapú tárolása. A stylo csomag a korpuszt, az elemzési eredményeket mind fájlok alakjában tárolja. Ez rosszul skálázódik, és nem nagyon alkalmas arra, hogy a segítségével korpuszokat építsünk és tartsunk nyilván. A Shtylo ezzel szemben a korpuszépítésben is segítséget nyújt. A felületen szövegeket lehet adott korpuszokban feltölteni, és a korpuszokat egy adatbázisban tárolja el. Ez jól skálázódik, tehát jelentős mennyiségű szöveg és sok felhasználó esetén a replikáció és terheléselosztás is megoldható. Ugyanez a fájlrendszerben tárolt fájlokkal jóval nehezebb lenne.

3.1 A Shtylo felépítése és technológiái

A Shtylo alapvetően egy kliens-szerver alkalmazás, amely kliens oldalon standard webböngészőt alkalmaz, szerver oldalon pedig az R környezet mellett egy webkiszolgálóra és egy korpusztárra (adatbázisra) épül.

2. ábra. *A Shtylo főbb részei*

Technológiai oldalról nézve a Shtylo a webes felület összeállítására a Shiny [14] keretrendszert használja, mely egy R-nyelvű webalkalmazások fejlesztésére alkalmas reaktív szoftverkönyvtár. Innen ered alkalmazásunk neve is: (Sh)iny + s(tylo).

Adatbázis-technológiának a MongoDB [15] dokumentumalapú adatbáziskezelőt választottuk, ugyanis adattárolási modellje jól illeszkedik a feladathoz és minden teljesítménybeli vagy funkcionális igényt kielégít.

3.2 A korpuszokról

A Shtylo rendszerben az adatréteg egy MongoDB adatbázis, amelyben a szövegek ún. gyűjteményekbe vannak szervezve és több gyűjtemény alkot egy adatbázist. Amikor egy szöveget feltöltünk, az a szöveg mindig egy MongoDB gyűjtemény, vagyis a mi értelmezésünkben egy korpusz tagja lesz. Amikor a Shtylo meghívja a stylo csomagot, akkor mindig az aktuálisan csatlakoztatott MongoDB gyűjteményt adja át neki elemzésre, azaz egy korpuszt. A gyűjtemények, vagyis korpuszok mindig egy adatbázisnak a tagjai, amelyeket felhasználókhoz, illetve projektekhez rendelhetünk.

3. ábra. *A Shtylo adatbázis-felülete. Balra az adatbázis és a gyűjtemény kiválasztása, jobbra lent a feltöltésre szánt fájlok megadása, felette az adatbázisnapló látható.*

3.3 A rendszer a felhasználó szemével

Felhasználóként a rendszer igen egyszerű. Az alábbi cselekvésekkel lépünk kapcsolatba a rendszerrel:

1. Kiválasztjuk, mely adatbázisnak mely gyűjteményén kívánunk dolgozni.
2. Ha ez még üres, akkor feltölthetjük a szövegeinket.
3. Elemzéseket futtatunk: kiválasztjuk, hogy jelenleg melyikre vagyunk kíváncsiak, és beállítjuk a paramétereit. A paraméterek a stylo beállítási lehetőségeitől függnek.
4. Az eredményeket a felületen látjuk, az adatokat, képeket változatos formátumokban elmenthetjük a lokális gépünkre.

4. ábra. *A csatlakozási űrlap és a napló egy sikeres csatlakozás esetén.*

5. ábra. *A stylo() függvény felülete a Shtylo webalkalmazásban egy elemzés lefuttatása után.*

A rendszer jelen pillanatban még csak a stylo() függvény számára biztosít felületet, mivel azonban ez a leggyakrabban használt funkció, ezért már így is hasznos lehet a stylo felhasználói számára. A továbbiakban három elemzési példát mutatunk be, melyek elvégzése során a Shtylo segített a stylo gördülékenyebb használatában.

4 Példák

A három példa közül kettő a Mikes-életművel foglalkozik. A harmadik egy kísérlet arra, hogy a stylo csomagot plágiumkeresésre használjuk egy egyetemi tárgy házi feladatára beküldött megoldásokon keresztül.

A Mikes-műveket a stilumetriai elemzés előtt egy szótárra épülő módszerrel normalizáltuk annak érdekében, hogy a szavak jellegzetesen sokféle írásmódja minél kevésbé torzítsa az eredményeket. A normalizálás részleteiről a [23] irodalomban számoltunk be.

4.1 Első kísérlet: saját művek és fordítások

Ebben a kísérletben arra a kérdésre kerestük a választ, hogy eldöntsük a stylo segítségével, a mikesi életműben hogyan viszonyulnak a saját művek a szerző által készített fordításokhoz. A kísérlet megtervezésekor az alábbi dolgokat tartottuk szem előtt:

- Mikes Kelemen saját művei egy fájlban voltak megadva, míg a fordítások külön-külön. Így az egyik osztályban több volt a mű, mint a másikban. Ilyenkor a Klasszikus Delta torzíthatja az eredményeket, így Eder deltájára került a választás, mely a nem izoláló nyelveken jobban teljesít.

- A konszenzusfa építése során magasra állítottuk a konszenzusküszöböt, mivel célszerű maximalizálni azt a hasonlóságot, ami mellett egy fordítás összekötésre kerül az eredeti művekkel.
- A műhosszak jelentős szórása miatt a mintavételezés mellett döntöttünk, és a szövegeket 10000-jelleg—hosszú darabokra osztottuk fel, ezzel csökkentve a statisztikai torzulást.

6. ábra. *Az első kísérlet konszenzusfája.*

Ennek az elemzésnek az eredményeképp a kísérlet során a művek jól láthatóan négy csoportba sorolódtak, amely árnyalja az életmű eddigi felosztásáról alkotott képet is. A beállítások finomításakor azt tapasztaltuk, hogy Az idő jól eltöltésének módja című fordítása mindig nagyon közel esik a Törökországi levelekhez (saját szerzőségű mű), bármilyen beállítást is használunk. Ezen kívül látni, hogy a többi fordítást külön konszenzuságakban találjuk meg, tehát csak ez az egyetlen fordítás volt olyan közel az eredetihez, hogy a futások 90%-ában volt közöttük közvetlen kapcsolat. Meglepetés az is, hogy az eddigi hagyományos kutatásokban a Törökországi levelekhez sok szempontból kötődő Mulatságos napok című fordítás nem kapcsolódik olyan szorosan a levelekhez.

4.2 Második kísérlet: művek tematikája

Ebben a kísérletben az volt a cél, hogy tematikai vizsgálatot végezzünk az életmű egészében. Arra voltunk kíváncsiak, hogy a stylo képes-e a különféle tematikájú szövegeket, így az egyházi, erkölcsi és élőbeszéd kategóriájába sorolható műveket különválasztani, illetve hogy ezek a művek stilisztikai szempontból megkülönböztethetőek-e. Három élőbeszédes, egy erkölcsi, és tíz vallási témájú szöveg alkotta a korpuszt. Ennél a kísérletnél:

- A Canberra távolság mellett döntöttünk, mert az érzékeny a ritka szókincsbeli különbségekre.
- A művek hosszeltérését itt is mintavételezéssel ellensúlyoztuk.
- A konszenzusküszöböt itt alacsony, mert itt az a kérdés, hogy a művek kapcsolatban lesznek-e egymással Canberra távolság mellett, vagy távol maradnak.

7. ábra. *A második kísérlet konszenzusfája.*

A kísérlet eredményeképpen két fő dolog látszódik világosan kirajzolódni. Először is az élőbeszéd jellegű szövegeket jól láthatóan külön lehet választani a többitől. Másodszor pedig az erkölcsi témájú szöveg besorolódik egy fába a vallási témájú szövegek egy részével. Tehát itt a stylo főleg arra volt képes, hogy az élőbeszédes szövegeket és a nem élőbeszédes szövegeket különválassza – amely a stilisztikai, szövegnyelvészeti kutatások számára nagy előnnyel bír –, s ez a legerősebb stilisztikai marker ebben a korpuszban.

4.3 Harmadik kísérlet: plágiumkeresés

Ez a kísérlet eltér az eddigiektől annyiban, hogy a stylo efféle használata nem dokumentált. Érdekes kérdés, hogy használható-e plágiumkeresésre is, ráadásul nem természetes nyelvű szövegeken, hanem forráskódokon, jelen esetben Java nyelven írt házi feladatokon. Az eredmények meglepően jók lettek, a referenciaként használt Sherlock [16] is azt a kettő párt adta valószínű plágiumként, mint a stylo. Emögött az ok valószínűleg az, hogy a plágiumok nem voltak túl szofisztikáltak. Az egyik eleve teljesen azonos másolat volt, ez nem tekinthető sikernek, a másik pedig relatíve egyszerűen azonosítható gépies módosításokat eszközölt a forráskódon. Ennek ellenére is biztató eredmény, hogy a stylo eredményesen működött.

8. ábra. *A harmadik kísérlet konszenzusfája.*

5 Összefoglalás

Célkitűzésünk egy olyan rendszer megvalósítása volt, amely az informatikában nem jártas kutatók számára is elérhetővé teszi az R nyelven írt stylo programcsomag alkalmazását. Használatához nincs szükség egyéni szoftvertelepítésre és -konfigurációra, mivel a program központi szerveren és ahhoz kapcsolódó böngésző

alkalmazás segítségével működik. A megvalósított programunk azonban többet nyújt egy egyszerű webes felhasználói felületnél.

Egyrészt olyan elemzési módszereket állítottunk össze, amelyek a stylo legfontosabb eszközeihez nyújtanak egyszerűen használható felületet azok bonyolult numerikus paraméterezését mellőzve. Ezek lehetővé teszik a szövegek stilometriai elemzését az R programozási nyelvet és a stylo programcsomagot nem ismerő felhasználók számára is.

Másrészt a rendszert kiegészítettük egy MongoDB-re épülő korpusztárral, amely a stylo fájlalapú korpuszbeviteli megoldásához képes lényegesen rugalmasabb tárolási és menedzselési lehetőségeket kínál. A felhasználók a webes felületen keresztül állíthatják össze a korpuszt, és adhatják meg annak kiválasztott részeit az elemzések számára.

A rendszer működését több kísérletben ellenőriztük történeti szövegek klaszterezése és programszövegek plágiumkeresése céljából.

A megvalósított rendszer nyílt forráskódú (https://github.com/dobijan/shtylo), a DHmine keretrendszer részét képezi (http://dhmine.mit.bme.hu). Közzétételével szeretnénk az egyre nagyobb népszerűségnek örvendő stylo eszköztárra felhívni a hazai kutatók figyelmét. Reményeink szerint az egyszerűen használható webes felülettel hozzájárulunk a stilometriai módszerek hazai elterjedéséhez.

Köszönetnyilvánítás

A projekt az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap társfinanszírozásával valósul meg (EFOP-3.6.2-16-2017-00013).

Bibliográfia

1. Lutosławski, W.: Principes de stylométrie. (1898)
2. Totty, R.N., Hardcastle, R.A., Pearson, J.: Forensic linguistics: The determination of authorship from habits of style. Journal of the Forensic Science Society, 27(1):13–28. (1987)
3. Brennan, M., Afroz, S., Greenstadt, R.: Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. ACM Trans. Inf. Syst. Secur., 15(3):12:1–12:22. (2012.)
4. Elliott, W.E.Y., Valenza, R.J.: And then there were none: Winnowing the Shakespeare claimants. Computers and the Humanities, 30(3):191–245. (1996)
5. Mosteller, F., Wallace, D.: Inference and Disputed Authorship: The Federalist. Addison-Wesley. (1964)
6. Lutosławski, W.: Sur une nouvelle méthode pour déterminer la chronologie des dialogues de Platon. Paris: H. Welter. (1896)
7. Millican, P.: The Signature Stylometric System http://www.philocomp.net/humanities/signature.htm (2017. november 24.)
8. Juola, P.: Java Graphical Authorship Attribution Program. https://github.com/evllabs/JGAAP (2017. november 24.)

9. Eder, M., Rybicki, J., Kestemont, M.: stylo: Functions for a Variety of Stylometric Analyses, https://cran.r-project.org/package=stylo. (2017. november 24.)
10. Eder, M., Rybicki, J., Kestemont, M.: computationalstylistics/stylo: R package for stylometric analyses https://github.com/computationalstylistics/stylo (2017. november 24.)
11. Eder, M., Rybicki, J., Kestemont, M.: Stylometry with R: a suite of tools. In Digital Humanities 2013: Conference Abstracts, pages 487–489. (2013)
12. Burrows, J.: 'Delta': A measure of stylistic difference and a guide to likely authorship. (2002)
13. Argamon, S.: Interpreting Burrows's delta: geometric and probabilistic foundations. Literary and linguistic computing, 17(3):267–287. (2002)
14. Rstudio: Shiny by RStudio http://shiny.rstudio.com/ (2017. november 24.)
15. MongoDB, Inc.: Introduction to MongoDB https://docs.mongodb.com/manual/introduction/ (2017. november 24.).
16. Department of Computer Science, University of Warwick: Sherlock – Plagiarism Detection Software https://www2.warwick.ac.uk/fac/sci/dcs/research/ias/software/sherlock/ (2017. november 24.)
17. Dobi, J. S.: Shtylo Wiki. https://github.com/dobijan/shtylo/wiki (2017. november 24.)
18. Jack Grieve, Quantitative Authorship Attribution: An Evaluation of Techniques, Literary and Linguistic Computing, Vol. 22, No. 3, 2007. 251–270.
19. *Craig DH, Kinney AF, Shakespeare, Computers and the Mystery of Authorship, Cambridge University Press, Cambridge, UK, 234 (2009)*
20. Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., Daelemans, W. (2016). Authorship Verification with the Ruzicka Metric. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 246-249.
21. van Hulle, D., Kestemont, M. Stylochronometry and the Periodization of Samuel Beckett's Prose. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 393-395. (2016)
22. Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? Literary and Linguistic Computing 1-7, (2011)
23. Margit Kiss, Tamás Mészáros, Creating an extended author's dictionary to support digital literary research In: DH Benelux 2016. Luxembourg (2016)

Természetes nyelvi interfész menetrend- és utazástervező szolgáltatásokhoz

Kemény Boldizsár, Recski Gábor

Budapesti Műszaki és Gazdaságtudományi Egyetem
Automatizálási és Alkalmazott Informatikai Tanszék

Kivonat A `Tervezz Velem` (`TeVe`) egy a Facebook Messenger platformon nyilvánosan elérhető chatbot-alkalmazás, mely természetes nyelven írt kérdésekre reagálva a budapesti tömegközlekedéssel kapcsolatos információkat jelenít meg. Az alkalmazás elemző modulja lehetővé teszi, hogy néhány szóból álló, töredékes bemenetekben is nagy pontossággal fedezzük fel a felhasználói szándékot. Az `e-magyar` nyelvfeldolgozó pipeline intergrációjának köszönhetően a rendszer képes a bemenetben szereplő főnévi esetragokat is figyelembevenni az elemzés során, a különböző szegmentációkat rangsoroló logika pedig lehetővé teszi, hogy a gyakran előforduló, többértelmű bemenetekhez a legmegfelelőbb elemzést társítsuk. A rendszernek jelenleg 160-nál több aktív felhasználója van, nyilvánosan elérhető a https://m.me/tevepp oldalon, a teljes forráskód pedig hozzáférhető a https://github.com/kebou/tevepp oldalon.

1. Bevezetés

Ebben a cikkben a `Tervezz Velem` (`TeVe`) nevű chatbot alkalmazást mutatjuk be, melynek segítségével a felhasználó természetes nyelven kommunikálva juthat információhoz a budapesti tömegközlekedésről. A rendszer a Facebook Messenger platformon nyilvánosan elérhető, és képes a chat-programokra jellemző rövid, töredékes felhasználói kérdések értelmezésére. A bevezetést követően először bemutatjuk a `TeVe` funkcionalitását és általános felépítését (2. fejezet), ezután vesszük sorra a fontosabb modulokat. A 3. fejezet a szövegfeldolgozó komponenst mutatja be, a 4. részben pedig az egyes felhasználói kéréseket értelmező és a válaszhoz szükséges információkat lekérdező komponens részleteit ismertetjük. Mivel a beérkező kéréseknek gyakran többféle elemzése is generálható, ezért a rendszer további fontos funkciója ezek rangsorolása, melyről az 5. fejezetben írunk részletesebben. Végül a 6. fejezetben bemutatjuk, miként integráltuk az alkalmazást a Facebook Messenger platformra és hogy eddig milyen tapasztalatokkal járt a nyilvános üzem. A Messenger alkalmazás a https://m.me/tevepp linken, a rendszer forráskódja a https://github.com/kebou/tevepp oldalon érhető el.

A chatbot (konverzációs ágens) olyan program, mellyel a felhasználó természetes nyelvet használva tud interakcióba lépni, és a chatbot ezt megértve tud válaszolni, információt szolgáltatni vagy különböző feladatokat végrehajtani. A chatbotokat gyakran azonosítják az általános intelligenciát megvalósító,

tetszőleges témában párbeszédet folytató programokkal, azonban ez csak egy lehetséges alkalmazási területe a konverzációs ágenseknek [1]. A chatbotok másik nagy csoportjába azok a nyelvi interfésszel rendelkező programok tartoznak, melyek egy előre jól meghatározott témakörben képesek feladatokat ellátni. Ezeket a programokat taszk-specifikus chatbotoknak is nevezzük [2,3]. Az utóbbi időben nagy érdeklődés övezi a virtuális személyi asszisztenseket, melyek a feladat-orientált programoknál szélesebb körben képesek különböző tevékenységeket elvégezni (pl. telefonhívások indítása, internetes keresések végrehajtása, útvonal tervezése, stb.), azonban az általuk megoldható feladatok száma még erősen korlátozott, általános intelligenciát nem valósítanak meg. Ebben a cikkben egy taszk-specifikus chatbotot támogató rendszer működését írjuk le. Az ilyen alkalmazásoknak manapság legelterjedtebb felületét az azonnali üzenetküldő alkalmazások adják, az általunk bemutatott alkalmazást így a Facebook Messenger platform segítségével teszteljük.

Az effajta integráció kulcsfontosságú azért is, mert – a Business Insider riportja szerint [4] – az üzenetküldő alkalmazások jelenleg már magasabb felhasználó számmal bírnak, mint az egyes közösségi hálókhoz tartozó alkalmazások. Ezt mára már a legtöbb üzenetküldő alkalmazásokat készítő cég felismerte, és egyre többen biztosítanak az automatizált üzenetküldéssel kapcsolatban integrációs lehetőséget a fejlesztők számára. A feladat-specifikus chatbotok felhasználási területei cégek esetén leggyakrabban a customer support automatizálása, a márkahűség mélyítése, webshop kialakítása és a felhasználók informálása a vásárlásokkal vagy egyéb eseményekkel kapcsolatos információváltozásokról. A cégekkel való kapcsolattartáson kívül számos kollaborációt segítő eszköz található, melyek az üzenetváltásban harmadik félként részt véve különböző feladatok – időpont-egyeztetés, emlékeztető lista készítés, stb. – végrehajtásában segítik a beszélgetőtársakat. Továbbá léteznek olyan alkalmazások, melyek egy „másik személyként” segítenek különböző információkhoz jutni. A mi rendszerünk egy budapesti tömegközlekedést segítő chatbot-alkalmazás, mely lehetővé teszi, hogy a felhasználók rövid szöveges üzenetek segítségével végezzenek menetrend- és utazástervezést.

2. A Tervezz Velem rendszer

2.1. Funkcionalitás

TeVe, azaz Tervezz Velem egy chatbot, mely a Facebook által készített Messenger Platformon keresztül nyilvánosan elérhető[1]. A chatbot a budapesti tömegközlekedéssel kapcsolatban szolgáltat információkat a felhasználóknak. Segítségével megkaphatjuk a közelünkben lévő megállókból induló járatok indulási idejét, valamint a tartózkodási hely elküldése helyett meg is írhatjuk a járat és a megálló nevét, melynek menetrendjére kíváncsiak vagyunk. Szöveges üzenet alapján, vagy a helyzetválasztó segítségével megjelölt – a BKK szolgáltatása által lefedett – címek között útvonalat is tervezhetünk. Az alkalmazásban továbbá elmenthetjük

[1] https://m.me/tevepp

kedvenc helyeinket, hogy azokra később az útvonaltervezés során könnyen hivatkozhassunk. A rendszer néhány alapvető társalgási formulát is támogat, így pl. köszönésre illedelmes választ és a funkcionalitást bemutató segédüzenetet kaphatunk. Az 1. ábra egy tipikus, útvonaltervezésre irányuló párbeszédet mutat be.

1. ábra: Útvonaltervezés a chatbottal

2.2. Felépítés

A Tervezz Velem szövegfeldolgozó rendszere egy nyelvfeldolgozó, egy elemző és pontozó, valamint egy rangsoroló komponensből áll össze. A nyelvfeldolgozó modul végigfuttatja a bemeneti szöveget az `e-magyar` nyelvfeldolgozó lánc tokenizáló, lemmatizáló és morfológiai elemző moduljain (ld. részletesebben a 3. fejezetben). A nyelvfeldolgozás után egy elemzési táblázat készül, mely során az egyes mezőknek megfelelő szókapcsolatok lekérdezésre kerülnek a felhasznált API-któl, és a találatok különböző szempontok alapján pontszámokat kapnak; ennek folyamatát a 4. fejezet írja le. Ezt követően legenerálódik a bemeneti szöveg összes lehetséges felosztása, majd az egyes felosztások az elemzési tábla alapján kiértékelésre kerülnek, végül kialakításra kerül az elemzések sorrendje. A szegmentáló és rangsoroló modul működéséről az 5. fejezet szól.

3. Nyelvfeldolgozás

A szövegfeldolgozó modul a keresett információk kinyeréséhez minden lehetséges szókapcsolatot külön-külön megvizsgál, melyhez a bemeneti szöveget darabokra kell bontani A rangsoroló folyamat során kimagasló pontszámot kapnak a megfelelő ragokat tartalmazó szókapcsolatok. Egy-egy megálló- vagy utcanév esetében pl. a *-tól/-től* vagy *-hoz/hez/höz* rag egyértelmű jele annak, hogy az adott entitás egy útvonaltervezés esetében a kiindulási vagy érkezési helyet jelöli. Az efféle információk kinyeréséhez szükség van a főnévi esetragok felismerésére, az e-magyar feldolgozási lánc morfológiai elemzője ezt is lehetővé teszi [5].

A nyelvfeldolgozó modul első lépéseként a bemeneti szöveg végigfut a GATE szövegelemző rendszeren, melyben az `e-magyar` `emToken`, `emMorph`, `emTag`, `emLem` moduljai kerülnek felhasználásra. Az `emToken` modul tokenizálja a bemeneti szöveget, melynek köszönhetően később lehetségessé válik a szöveg szavankénti vizsgálata. Ezután a morfológiai elemzés következik, melyet az `emMorph` modul végez. A modul a bemenet minden szóalakjához hozzárendeli annak összes lehetséges morfológiai elemzését. Mivel minden szónak több lehetséges morfológiai elemzése is lehet, a szöveg egészét vizsgálva ki kell választani ezen elemzések közül a megfelelőt. Ezt a morfológiai egyértelműsítést az `emTag` modul végzi. Az egyes tokenek szótövezését és a szótövek szófaját az `emLem` modul határozza meg. A GATE szerver eléréséhez készítettünk egy API csomagoló klienst[2]. Ennek a kliensnek a feladata a feldolgozandó szöveg elküldése a GATE szervernek és a visszakapott válasz a továbbiakban könnyen kezelhető formátumra hozása. A GATE kliens a feldolgozási lánc `parseText` moduljából kerül meghívásra, ami az e-magyar elemzés lekérdezése után kiszűri a névelőket, írásjeleket, kötőszavakat és a kérdőnévmásokat a tokenek közül, ugyanis ezek elhagyása megkönnyíti a következő modulok feladatát.

4. Elemzés

A fő elemző modul elkészítése előtt a következő feltevésekkel éltünk:

- a szövegből kinyerendő információk az *indulási hely*, *érkezési hely* és a *járatnév*;
- ezen információk közül bármelyik kettő egyértelműen meghatározza a választ:
  - indulási és érkezési hely együttes jelenléte esetén útvonaltervezés,
  - indulási hely vagy ragot nem tartalmazó hely és egy járatnév találata esetén járatindulási információk,
  - érkezési hely és járatnév együttes előfordulásakor a kiindulási helyre irányuló kérdés;
- úgy tekintjük a szöveget, mintha abban az összes szükséges információ szerepelne – tehát egy-egy lekérdezést követően nem törekszünk további interakcióra, minden esetben előnyben részesítjük azokat az elemzéseket, melyek alapján a kérdés már megválaszolható.

[2] https://github.com/kebou/emagyar

A rendszer egy korábbi verziójában az elemzőlánc egyes moduljai csak az előttük már lefutott komponensek kimenetéhez fértek hozzá - nem elemezhették tehát azokat a tokeneket, melyeket egy magasabb precedenciájú komponens már feldolgozott. Így a rendszer nem volt képes önkorrekcióra, egy korai modul hibája a teljes elemzésre nézve végzetesnek bizonyult. Ezen hibák javítására készítettük el a feldolgozási lánc módosított változatát, melyben az összes lehetséges szókapcsolat teljes elemzését elvégezzük. Ehhez szükség van egy olyan adatstruktúrára, mely tárolja az egyes szókapcsolatokat és a hozzájuk tartozó elemzéseket is. Ennek a nyilvántartásnak létrehozása és adatokkal való feltöltése a `generateMap` modul feladata.

A nyilvántartásra használt tárolót egy kétdimenziós tömbként implementáltuk. A tömb és minden altömb elemszáma megegyezik a modul bemeneteként kapott tokenek számával, azaz egy négyzetes mátrixnak is tekinthető. Erre a kétdimenziós tömbre a továbbiakban elemzési táblázatként fogunk hivatkozni. A tömb első dimenziója a szókapcsolat kiindulási tokenjének, második dimenziója a záró tokenjének indexeit tartalmazza. Tehát például a *„Batthyány tér 1 Oktogon”* szöveg *tér 1 Oktogon* szókapcsolatának elemzését az elemzési tábla `[1][3]` indexelésével érhetjük el, a *Batthyány tér* szókapcsolatét pedig a `[0][2]`-vel. Lehetőséget teremtettünk a szókapcsolatok maximális tokenszámának korlátozására, tehát megszabhatjuk, hogy egy-egy feldolgozandó egység maximum hány szóból állhat. Ez a beállítás gyorsíthat a feldolgozáson, mivel nem vizsgáljuk meg a fölöslegesen hosszú szókapcsolatokat. Az elemzési táblázat egy kitöltésére az 1. táblázat mutat példát - a maximális tokenszám beállítása miatt a szürkével jelölt mezők üresen maradnak, ezekre nem fut le az elemző algoritmus.

	Batthyány	tér	1	Oktogon
Batthyány	(0,0)	(0,1)	(0,2)	
tér		(1,1)	(1,2)	(1,3)
1			(2,2)	(2,3)
Oktogon				(3,3)

1. táblázat. A *„Batthyány tér 1 Oktogon”* szöveg elemzési táblájának vázlata

A táblázat minden érvényes mezőjében az adott szókapcsolatot reprezentáló `MapNode` csomóponti objektum található. Ez az objektum tartalmazza a mezőhöz tartozó tokeneket, a szókapcsolathoz tartozó elemzési találatokat és a csomópont jóságát jelző pontszámot. Az egyes elemzési találatokat egy `MapElement` objektum írja le, mely tartalmazza a találat típusát, tartalmát és pontszámait. Egy csomóponti objektum értékeinek feltöltése két részben történik. Először elemzésre kerül a szövegrész, majd az összes lehetséges elemzés átfut a pontozó logikán.

4.1. Elemzés

Első lépéseként ragok keresése történik a csomóponthoz tartozó utolsó token morfológiai elemzése alapján. A megfelelő ragok jelenléte és felismerése megkönnyíti az egyes információ-egységek elhatárolását, és segítenek az adott cím vagy megálló mondaton belüli szerepének (kiindulás vagy érkezés) megállapításában is. Kiindulási helyet az *elativus (-ból, -ből)*, *delativus (-ról, -ről)* és *ablativus (-tól, -től)* ragok jelenthetnek, érkezési helyet pedig az *illativus (-ba, -be)*, *sublativus (-ra, -re)*, *allativus (-hoz, -hez, -höz)* és *terminativus (-ig)* ragok. Bizonyos esetekben előfordulhat, hogy a morfológiai elemző nem talál ragot az adott tokenen, pedig az mégis tartalmazza azt. Jó példa erre a *„Nyugati pályaudvarról”* szókapcsolat, ugyanis jelen írásmód esetén a morfológiai elemzés helyesen megadja, hogy a *pályaudvarról* szón megtalálható a *-ról* delativusi esetrag, azonban a rag ékezetének elhagyásával, a *pályaudvarrol* morfológiai elemzéséről már nem mondható el ugyanez. Az ilyen esetek helyes kezelése érdekében a rag meglétének és típusának vizsgálata reguláris kifejezésekkel is megtörténik. A rag jelenlétét és szerepét jelző információ a csomóponti objektumban kerül nyilvántartásra.

Következő lépésként lekérdezésre kerül a szókapcsolat a BKK Futár API-tól mint megálló és a Google Maps Geocoding API-tól mint cím. Sikeres lekérdezés esetén a találat elhelyezésre kerül a csomóponti objektum lehetséges elemzéseit tartalmazó tömbjébe. Egyetlen tokent tartalmazó csomópont esetén egy reguláris kifejezés ellenőrzi, hogy a token értelmezhető-e járatnévként, és ha igen, akkor ez is bekerül a lehetséges elemzések közé. Végül ha a szövegrész megegyezik a felhasználó által elmentett kedvenc helyek valamelyikével, akkor ez a hely szintén mentésre kerül az elemezések közé.

4.2. Pontozás

Az elemző komponensek lefutása után a csomóponti objektum a pontozó logikához kerül, aminek feladata a csomópont különböző szempontok szerinti pontozása a teljes bemenet lehetséges elemzéseinek későbbi rangsorolása érdekében. A szempontok súlyozásához felhasznált értékek jelenleg manuális próbálgatás eredményeként kerültek meghatározásra. Az értékelés két szinten történik: pontozásra kerülnek a csomópont szókapcsolatához tartozó elemzések, és kiértékelődik maga a csomópont is. A logika mindkét szinten moduláris felépítésű, így könnyen bővíthető, a jövőben tetszőleges újabb tulajdonságok szerint kiértékelhetőek a találatok.

Jelenleg az elemzési eredmények szintjén három, csomópontok szintjén pedig két szempont alapján történik a pontozás. Az elemzési szinten megállapításra kerül a csomóponthoz tartozó szókapcsolat és a elemzés eredményeként megkapott szöveg közti *Levenshtein-távolság*. Ez azt jelenti, hogy ha például valamelyik API a *Nyugati* sztringre a *Nyugati pályaudvar* kimenetet adja vissza, akkor e két sztring hasonlóságát vizsgáljuk. A távolság meghatározásakor a kis- és nagybetűk, illetve az ékezetes karakterek és azok ékezet nélküli párjai azonos karakternek számítanak, mivel a bemeneti szövegben ezek gyakran keveredhetnek. A

Levehnstein-távolság alapján adott hasonlósági pontszámot az (1) képlet alapján számoljuk, ahol a és b a két összehasonlítandó szövegrész, $levenshtein(a, b)$ a két szövegrész közötti szerkesztési távolság, w a pontozási szempont súlya, $|a|$ és $|b|$ pedig a két sztring karakterszámát jelöli.

$$s = w * \left(1 - \frac{levenshtein(a, b)}{max(|a|, |b|)}\right) \tag{1}$$

Szintén az egyes elemzések jóságát jelző pontszám a **relatív hossz**, mely a keresési szövegrész és az eredmény szövegrész közötti szószám összehasonlítására szolgál. Értéke a szerkesztési távolsághoz hasonlóan a (2) képlet alapján áll elő - ezúttal $|c|$ és $|d|$ a két sztring szavainak számát jelöli. A relatív hossz segítségével jutalmazhatjuk a kereséshez szószámban legközelebb álló találatot, és büntethetjük a nagyon eltérőt.

$$s = w * \left(1 - \frac{||c| - |d||}{max(|c|, |d|)}\right) \tag{2}$$

Az egyes elemzések a **találat típusa** alapján is értékelésre kerülnek. A típusra vonatkozó értékekkel megadható a találatok fontosságának sorrendje. A sorrend kialakítása történhet egyéni preferencia szerint is, azonban jelen alkalmazás esetén ezen pontozási szempont segítségével korrigálható az API-k lazán kapcsolódó találatainak kezelése is. Ugyanis a Google Maps Geocoding API-ja sokkal megengedőbb, mint a BKK Futáré, azaz kis egyezés esetén is ad vissza találatokat, ezért a megálló típusú találatok több pontot kapnak az egyes címeknél. Szintén ezzel tudjuk beállítani, hogy egy más elemzéssel egyező nevű kedvenc hely találatakor a kedvenc kerüljön kiválasztásra.

A csomóponti értékelési szempontok egyike a **rag alapú** pontozás. Ezen szempont alapján abban az esetben kap pontot a csomópont, ha tartalmaz a kiindulási vagy érkezési helyeknek megfelelő esetragot. Külön értékkel súlyozható a morfológiai elemzés révén megállapított és a reguláris kifejezések által megtalált rag, így nagyobb mértékben jutalmazhatjuk a megbízhatóbb morfológiai elemzés találatait és például a csak látszólag ragozott *Szentendre* szó nem kap túl magas pontszámot.

Az egyik legfontosabb és legérzékenyebb pontozási szempont az **abszolút hossz** alapú értékelés. Az abszolút hossz megegyezik a csomóponthoz tartozó tokenek számával. Az ezen szempont alapján adott pontszám befolyásolja, hogy milyen súllyal szerepeljenek a meghatározott hosszú szókapcsolatok a rangsorolás folyamán. Vegyük példaként a *„Fehérvári út 17”* szöveg elemzését. Abban az esetben, ha nem különböztetjük meg a csomópontokat a hozzájuk tartozó tokenek száma alapján, vagyis az abszolút hosszra mindegyik azonos pontszámot kap, akkor a rangsorolás után a legjobb eredmény ebből a három komponensből fog állni: `1) Fehérvári`, `2) út`, `3) 17`. Ennek oka egyszerűen az, hogy mindegyik szóra külön-külön is lett találat, és mivel a több szóból álló szókapcsolatok nem érnek jelentősen többet az egyszavasaknál, a három szóra kapott pontszám összege magasabb értékű, mint egy kétszavas és egy egyszavas találat pontjainak összege. Azonban ha a kétszavas szókapcsolatoknak magasabb pontszámot

adunk, már más a helyzet. A jól megválasztott pontszámmal elérhetjük, hogy három egyszavas elemzési találat pontjainak összege kevesebb legyen, mint egy kétszavas és egy egyszavas összege. Ebben az esetben a rangsorolás után a legjobb eredmény a `1) Fehérvári út` és a `2) 17` komponenseket fogja tartalmazni. Ha a három tokenből álló szókapcsolatokat jobban előnyben szeretnénk részesíteni, mint a kettőből állókat, akkor – a mintát követve – ezek abszolút hosszára olyan pontszámot kell adnunk, amely egy egyszavas és egy kétszavas szókapcsolat pontjainak összegénél magasabb. Azonban könnyen eljuthatunk egy olyan szószámhoz, ahol már nem szabad ennyire jutalmazni a hosszúságot. Már négy szó esetén is adódhatnak problémák. Tekintsük a *„Fehérvári út 17 Andrássy út”* mondatot. A Google API megengedő tulajdonsága miatt az *út 17 andrássy út* szókapcsolatra is ad vissza találatot, ami négy szóból áll. Vagyis ha a négyszavas szókapcsolatoknak magasabb pontszámot állítanánk be, mint a két és háromszavasak összege, akkor a rangsor végén a találat így alakulna: `1) Fehérvári`, `2) út 17 andrássy út`. Azonban jól látható, hogy egyáltalán nem ezt a szegmentálást szerettük volna elérni. Javíthatunk ezen a helyzeten, ha négyszavas szókapcsolatok pontszáma nem haladja meg a két és három szavasok pontszámának összegét.

Annak érdekében, hogy tartsuk magunkat ahhoz az alapfeltevésünkhöz, miszerint úgy kezeljük a szöveget, mintha abban minden információ-egység rendelkezésünkre állna, külön esetként kerül kezelésre az a szókapcsolat, mely a teljes szöveget lefedi. Az ilyen szókapcsolat kevesebb pontot kap, mint amennyit a tokenszáma alapján kapna, ugyanis előfordulhat, hogy valójában több információegységet is lefed. Visszatérve korábbi példánkhoz, bár a *„Fehérvári út 17”* szöveg önmagában is megfeleltethető egy cím típusú információ-egységnek, mivel a teljes szöveget lefedi, ezért a kisebb pontszám miatt hátrébb kerül a rangsorban és eredményül a `1) Fehérvári út` megállót és a `2) 17` járatnevet kapjuk, így a felhasználói kérdésre tudunk válaszolni: válaszul a 17-es villamos következő indulási idejét adjuk meg a Fehérvári út megállóból.

Következő lépésként a teljes felhasználói bemenetnek összes lehetséges szegmentációját rangsoroljuk a bennük szereplő csomópontok pontszámaiból kiindulva. A rangsorolást részletesen az 5. fejezet mutatja be. Az egyes csomópontokat ebben a folyamatban a legmagasabb pontszámot elért elemzésük és a csomópont saját pontértéke alapján vesszük majd figyelembe. A csomópontok pontozásakor az egyes szempontok súlyozása egyelőre „trial-and-error” módszerrel történt (ld. a 2. táblázatot), tanítóadat birtokában azonban később lehetőség nyílhat valamely *learning-to-rank* algoritmus alkalmazására is, így ezeket a paramétereket úgy tudjuk majd beállítani, hogy a lehetséges elemzések rangsorolása minél jobban megfeleljen a tényleges felhasználói szándéknak.

4.3. Elemzési példa

A 3. ábrán az elemzési táblázat feltöltésére és a pontozó logika működésére adunk példát. Látható, hogy a táblázat oszlopainak és sorainak száma megegyezik az elemzendő szöveg szószámával, illetve a főátló alatt nem tartalmaz elemeket. Minden mező első sorában a csomópontra vonatkozó információk találhatóak.

Pontozási szempont	Típus	Súlyozás
rag	morfológiai elemzés	22
	reguláris kifejezés	20
abszolút hossz	1 token	80
	2 token	220
	3 token	550
	4 token	690
	5 token	300
relatív hossz	-	10
szerkesztési távolság	-	1
elemzés típusa	járatnév	10
	megálló	10
	kedvenc hely	15
	cím	1

2. táblázat. A pontozási szempontok súlyozása

A mezőhöz tartozó indexek után a csomópont pontszáma szerepel. Ezen információkon kívül megtalálhatóak a mezőben a csomóponthoz tartozó szókapcsolat elemzései, melyeknél megálló esetén `STOP`, cím találata esetén `LOC` felirat után a megfelelő API-tól visszakapott szöveg, majd az adott elemzés két tizedesre kerekített összpontszáma szerepel. Az elemzésekkel egy cellában találhatóak még az egyes értékelési szempontok és a hozzájuk tartozó pontszám is.

A `(0,1)` indexű csomóponthoz a *Silvanus setanyrol* szöveg tartozik. Láthatjuk, hogy a szókapcsolatra két elemzési találat született, egy megálló és egy cím. Ezek közül a megállónév összpontszáma nagyobb, ezt aláhúzással jelöl a táblázat. A 21-es összpontszám komponensei: a találat típusára 10 pont, a relatív hosszra 10 pont, a szerkesztési távolságra 1 pont. A relatív hossz pontszám értéke maximális, mivel a keresendő *Silvanus setany* és a visszakapott *Silvanus sétány* szószáma egyezik. A szerkesztési távolság pontszáma szintén maximális, mivel rag találata esetén a rag nélküli szövegrész kerül keresésre és a keresett és visszakapott szövegben csupán a kis- és nagybetűk, illetve az ékezetek különböznek, melyek ilyenkor nem számítanak különböző karakternek. A csomópont a ragnak köszönhetően 20 pontot, a hozzá tartozó tokenek száma alapján abszolút hosszára pedig 220 pontot kap. Így a végső pontszáma ezen két érték és a legmagasabb elemzési találat pontszáma alapján 261 pont lesz. A többi csomópont értéke hasonlóan kerül kiszámításra. Látható, hogy a harmadik oszlop első és második sorának pontszáma negatív, mivel a *Silvanus setanyrol deak* és a *setanyrol deak* szókapcsolathoz nem tartozik elemzési találat. A táblázatban dupla aláhúzással szerepelnek azon csomópontok pontszámai, melyek a rangsorolás (ld. 5. fejezet) után a legtöbb pontot elérő szegmentációban szerepelnek.

	0 - Silvanus	1 - setanyrol	2 - deak
0	(0,0) - **95.53** - absoluteLength: 80	(0,1) - **261** - absoluteLength: 220 - hasRole: 20	(0,2) - **-10000** - absoluteLength: 80
	STOP: Silvanus sétány: **15.53** - editDistance: 0.53 - relativeLength: 5 - typeBasedScore: 10	STOP: Silvanus sétány: **21** - editDistance: 1 - relativeLength: 10 - typeBasedScore: 10	-
	LOC: Visegrád: **11.13** - editDistance: 0.13 - relativeLength: 10 - typeBasedScore: 1	LOC: Bp., Silvanus sétány- **8.27** - editDistance: 0.6 - relativeLength: 6.67 - typeBasedScore: 1	
1	-	(1,1) - **135.43** - absoluteLength: 80 - hasRole: 20	(1,2) - **-10000** - absoluteLength: 220
		LOC: Budapest: **11.13** - editDistance: 0.13 - relativeLength: 10 - typeBasedScore: 1	-
2	-	-	(2,2) - **93.6** - absoluteLength: 80
			STOP: Deák Ferenc tér: **13.6** - editDistance: 0.27 - relativeLength: 3.33 - typeBasedScore: 10

3. táblázat. A *„Silvanus setanyrol deak”* szöveg elemzési táblázata

5. Rangsorolás

Az elemzéseket és pontozásokat tartalmazó táblázat elkészítése után a következő lépés a rangsorolás, mely során a teljes bemeneti szöveg összes lehetséges felosztása közül választjuk ki a megfelelőt. Elsőként legeneráljuk a bemeneti szöveg összes lehetséges szegmentálását. Ahogy az elemzési táblánál, itt is beállítható egy maximális hossz, ekkor egyetlen szegmentum sem lehet ennél hosszabb, így sok szegmentáció kizárható. A szövegnek megengedett olyan felosztása is, amely nem tartalmazza az összes tokent: a természetes nyelven megfogalmazott útvonal-tervezés például a kiindulási és érkezési helyeken felül további, az információ kinyerése szempontjából közömbös szót tartalmazhat (*Hogyan jutok el...*, stb.).

Hogy megértsük, miért van szükség minden lehetséges felosztásra, tekintsük az *„Albertfalva utca 17”* szöveget. Ennek a bemenetnek kétféle olvasata is lehetséges: ha egyetlen szegmentumként kezeljük, egy pontos címet értünk alatta, egy másik értelmezés szerint a felhasználó a *17-es* villamos indulására kiváncsi az *Albertfalva utca* megállóból. Ahhoz, hogy mindkét értelmezés később pontozásra kerüljön, elő kell állítani az összes értelmezési lehetőséget. A két lehetséges ér-

telmezés szürke színnel van jelölve az összes lehetséges felosztást tartalmazó 4. táblázatban.

Index párok	Infromáció-egységek
(0,0)	[Albertfalva]
(0,1)	[Albertfalva utca]
(0,2)	[Albertfalva utca 17]
(0,0), (1,1)	[Albertfalva], [utca]
(1,1)	[utca]
(0,0), (1,2)	[Albertfalva], [utca 17]
(1,2)	[utca 17]
(0,0), (2,2)	[Albertfalva], [17]
(0,1), (2,2)	[Albertfalva utca], [17]
(0,0), (1,1), (2,2)	[Albertfalva], [utca], [17]
(1,1), (2,2)	[utca], [17]
(2,2)	[17]

4. táblázat. Az *„Albertfalva utca 17”* szöveg lehetséges particionálásai

A rangsoroló logika minden egyes szegmentációhoz az elemzési tábla alapján egy számot rendel. Egy szegmentációhoz úgy rendelünk pontszámot, hogy az abban szereplő indexpároknak az elemzési táblában megfelelő csomópontok pontszámait összeadjuk. Tehát az `[Albertfalva utca], [17]` sor értéke pl. az elemzési tábla `[0][1]` és `[2][2]` mezőiben található pontszámok összege.

6. A chatbot alkalmazás

Befejezésül bemutatjuk, hogyan integráltuk az eddigiekben leírt architektúrát egy a Facebook Messenger platformon működő chatbot-alkalmazásba és milyen tapasztalatokkal gazdagodtunk a nyilvános tesztelés során.

Az alkalmazáslogika Node.js – szerver oldali JavaScript – nyelven íródott és egy virtuális gépen fut. Amikor a felhasználó üzenetet küld a Messenger alkalmazáson keresztül, a Facebook továbbítja azt az általunk megadott webhook URL irányba. Az URL-re érkező kéréseket az alkalmazáslogika üzenettípusok szerint dolgozza fel. A Messenger lehetőséget kínál a szöveges üzenet mellett egyéb üzenettípusok küldésére is: a felhasználó küldhet képeket, videókat, különböző csatolmányokat, illetve választhat a chatbot által kínált gyors-válaszok közül, és használhatja az egyes üzenetsablonokon elhelyezett gombokat is.

Az alkalmazás magját a BKK Futár API-ja képezi, mely az egyes megálló és járatnevek validációjáért, az indulási információkért és az utazástervezésért egyaránt felel. Használatával számos további információt elérhetünk a budapesti tömegközlekedéssel kapcsolatban: forgalmi változások, járatok aktuális pozíciója, tervezett és tényleges indulások. A megállón kívüli egyéb helyek keresésére a

Google Maps Geocoding API-ját használjuk, ennek segítségével a szövegesen megadott címekből egységesen strukturált, koordinátákkal ellátott objektumokat kaphatunk.

Az adatok tárolása relációs adatbázis helyett MongoDB NoSQL dokumentum alapú adatbázisban történik, mely jelentősen megkönnyíti a különféle forrásból származó adatok konzisztens mentését. A bemeneti szöveg feldolgozása során jelentős szerepet játszik a program mellett futó GATE szerver és a hozzá kapcsolódó e-magyar feldolgozási lánc moduljai is.

6.1. Az elemző modul integrációja

A chatbot szövegelemző komponensének elkészítése során az átstrukturálhatóság és a későbbi bővíthetőség érdekében egy könnyű, moduláris feldolgozási lánc megvalósítását tartottuk szem előtt. A feldolgozási sor egyes komponensei azonos interfészt valósítanak meg: minden modul meghíváskor két paramétert kap, az egyik egy, a feldolgozási láncon végigmenő, közös kontextus objektum, amelybe minden modul tetszőlegesen helyezhet új értékeket és amelyből bármely értéket felhasználhat. A másik paraméter egy, a következő modul hívására szolgáló függvény. Amennyiben ez a függvény nem kerül meghívásra, illetve nincsenek további modulok, a feldolgozási sor véget ér.

A chatbot alkalmazásához beérkező üzenet először végigfut a `wit.ai` szolgáltatáson, mely az alapvető felhasználói szándék felismerésére szolgál. Ennek az elemezésnek köszönhetően kerülnek feldolgozásra a köszönések, funkció hívások és segítségkérések. Ha ilyen felhasználói szándékot nem azonosítunk, működésbe lép a korábbi fejezetekben ismertetett logika. Először a nyelvfeldolgozó modul az `e-magyar` segítségével feldolgozza a szöveget, majd feltöltésre kerül az elemzési táblázat. Végül legenerálódnak a lehetséges felosztások és megtörténik a rangsorolás.

A kialakult rangsor alapján a következő modul meghatározza a kiindulási helyek, érkezési helyek és járatnevek lehetséges értékeit és elhelyezi azokat a kontextusban. Ezen lépés után olyan modulok futnak le, melyek ha megtalálnak minden, a működésük számára szükséges értéket a kontextusban, akkor feladatuknak megfelelő üzenetet küldenek a felhasználónak. Amennyiben egyik komponens sem talál elegendő információt, azaz nem sikerült értelmezni a felhasználó üzenetét, egy erről szóló üzenet kerül kiküldésre.

6.2. A chatbot használat közben

A chatbot nyilvános tesztelése 2017. októberében indult, az első 6 hét után több, mint 170 aktív felhasználóval rendelkezik és ez idő alatt 16000 beérkező üzenetet dolgozott fel. Átlagosan napi két-három felhasználó tervez utazást vagy keres menetrendi információkat segítségével. A Tervezz Velem chatbot az elemző modul egy korábbi – a rangsoroló logikát még nem tartalmazó – verziójával kezdte meg a nyilvános működést. A cikkben bemutatott új modul bekötése után jelentős javulást várunk az egyes információ-típusok felismerésével kapcsolatban. A

lehetséges továbbfejlesztések közül az utazástervezés időpontjának beállíthatósága szerepel még rövidtávú terveink között, a felhasználói visszajelzések alapján ugyanis erre mutatkozik a legnagyobb igény.

7. Köszönetnyilvánítás

Köszönjük a SZTAKI Nyelvtechnológiai Kutatócsoportjának, hogy biztosították a nyilvános teszteléshez szükséges infrastruktúrát, valamint két névtelen bírálónknak a hasznos észrevételeket.

Hivatkozások

1. Carpenter, R.: Cleverbot. http://www.cleverbot.com (2017) Letöltés ideje: 2017.12.05.
2. Weizenbaum, J.: ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine. Communications of the ACM **9**(1) (1966) 36–45
3. Colby, K.M., Weber, S., Hilf, F.D.: Artificial paranoia. Artificial Intelligence **2**(1) (1971) 1–25
4. BI Intelligence: Messaging apps are now bigger than social networks. http://www.businessinsider.com/the-messaging-app-report-2015-11 (2017) Letöltés ideje: 2017.11.13.
5. Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholcz, I., Novák, A., Indig, B., Prószéky, G., Farkas, R., Vincze, V.: Az e-magyar digitális nyelvfeldolgozó rendszer. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017), Szeged, Szegedi Tudományegyetem (2017) 49–60

TANIT – Magyar nyelvű szövegeket elemző eszköz összehasonlító digitális bölcsészeti feladatokhoz

Labádi Gergely[1], Farkas Richárd[2], Nagy Roland[2], Péter Róbert[3]

[1] Szegedi Tudományegyetem,
Magyar Nyelvi és Irodalmi Intézet
[2] Szegedi Tudományegyetem,
Informatika Intézet
rfarkas@inf.u-szeged.hu
[3] Szegedi Tudományegyetem,
Angol-Amerikai Intézet
robert.peter@ieas-szeged.hu

Kivonat: Cikkünk a TANIT (Text ANalysIs Tools) rendszer célkitűzését, funkcióit és használatát mutatja be. A TANIT rendszer célja, hogy magyar nyelvű szövegek számítógépes nyelvészeti feldolgozásával dokumentumok összehasonlító elemzéséhez szükséges statisztikákat kigyűjtsön. Ez a webszolgáltatás létező nyelvtechnológiai elemzőlánc kimenetére épülő aggregált statisztikákat számít ki, témamodelleket épít és ezeket olyan formátumban adja át a digitális bölcsész felhasználónak aki utána programozói ismeretek nélkül is fel tudja ezt használni kutatásaiban.

1 Bevezetés

Mindannyian tudjuk és tapasztaljuk, az információs társadalom korszaka jelentős kihívás a humán tudományok számára: a változás egyaránt érinti a tudás társadalmi intézményeit, legitimációját, ám ugyanúgy kihat mindennapjainkra, tudományos gyakorlatunkra. Mindez alkalmat nyújt – valójában sürget – a humántudományi kutatások tárgyának, módszerének és közegének újragondolására [1]. Ez az egyik legfontosabb célja a digitális bölcsészetnek (*digital humanities*). A nyugati tudományosságban a digitális bölcsészet a 2000-es évek második felére szervezetileg, intézményileg végleg és egyértelműen áttört, azaz vannak képzések, folyóiratok, kutatóközpontok, konferenciák, a bölcsész állásoknál rendszerint elvárás valamiféle DH-képesség vagy -gyakorlat.

A digitális tudományos kutatás formabontó és úttörő kutatási lehetőségekkel kecsegtet a bölcsészettudományokban, mivel a szövegek és az ezekhez rendelt metaadatok elemzéséhez számtalan új lehetőséget kínálnak, melyeket e tudományterületen korábban nem alkalmaztak. Az új digitális módszerek és eszközök használatát a világszerte gomba módra szaporodó digitális bölcsészet tanszékeken oktatják. A hallgatók megtanulják például miként kell adatbázisokat létrehozni, szövegeket felcímkézni. Született már olyan egyetemi tankönyv is, amely kifejezetten a nyelv és irodalom szakosokat célozza meg a hatalmas szövegkorpuszok és

adatbázisok statisztikai elemzéséhez és ábrázolásához használt R programnyelv megismertetésével, amelynek van kifejezetten nyelvészeti és irodalmi elemzésekhez használható csomagja is [2].

A Szegedi Tudományegyetemen létrejött – jelentős részben a 2017. szeptemberében elhunyt Labádi Gergelynek köszönhetően – egy digitális bölcsészet fókuszú informális együttműködés különböző karok különböző intézetei közt: Magyar Nyelvi és Irodalmi Intézet (BTK), Angol-Amerikai Intézet (BTK), Informatika Intézet (TTIK), Klebelsberg Kuno Könyvtár.

Laptopos bemutatónkban demonstráljuk a TANIT (TANIT – Text ANalysIs Tools) rendszert, ami ennek az együttműködésnek az első materializálódott eredménye. A TANIT rendszer célja, hogy kis mennyiségű, tipikusan néhány tucat, magyar nyelvű dokumentum mélyebb számítógépes nyelvészeti feldolgozásával dokumentumok összehasonlító elemzéséhez szükséges statisztikákat kigyűjtsön. A TANIT egy publikusan elérhető webes felület (http://dighum.bibl.u-szeged.hu/tanit), ahol a bemeneti dokumentumok feltöltése után lehetőség van az elemzés néhány paraméterének beállítására, majd a felhasználó az összehasonlító statisztikákat webes felületen böngészheti és táblázatos formában letöltheti további elemzés céljára. A TANIT célfelhasználója az a digitális bölcsész, aki ismeri a számítógépes nyelvészetben rejlő lehetőségeket, de nem akar saját nyelvtechnológiai pipelinet építeni, csak a számára legfontosabb statisztikákat megkapni.

A TANIT funkcióinak tervezésekor néhány valós összehasonlító digitális bölcsészeti kutatás nyelvtechnológiai igényeiből indultunk ki. Például Labádi Gergely egyik összehasonlító irodalomtudományi tanulmányában [3] azt a kérdést igyekszik megválaszolni, hogy Jókai *A kőszívű ember fiai* című művének különböző átdolgozásai mennyire hasonlítanak az eredeti műhöz, mennyire tekinthetőek ugyanazon műnek. Például Nógrádi átdolgozása egy „modernizált” változat. Pedagógusok szerint ugyanis a mai diákok nem értik – és így nem is olvassák – a klasszikus (és kötelező) irodalmat, kifejezetten a regények nyelvezetét találják problémásnak, a nyelvre helyezik a hangsúlyt, ami érinti az igeidőket, a latin, német kifejezéseket, és természetesen magukat a mondatstruktúrákat is.

A kőszívű ember fiai különböző változatainak összehasonlító kutatásához olyan nyelvtechnológiai kérdések megválaszolására van szükség, mint

- Az egyes művekben milyen a szavak és mondatok hosszának eloszlása?
- Mennyire dinamikus/eseményorientált a mű? Ez az igék és egyéb szófajok arányával közelíthető.
- Az egyes témák ugyanolyan súllyal szerepelnek-e a változatokban?

2 Kapcsolódó munkák

Nemzetközi szinten számos tudományos publikáció [4,5,6,7] és elérhető digitális eszköz[1] született hasonló célzatú összehasonlító digitális bölcsészeti kutatások

[1] Mallet: http://mallet.cs.umass.edu
Interactive Text Mining Suite: https://languagevariationsuite.shinyapps.io/TextMining/

támogatására. Az elérhető eszközök általában az angol nyelvre koncentrálnak. Vannak köztük olyanok, amelyek akár tucat nyelv kezelését is ígérik, de ez általában egyszerű szótövezést takar csak. Megvizsgáltuk ezen eszközök pontosabb és mélyebb magyar nyelvre kidolgozott eszközökkel való kibővítésének lehetőségét, de technológiai oldalról ez nem volt kivitelezhető. Ezen vizsgálódások folyományaként döntöttünk úgy, hogy saját rendszert implementálunk. A TANIT tervezése során azonban erősen építettünk a megismert nemzetközi megoldásokban alkalmazott elemzési lépésekre.

Magyar nyelvtechnológiai oldalról tekintve is több kapcsolódó tudományos publikáció és elérhető eszköz létezik (például e-magyar[2] [8], magyarlanc[3] [9], hun* eszközök[4] [10]). Az elérhető magyar nyelvű szövegek számítógépes nyelvészeti feldolgozását támogató elemzők és elemzőláncok kimenetele már strukturált metaadat, például egy mondat szavainak morfológiai leírása, de még nem kezelhető könnyen, programozói ismeretek szükségesek a kiment feldolgozásához, ugyanis ezekből még aggregált statisztikák leszámolására van szükség egy összehasonlító digitális bölcsészeti kutatáshoz. Továbbá az összehasonlítások talán legfontosabb szempontja a dokumentumok tartalmi elemzése és legjobb tudomásunk szerint nem érhető el szabadon felhasználható témamodellező (topic modelling) eszköz magyar nyelvű szövegek elemzésére. A TANIT célja, hogy létező nyelvtechnológiai elemzőlánc kimenetére épülő aggregált statisztikákat kiszámítsa, témamodelleket építsen és ezeket olyan formátumban adja át a digitális bölcsész felhasználónak aki utána programozói ismeretek nélkül is fel tudja ezt használni kutatásaiban.

3 Alapstatisztikák

A feltöltött dokumentumokon a TANIT mondat és szószegmentációt, majd szófaji egyértelműsítést futtat. Egyéb konfigurálható előfeldolgozási lépések után a TANIT minden dokumentumra külön leszámolja majd egyetlen táblázatba szervezve visszaadja az alábbi egyszerű statisztikákat:

- mondatok száma
- szavak száma (token)
- átlagos mondathossz (szó/mondat)
- különböző szóalakok száma (type)
- különböző szótövek száma (lemma type)
- minden egyes fő szófaji kódra az adott morfológiai elemzéssel ellátott szavak száma

Az alapstatisztikák mellett a művek szókincsgazdagságának jellemzésére két statisztikát számol a TANIT. A magyar irodalomtudományi szakirodalomban elfogadott Guiraud-index:

Stylo: https://sites.google.com/site/computationalstylistics/

[2] http://e-magyar.hu/

[3] http://www.inf.u-szeged.hu/rgai/magyarlanc

[4] http://mokk.bme.hu/en/eszkozok/

$$\frac{type}{\sqrt{token}}$$

A Guiraud-index nem alkalmas a szöveghosszból fakadó különbségek semlegesítésére. Ennek feloldására javasolt Herdané [11] egy másik képletet:

$$\frac{\log type}{\log token}$$

4 Témamodellezés

Az összehasonlító elemzések tartalmi szintjéhez a TANIT két különböző eszközzel járul hozzá. Mindkét esetben a Latent Dirichlet Allocation (LDA) valószínűségi témamodellező (topic modelling) algoritmust [12] alkalmazzuk. Az LDA egy nagy dokumentumhalmazból képes előre megadott számú téma automatikus azonosítására. A látens témákat leírhatjuk azon szavak listájával amelyek a kérdéses témában szereplésének feltételes valószínűségei a legnagyobbak. Továbbá az LDA minden, a modellező halmazban szereplő és abban nem szereplő dokumentumhoz a topikok egy eloszlását rendeli. Különböző művek ezen eloszlásainak összevetése érdekes tartalmi különbségekre világíthat rá.

A TANIT két különböző megközelítése témamodellezésre abban tér el, hogy milyen dokumentumhalmazon modellezi a témákat. A témamodell alkalmazása (dokumentumok topik eloszlásának számítása) már ugyanúgy működik:

- Lehetőségünk van egy előre betanított LDA használatára a feltöltött dokumentumainkon. Ehhez jelenleg egy az MNSz2 [13] szépirodalmi alkorpuszán tanított modellt biztosítunk. Folyamatban van a Magyar Elektronikus Könyvtár (MEK) feldolgozása, ami eggyel nagyobb nagyságrendű szöveghalmazt biztosít majd. Ezen a korpuszon már érdemes lehet különböző szempontok szerint szűrt részkorpuszokon – például 19. sz. magyar irodalom – különböző LDA modelleket tanítani.
- Vagy magán a feltöltött dokumentumokon is taníthatunk egy LDA modellt. A TANIT kevés számú mű elemzését célozza meg, ami nem elégséges mennyiség az LDA tanításához. Ezért lehetőség van egy csúszó ablakkal a művek kisebb kvázi-dokumentumokra bontására majd ezeken a kvázi-dokumentumokon LDA tanítására. Azt tapasztaltuk, hogy néhány száz szó méretű ablak már elégséges együtt előfordulási kontextust biztosít az LDA számára, de az ablakméretet a felhasználó is állíthatja. Ennek a megközelítésnek az az előnye, hogy a témák kizárólag a feltöltött műveket bontják le témákra. A jövőben tervezzük a csúszó ablakokból a mű cselekményének időbeli fejlődésének vizualizációját és dokumentumok közti összevetését is.

A második megközelítés alkalmazásával Labádi [3] tudta azonosítani hogy Nógrádi átdolgozott, modernizált verziója “a Baradlay-gyerekek mellé az édesanyjukat érintő

eseményszálakat domborította ki", míg az eredeti Jókai "a család köré fonódó intrikát emeli ki".

5 Technológiai megvalósítás

A TANIT egy böngészőből elérhető szolgáltatás. A vékony front-end egy JSP weboldal amin keresztül feltölthetjük a szerverre az elemzés alanyául szolgáló dokumentumokat. A dokumentumokat egyesével megadva vagy egy .zip állományba csomagolva is feltölthetjük. Nyers szövegfájlok (.txt) feldolgozására van a rendszer jelenleg felkészítve aminek karakterkódolását előre beállíthatja a felhasználó.

A felhasználó az alábbiakat konfigurálhatja a weboldalon ha az alapértelmezett beállításokkal nem elégedett:

- saját stopword szótár használata (feltöltése)
- kiegészítő szótőtár hozzáadása
- kisbetűsítés, számok cseréje
- LDA-nál:
    - előre betanított modellek közül választás
    - ha saját belső modellt tanul, akkor a csúszóablak mérete
    - topikok száma

Az elemzés elindításával a dokumentumok feltöltésre kerülnek egy egyetemi szerverre. A dokumentumokat a nyelvtechnológiai elemzések elkészülte után azonnal töröljük a szerverről. Az egyidejű felhasználók kiszolgálása session alapú kiszolgálása biztosítja, hogy azok nem férhetnek hozzá mások adatához.

A server oldal egy Tomat web application server, ami Java nyelvű kódot futtat. A nyelvtechnológiai elemzésekhez a magyarlanc [9] szó- és mondatszegmentációját valamit szófaji egyértelműsítését (ami a környezetfüggő szótöveket is megadja) használtuk fel. A saját stopword listák és kiegészítő szótőtárak miatt kisebb módosításokat kellett végrehajtanunk a magyarlancon. Az LDA témamodellezést a Mallet[5] [14] könyvtárainak felhasználásával végezzük. A leszámolt statisztikákat visszaadjuk a front-endnek megjelenítésre és egy .xlsx fájlt is megkonstruálunk belőlük. Ebben a fájlban különböző lapokra (spreadsheet) kerülnek a különböző elemzések, az egy lapon szereplő táblázatok sorai pedig az egyes dokumentumoknak felelnek meg. Az .xlsx fájlt a felhasználó letöltheti és azon könnyen további aggregált statisztikákat számolhat. Az eredmény fájlt 24 óráig tároljuk a szerveren, utána törlésre kerül.

6 Összegzés és további tervek

Laptopos bemutatónkban bemutatjuk a http://dighum.bibl.u-szeged.hu/tanit rendszert, ami összehasonlító digitális bölcsészeti kutatások támogatására készült

[5] http://mallet.cs.umass.edu

nyelvtechnológiai rendszer. A rendszer célja, hogy bármiféle programozói vagy nyelvtechnológiai ismeret nélkül használhassák azt bölcsész kutatók és hallgatók.

A TANIT első verziója elkészült és publikusan elérhető. A jövőben tervezzük funkcionalitásának bővítését, például különböző témamodellek biztosításával, szintaktikai elemzőkre építve grammatikai hasonlósági mértékek alkalmazása művek közt, plágiumkeresők implementálása stb.

Köszönetnyilvánítás

A munka az EFOP-3.6.1-16-2016-00008 azonosítójú, EU társfinanszírozású projekt támogatásával készült el. Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Bibliográfia

1. http://digibolcsesz.ek.szte.hu/
2. Jockers, M. L. (2014): Text Analysis with R for Students of Literature (Quantitative Methods in the Humanities and Social Sciences). Springer-Verlag, Cham (2014);
3. Labádi, G.: Latra vetett szavak: A kőszívű ember fiai kiadásairól és átiratairól. Digitális Bölcsészet (2018) in press.
   online változat: http://www.academia.edu/34473500/Labadi_Latravetettszavak.pdf
4. Ramsay, S.: Reading Machines: Toward an Algorithmic Criticism University of Illinois Press, Champaign, IL (2011)
5. Jockers, M. L.: Macroanalysis: Digital Methods and Literary History. University of Illinois Press, Urbana (2013)
6. Moretti, F. Distant Reading. Verso, London (2013)
7. Schreibman, S., Siemens, R. and Unsworth, J. (eds) A New Companion to Digital Humanities. Blackwell, Oxford (2016)
8. Váradi T., Simon E., Sass B., Gerőcs M., Mittelholcz I., Novák A., Indig B., Prószéky G., Farkas R., Vincze V. Az e-magyar digitális nyelvfeldolgozó rendszer. Magyar Számítógépes Nyelvészeti Konferencia (2017)
9. Zsibrita, J., Vincze,V., Farkas,R. magyarlanc: a toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP 2013, pp. 763–771 (2013)
10. Halácsy P., Kornai A., Oravecz Cs. Hunpos an open source trigram tagger. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (2007)
11. Herdan, G. The Advanced Theory of Language as Choice and Chance. Springer-Verlag, Berlin, Heidelberg, New York. (1966), 75–77.
12. Blei, D., Ng A., Jordan, M. Latent dirichlet allocation. Journal of Machine Learning Research 3 (2003), 993-1022.
13. Oravecz Cs., Váradi T., Sass B.: The Hungarian Gigaword Corpus. In: Proceedings of LREC (2014)
14. McCallum, A. MALLET: A Machine Learning for Language Toolkit. (2002)

Nagyfelbontású pragmaszemantikai igazságértékelés egy játékprogramban

Nőthig László, Szeteli Anna

Pécsi Tudományegyetem, ℜeALIS Elméleti, Számítógépes és Kognitív Nyelvészeti Kutatócsoport
{nothig.laszlo, anna.szeteli}@gmail.com

Kivonat: Tanulmányunk azt mutatja be, hogy előkészületben lévő programunk megsokszorozott igazságértékelések eredményeit figyelembe véve hogyan szimulálja diskurzusok kimenetelét és perlokúciós hatásmechanizmusának érvényesülését lépésről lépésre, továbbfejlesztve Benz DRT alapú multiágens rendszerét. Az extenzionális szempontok mellett figyelembe veszünk olyan intenzionális állapotváltozásokat, mint a beszélő és a hallgató információállapotának, (másikról való) hiedelmeinek, vágyainak, szándékainak megváltozása, melyek közül utóbbi már a külső világmodell változására is hatással lehet.

Kulcsszavak: diskurzusreprezentáció, formális szemantika, pragmatika

1 Bevezetés

Ahogyan egyszerű, referenciális kifejezések – mint például a *nap* vagy az *iskola* szó – poliszém jelentéshálózatát is szükséges kidolgozni és implementálhatóvá tenni [1], illetve a gép számára a különböző jelentéseket felismerhetővé tenni, úgy mi a különböző, korábban töltelékszónak tartott diskurzusjelölők [2], illetve formailag és/vagy funkciójukban hasonló mondattípusok [3] közötti jelentés-egyértelműsítésre [4] törekszünk magunk is [22].

Célunk egy olyan program létrehozása, amely egy világmodell szimulációja, és annak részeként tartalmaz beszélőket és hallgatókat az ő információállapotukkal együtt, melynek részei a potenciális feladói és címzetti szerepeik. Ezekkel a modellben fennálló eseményszerűségekről tesznek állításokat úgy, hogy eközben akár vissza is hathatnak az eseményszerűség állapotára. Modellünkben ötvözni kívánjuk a pragmatikából ismert sikerességi feltételeken [5] nyugvó elképzeléseket [6][7] a formális szemantika igazságértékelő eljárásával [8].

Végső soron Anton Benz [9] mesterségesintelligencia-kutató elképzelését szeretnénk továbbfejleszteni, aki egy DRT alapú [10] *multiágens* [11] rendszert dolgozott ki. A lényeg abban áll, hogy interpretálók egy csoportja egy (egyes tagjai által észlelt) eseménnyel kapcsolatban folyamatosan bővíti / módosítja a külvilágban

történtekről való tudását, illetve (legalábbis az adott témakörben) az egyes részcsoportokhoz rendelhető *közös* tudást – ami helyett mi a ℜeALIS még finomabban disztingváló megközelítését követve inkább a (rekurzívan gyarapítható) „*kölcsönös* tudásokra" építjük majd az interpretációs kalkulust, illetve a külvilágmodellben való változtatásokat is implementáljuk. Ha ugyanis számot akarunk adni dialógusok jól formáltságáról adott interpretálói információállapotokban, akkor nem kerülhetjük el az interpretálók egymásról (és ezen belül kiemelten egymás tudásáról) való *kölcsönös* (de nem feltétlenül szimmetrikus) tudásának ábrázolását.

Törekvéseink a DRT pragmatikai továbbfejlesztését célul tűző SDRT [12] elmélet elképzeléseit implementáló RUDI szoftver [13] alkotóinak törekvéseivel is rokoníthatóak. Közös vonás a pragmatikai tényezők és a külvilágmodell közötti kapcsolatrendszerre irányuló kitüntetett figyelem. A RUDI rövidítés mögött a következő név áll: *Resolving Underspecification with Discourse Information*, kiemelve a diskurzusbeli információ szerepét az alulspecifikált nyelvi formák egyértelműsítésében, különös tekintettel az időre utaló kifejezések, a beszédaktusok és a háttérben meghúzódó célok háromszögére. Az is közös a ℜeALIS és a RUDI alkotóinak felfogásában, hogy a nyelv finom és gyakran rejtett eszközeinek jelzéseit minden korábbi "plan-recognition" megközelítésnél mélyebben kívánja figyelembe venni [14], a legfrissebb formális szemantikai és pragmaszemantikai eredmények bevonásával.

Térjünk vissza a saját, ℜeALIS alapú modellünk ismertetésére! Egy mondatban – például *A fiam itthon van.* – igazságértékelésnek vethető alá, hogy a világmodellben a fiaimként definiált entitásokat tartalmazó halmaz részhalmaza-e az itthon levők halmazának, illetve pontosan egy ilyen entitás van-e – több entitás jelenlétének esetén a határozott névelővel nem egyértelműen referáltunk. Egy dinamikus diskurzusreprezentációt megvalósító rendszerben ugyanígy egy sor tényező vizsgálatával, azaz mintaillesztést végezve igazságértékelhetünk. A sikerességi feltételek bevonása megköveteli, hogy az extenzionális értékelés mellett intenzionálisan, az elmék és a világmodell között, illetve elmék és diskurzusok reprezentációja között is értékeljünk.

A következőkben bemutatjuk a mintaillesztési eljárást, majd azt, hogy a ℜeALIS-keret [7] formális eszköztára egy diskurzusokat építő játékprogramban hogyan tudná szimulálni egy valós diskurzus lehetséges kimeneteleit, illetve perlokúciós hatásmechanizmusait, ténylegesen modellezve a külvilágban megvalósuló változásokat is.

2 Pragmaszemantikai mintaillesztés

Mintaillesztési eljárásunk pragmatikai vonatkozása a sikerességi feltételek [5][15], társalgási maximák [16] formalizálásában és bővítésében nyilvánul meg, mely eljárást más reprezentacionalista keretek is alkalmazzák [12]. Emellett célunk, hogy a kimenet egy formális, montagoviánus gyökerű [8] igazságértékelő eljárásból fakadjon.

A fenti elméletekből következik egy olyan elgondolás, hogy a beszélő jogosultan ölthet magára egy feladói szerepet, hallgatójára pedig címzetti szerepet oszthat,

amennyiben ezt információállapota lehetővé teszi [6]. E jogosultság vizsgálatához az üzenet feladóját el kell választani a gondolkodó beszélőtől, amely szemléletet a ℜeALIS-keret már aktívan alkalmazza, egyrészt konvencionális feladói profilok definiálására, tehát nyelvészeti jelenségek valamilyen fokon szituációfüggetlen értelmezésére [2][4]. Másrészt az elválasztásból adódóan a feladói szerep – akár szándékosan – jogosulatlan felöltése gyakran a pragmatika tárgykörébe eső jelenségeket eredményez, a mintaillesztési hiba megtalálásával pedig ezekre a jelenségekre egzakt formális definíció adható [17].

Emellett az általunk szimulálni kívánt világmodellnek része a hallgató információállapota is, ami meghatározza őt abból a szempontból, hogy alkalmas-e a beszélő által ráosztott címzetti szerep felvételére, képes-e olyan módon akkomodálni a feladói profil elemeit, hogy az eredeti feladói szándék megvalósulhasson. Ez a címzetti szerep és a hallgató közötti kompatibilitás vizsgálatát kívánja meg.

A sikerességi feltételek figyelembe vétele által megsokszorozódott axiómákból adódó profilok ezáltal kettős elbírálás alá esnek az igazságértékelési folyamatban: a beszélő és a hallgató alkalmassága is kérdéses, mielőtt a feladó szándékát megvalósulni látjuk.

A fenti feladatot úgy valósítjuk meg, hogy a Kamp-féle logikai alapú diskurzusszemantikai reprezentációt (DRT) [10] ontológiailag újjászervező ℜeALIS-keretben kompozicionálisan [18] értelmezett profilokat vetjük össze az információállapotokkal és a világmodellel. A ℜeALIS gyökeresen új ontológiája ezt lehetővé teszi, mivel abban a beszélők elméje a világmodell része, így beszélői információállapotukhoz tartoznak az élethossziglani diskurzusreprezentációk is. Címkerendszer szerinti részben rendezés révén feladói hiedelmekre, vágyakra, autoritásokra és szándékokra utaló jegymátrixok alkotják az intenzionális profilt, de a beszélő és a hallgató hiedelmei, vágyai is ezen a módon ábrázoltatnak, ezáltal összevethetjük őket.

A tanulmányban bemutatásra kerülő játékprogramban a világmodell változása mintaillesztések eredményéből adódik, célja megragadni néhány alapvető nyelvi jelenség funkcióját, illetve beszélő–hallgató interakció formájában visszaadni ezeket a funkciókat. Hosszabb távú célunk egy olyan felület, ahol a finomhangolók/mondattípusok repertoárja bővíthető, hogy ezáltal egyre bonyolultabb diskurzusokat kezelhessünk.

3 Kettős illeszkedés a sikeres diskurzusban

A diskurzus sikerességét a kettős illeszkedésből fakadóan meghatározzuk egyrészt aszerint, hogy a feladó saját beszélői állapotára való tekintettel megfelelő intenzionális profilt választott-e. Másrészt pedig aszerint, hogy a hallgatója képes-e a címzetti szerepnek eleget téve a profil elemeit felismerni, a rá vonatkozó részeket pedig magára vonatkoztatni, vagy akkommodálni.

A Delphi fejlesztői környezetben írt, Windows alatt futó játékprogram alapja, hogy a felhasználónak szavakkal kell megcselekednie egy kapott feladatot, például Elemér – a világmodell egyik entitása – tudja meg, hogy Cili – egy másik entitás – otthon van!

$\langle B,M,E,\tau^{+},+\rangle$[1] Megjelenik egy feladó- és egy címzettválasztási lehetőség, hogy mely beszélők és hallgatók vegyék magukra az üzenet átadásához és átvételéhez kötődő szerepeket. Emellett megjelennek a világmodellben lehetséges eseményszerűségek és a modalitások/finomhangoló kifejezések. A felhasználó egy ikon segítségével a folyamat során bővítheti tudását a világmodellről, ha nincs ötlete, hogy kit válasszon és milyen szerepekre.[2] Mivel Elemér az, aki még nem tud a propozicionális tartalomról, a felhasználó megpróbálhatja először őt címzettnek választani. Eltérő lesz azonban a sikeresség attól függően, hogy kit választ feladónak. Lehetnek olyan feladók, akik nem beszélhetnek az eventualitásról, mert például nem tudnak róla (táblázat kijelentő mondatának autoritás és hiedelem axiómái). Ha a feladónak meg is van a lehetősége elmesélni Elemérnek, amit az hallani szeretne, akkor se vehető biztosra, hogy adott feladónak vágya elmesélni azt, illetve nem biztos, hogy tudja, Elemér hallani szeretné az adott információt. Ha hibás mintaillesztés történik, a felhasználó megtudja annak okát, emellett az információk segítenek neki abban, hogy a megfelelő feladót válassza. A feladat legjobb, de nem egyetlen megoldása először Elemért feladónak felvenni (címzett helyett) és magát Cilit megkérdezni, aki ezután tudni fogja, hogy Elemér szándéka megtudni a dolgot őáltala (ld.: kérdő mondat szándék paramétere), így a következő fordulóban feladóként elmondhatja neki. Egy nehezített pályán persze megtiltható, hogy éppen Cilit valaki címzettnek vegye fel.

Ezen a ponton az „Ez téged miért érdekel?" esettel árnyaljuk a mintaillesztés jelentősége mellett ezúttal a jellegéből adódó határait. Azt szemléltetjük, hogy a profilok egyes elemei a beszélő és a hallgató elméjében eltérő értéket vehetnek fel. A fenti reakció érkezhet például a „Cili otthon van?" kérdésre, amennyiben Elemér Dénesnek, Cili férjének teszi fel azt. Elemér a kérdő profil értelmében felteheti a kérdést azért, mert jóban vannak Dénessel / jóban érzi magát vele, és azt gondolja, barátja szívesen mesélne a családjáról. Az információ címzettnek való átadásához fűződő vágyat mindkét félnek tulajdoníthatjuk kijelentés és kérdés esetén is. Dénesnek viszont nem jutott volna eszébe erről mesélni, ezért úgy értelmezi a szituációt, hogy mivel számára érdektelen a téma, Elemér az, aki kifejezetten meg szeretné tudni az információt. Ugyanez megeshet felszólító profil esetén például a *Menjünk haza!* mondat használatakor, amikor a beszélő majd a hallgató is elvégzi a mintaillesztést a saját elméje és a felszólító profil között, mindketten a másikról gondolva, hogy az haza szeretne menni, miközben egyikőjüknek sincs ilyen vágya. A profilok alulspecifikáltságából adódó rugalmasság tehát egyúttal félreértésekhez is vezethet.

Olyan feladat is lehet a világmodellben, ahol nem csupán az elmékben módosulnak információállapotok eventualitásokról, hanem a világban is megváltozik az igazságértékük. Ilyen feladat például, hogy Anti, a világmodell egyik entitása vásároljon be. Ugyanúgy megjelennek a feladó- és címzettválasztási lehetőségek, a

---

[1] Az információ megjelenik/megjelenhet magyarázó jegymátrixok formájában is, a képlet jelentésének kibontása sorban a következő: Az eventualitásra nézve, hogy Cili otthon van-e (e_3), a tudása (**B**elief) legyen biztos erősségű (**M**aximal) Elemérnek (**E**) a következő pillanatban ($\boldsymbol{\tau}^{+}$) és pozitív (+).

[2] A felhasználó megoldásait pontozva, a pontozási rendszerrel egyensúlyba hozható, hogy ugyanúgy ne érje meg számára "találomra" párbeszédet kezdeményeztetni, mint folyamatosan a segítséget igénybe venni.

világmodellben lehetséges eseményszerűségek és a modalitások/finomhangoló kifejezések. A felhasználó szokás szerint bővítheti tudását a világmodellel kapcsolatban, ha nincs ötlete, hogy kit válasszon és milyen szerepekre. A program például arra figyelmezteti, hogy Bea Anti felesége (ő azt szeretné, ha Anti bevásárolna, illetve valamekkora hatalommal is bír felette ezen a téren). s)he should trust in others.

A felhasználó, ekkor megpróbálhatja Beát választani a *bevásárlás* eventualitás tekintetében például egy felszólító feladói szerepre, de ha ezt teszi, a dolog nem valósul meg. A program igazságértékelések sorozatával arra jut, hogy Bea ebben az esetben nem tudott szándékot ébreszteni Antiban, mert valamilyen elvárásának nem felelt meg a kínált profil.

1. táblázat: Kijelentő és kérdő profilok a ℜeALIS-ben: $\lambda=\langle B,\gamma,AR,\tau,+\rangle$; $\lambda'=\langle D,\gamma_{r^*},r^*,\tau,+\rangle$, ahol $\Sigma\gamma_x \geq 1$ ($x \in r^*$) és pref.: $r^*=\{AR,ae\}$; $\lambda''=\langle I,M,AR,\tau,+\rangle$; $\lambda'''=\langle I,M,ae,\tau^+,+\rangle$

	Ugye-Kérdő	Kijelentő	Kérdő
HIEDELEM	$\langle B,M,AR,\tau,0\rangle$	$\langle B,M,AR,\tau,+\rangle$	$\langle B,M,AR,\tau,0\rangle$
	$\langle BD,\gamma_u,AR,\tau,+\rangle$ $\gamma_u \geq gr$		
	λ^$\langle B,\gamma'',ae,\tau,+\rangle$ ^$\langle B,M,AR,\tau,0\rangle$	λ ^$\langle B,M,ae,\tau,0\rangle$	λ^$\langle B,\gamma'',ae,\tau,+\rangle$ ^$\langle B,M,AR,\tau,0\rangle$
VÁGY	λ^λ' ^$\langle B,M,AR,\tau^+,+-\rangle$	λ^λ' ^$\langle B,M,ae,\tau^+,+-\rangle$	λ^λ' ^$\langle B,M,AR,\tau^+,+-\rangle$
AUTORITÁS	λ^$\langle A,\gamma',ae,\tau,+\rangle$ ^$\langle B,M,AR,\tau^+,+-\rangle$ szigorúbb, mint λ^$\langle B,M,ae,\tau,+-\rangle$	$\langle A,M,AR,\tau,+\rangle$ ^$\langle B,M,ae,\tau^+,+\rangle$	λ^$\langle A,\gamma',ae,\tau,+\rangle$ ^$\langle B,M,AR,\tau^+,+-\rangle$ szigorúbb, mint λ^$\langle B,M,ae,\tau,+-\rangle$
SZÁNDÉK	λ''^λ''' ^$\langle B,M,AR,\tau^+,+-\rangle$	λ'' ^$\langle B,M,ae,\tau^+,+\rangle$	λ''^λ''' ^$\langle B,M,AR,\tau^+,+-\rangle$

4 Tetten ért udvarias szavak

A felhasználó minden fordulóban információt kap, hogy miért nem volt sikeres a mintaillesztés, mielőtt újból próbálkozhat. Anti nem szereti, ha parancsolgatnak neki. Általában elfogadja azonban az utasításokat, ha úgy gondolja – esetünkben Bea megnyilatkozása nyomán elméjének részévé válik a hiedelem – hogy az üzenet feladója figyelembe veszi az eventualitással kapcsolatos vágyait, meg akarja ismerni azokat. Ezt fejezik ki a „Vásárolj be, ha nincs ellenedre!”, illetve a *Bevásárolnál?* [18] mondatok.

1. ábra: Példa egy mintaillesztési hibára

Emellett Anti „hős” típusként akkor is nagy eséllyel segít, ha kellően hangsúlyosnak érzi, hogy a feladó vágyát teljesítve mekkora szívességet tesz. Ehhez az szükséges, hogy a feladó explicitté tegye saját maximális vágyát:

(1) a. *Bárcsak* bevásárolnál!
b. *Ugye* bevásárolsz?
c. Vásárolj be *nekem*!

A *nekem* szócskát a fenti mondatban tekintsük diskurzusjelölőként, ami arra utal, hogy a feladó vágya pozitív a megvalósulás tekintetében. Ezzel kontrasztba állítható a *Nyugodtan* vásárolj be! mondat, amelynek használata az adott szituációban nem vezetne sikerre, mivel leginkább autoritás adást fejez ki [4], tehát a másik részéről feltételez vágyat.

2. táblázat: Felszólító profilok a ℜeALIS-ben: $\lambda=\langle B,\gamma,AR,\tau,+\rangle$; $\lambda'=\langle D,\gamma_{r^*},r^*,\tau,+\rangle$, ahol $\Sigma\gamma_x \geq 1$ ($x \in r^*$) és pref.: $r^*=\{AR,ae\}$; $\lambda''=\langle I,M,AR,\tau,+\rangle$; $\lambda'''=\langle I,M,ae,\tau^+,+\rangle$

	Feltételes (-nÁl) **kérés**/*kérdés*	Felszólító	Óhajtó (*címzett*)
HIEDELEM	$\langle B,M,AR,\tau,-\rangle$; $\langle B,M,AR,\tau,0\rangle$^$\langle D,M,ae,\tau,+-\rangle$	$\langle B,M,AR,\tau,-\rangle$	$\langle B,M,AR,\tau,-\rangle$
	λ^$\langle B,M,ae,\tau,-\rangle$	λ^$\langle B,M,ae,\tau,-\rangle$	λ^$\langle B,M,ae,\tau,-\rangle$
	λ^$\langle B,\gamma'',ae,\tau,+\rangle$^ $\langle B,M,AR,$^$\tau,0\rangle$^$\langle D,M,ae,\tau,+-\rangle$		
VÁGY	**$\langle D,M,AR,\tau,+\rangle$**; $\langle D,M,AR,\tau,+\rangle$^$\langle B,M,AR,\tau^+,+\rangle$ ^$\langle D,\gamma_{ae},ae,\tau,+-\rangle$	λ^λ'	$\langle D,M,AR,\tau,+\rangle$
AUTORITÁS	triviálisan teljesül: λ^$\langle B,M,ae,\tau,+-\rangle$ ^$\langle D,M,ae,\tau,+-\rangle$; λ^$\langle A,\gamma',ae,\tau,+\rangle$	λ^$\langle A,\gamma',ae,\tau,+\rangle$; λ^$\langle A,M,AR,\tau,+\rangle$ ^$\langle I,M,ae,\tau^+,+\rangle$	$\langle A,M,AR,\tau,0\rangle$
SZÁNDÉK	λ''^λ'''^$\langle B,M,AR,\tau^+,+\rangle$ ^$\langle D,M,ae,\tau,+-\rangle$; **$\lambda''$^$\lambda'''$ λ''^λ'''^$\langle B,M,AR,t^+,+\rangle$ ^$\langle D,M,ae,\tau^+,+-\rangle$**	λ''^λ'''	λ'' ^$\langle B,M,ae,\tau^+,+\rangle$ ^$\langle D,M,AR,\tau,+\rangle$

A második táblázat illusztrálja a felszólító mondat és kompozicionális jelentésüket tekintve nem felszólító mondatok ℜeALIS-beli intenzionális profiljait. Szoftverünk a felhasználónak annyi szabadságot hagy, hogy ugyan elvégzi a mintaillesztést, de az maga dönthet a végső perlokúciós hatás felől. A világot ugyanis – beleértve a hallgató elméjének világocskáit – végül a hallgató által akkomodált profilelemek definiálják újra. Az akkomodációkat tekintve négy mechanizmust veszünk most számba, mivel amellett, hogy (i) a profil centrális elemét, az elsődleges feladói szándékot – helyes illesztés esetén – a világban végrehajtottnak tekintjük, amellett lehetséges (ii) nem teljesen sikeres mintaillesztés esetén a hiányzó axiómák pótlása. Vegyünk példának

egy felszólítást, ahol a feladó a címzettről – a második axióma értelmében – feltételezi, hogy az tisztában van vele, az eventualitás nem valósult meg. Amennyiben elhangzik a *Zárd el a gázt!* mondat, nem szükséges magyarázkodni, hogy a gáz még nincs elzárva, a címzett egyszerűen pótolja a korábban hiányzó premissza világocskát az eventualitással kapcsolatban. Emellett (iii) bármely profilelem egy *a feladó szándéka, hogy a címzett megtudja* világocskába ágyazódhat, ahogyan azt például a táblázatbeli óhajtó profil utolsó eleme szemlélteti. Címzett felvételekor a vágy közlése történik meg. Kijelentés esetén pedig ez a fajta beágyazás azt az érdekes jelenséget eredményezi, hogy míg egy becsületes beszélő biztos tény közlése közben a *szándékom, hogy tudd* feladói világocskát tárja a címzett elé, a hallgató mégis akkomodálhatja a kijelentés első axiómájára vonatkozó *a feladó szándéka, hogy tudjam, hogy ő ezt tudja* világocskát, ami mögött ártatlan hencegést, de akár hazugságot is sejthet. Végül ezzel szimmetrikusan (iv) megtörténhet a szándék centrális komponens redukciója is, melyre jó példa a *szerint(em)* diskurzusjelölő interpretációja. A *Peti szerint Mari otthon van.* mondat esetén a feladói szándék, hogy a címzett megtudja, Peti hiedelemvilágocskája pozitív abban a tekintetben, hogy Mari otthon van [20]. A hallgatónak a feladó és Peti tudása felé tanusított nagymérvű bizalma lehetővé teheti számára adott esetben, hogy a fenti (iii) „szándékemeléssel" éppen ellentétesen ne a „valaki szerint valami" alakban tárolja az információt, hanem elfogadja azt igaznak.

5 Összegzés, továbbfejlesztési lehetőségek

A bevezetőben beszámoltunk róla, hogy a korai, a természetes nyelveket – a maguk komplexitásában való – megragadásához még steril, extenzionálisan igazságértékelő eljárásokat a diskurzusreprezentáción keresztül pragmatikai jelenségek felé nyitva kezdték el fejleszteni. A ℜeALIS elméleti keretet és implementációját is a hasonló törekvésű reprezentacionalista kezdeményezések soraiban helyeztük el, a második egységben pedig a keret pragmatikai gyökereit és ontológiai újítását ismertettük.

A harmadik egységben megmutattuk, hogy a mintaillesztési-eljárás axiómáról axiómára igazolhatja a beszélő jogosultságát feladói szerepére, illetve a hallgató és az átvett profil közötti illesztés eredménye a címzettválasztás helyességéről is állítást tesz. A negyedik egységben felgöngyölítettük az esetet, amelyben a választott intenzionális profil nem felelt meg a hallgató információállapotának. Végül számot adtunk róla, hogy a szándékolt feladói profilok perlokúciós hatásmechanizmusa hogyan variálódhat a hallgató általi akkomodációktól függően.

Projektünk jelenlegi, első fázisában a külvilág modelljében a figyelembe vehető eventualitások még csak olyan egyszerű állapotok (vagy legalábbis ilyenként vannak kezelve), amelyek a "játék" egy adott pillanatában vagy fennállnak, vagy nem. Például Péter egyes világfázisokban nős, másokban nem. Bea egyes fázisokban dolgozik, másokban nem. Péter egyes pillanatokban imádja Beát, másokban nem. Időnként havazik, időnként nem. A ℜeALIS-keretben ennél sokkal finomabb megközelítés létezik az eventualitások eseményszerkezetére nézve [21]: az időtengelyt öt szakaszra osztják fel. Az első és az utolsó szakaszban nem áll fenn az eventualitás. A második a

felkészülési szakasza: például a *hazautazófélben* van kifejezés arra utal, hogy valaki csomagol, jegyeket vásárol, búcsúzkodik. A harmadik szakasz a megvalósításé: az iménti példában ilyenkor utazik a cselekvő. Ezt így fejezi ki a nyelv: *éppen utazik haza*. Végül a negyedik az eredményszakasz: pl. *már hazautazott / otthon van*. Nemcsak mennyiségi kérdés persze, hogy egy eventualitásnak egy adott világfázisban kettő helyett öt állapotát tekintjük-e, hanem a vizsgált intenzionális profilok elméleti kérdéseit is érinteni fogja egy ilyen kiterjesztés. *Utazz csak haza!* – egy ilyen *csak*-kal finomhangolt biztató (vagy más intonáció mellett éppen fenyegető) felszólítás például inkább a felkészülési szakaszban lévő cselekvőnek szólhat, szemben az *Utazz haza!* bázisváltozattal, amit a legelső, még a felkészülési szakaszt megelőző állapotban használhatunk. Egy másik továbbfejlesztési lehetőség: két fokozat helyett egyes állapotok esetében –1 és +1 közötti számmal jellemezni, hogy milyen mértékben áll fenn az adott állapot. Hiszen így igazságértékelhető például a *Péter nagyon szereti Beát* mondat.

Köszönetnyilvánítás

Köszönettel tartozunk az EFOP-3.4.3-16-2016-00005 számú, „Korszerű egyetem a modern városban: Értékközpontúság, nyitottság és befogadó szemlélet egy 21. századi felsőoktatási modellben” című projektnek, mely a Széchenyi 2020 program keretében, az Európai Szociális Alap támogatásával lehetővé tette a szerzők számára a konferenciarészvételt.

Köszönjük a névtelen bírálók hasznos tanácsait, megjegyzéseit!

Hivatkozások

1. Pustejovsky, J.: The Generative Lexicon. MIT Press, Mass. & London, Cambridge (1995)
2. Kleiber, J., Alberti, G.: Compositional Analysis of Interrogative Imperatives in Hungarian. Manuscript (Univ. of Pécs, Dept. of Linguistics) based on a talk at the Linguistics Beyond and Within Conference. Lublin, 20–21 October (2016)
3. Halm, T.: The Semantics of Weak Imperatives Revisited: Evidence from Free-Choice Item Licensing. Talk at the 13th International Conference on the Structure of Hungarian (ICSH13), Budapest (2017)
4. Kleiber, J.: Similar Intentions with Different Underlying Wishes: Intensional Profiles of Imperatives in Hungarian. Poster presented at SinFonIJA X. (2017)
5. Austin, J. L.: How to Do Things with Words. Clarendon Press, Oxford (1975/1962)
6. Oishi, E.: Discursive functions of evidentials and epistemic modals. In: Cantarini, S. et al. (eds.) Certainty-uncertainty – and the attitudinal space in between. Studies in Language Companion Series 165. pp. 239–262 Benjamins, Amsterdam (2014)
7. Alberti G.: ℜeALIS: Interpretálók a világban, világok az interpretálóban. Akadémiai Kiadó, Budapest (2011)
8. Dowty, D. R., Wall, R. E., Peters, S.: Introduction to Montague Semantics. Reidel, Dordrecht (1981)
9. Benz, A.: Chains and the Common Ground. In: Poesio M., Traum D. (eds.) GötaLog 2000 — Gothenburg Papers in Computational Linguistics 00–5 pp. 181–184 (2000)

10. Kamp, H., Genabith, J., Reyle, U.: Discourse Representation Theory. In: Gabbay D., Guenthner F. (eds.) Handbook of Philosophical Logic 15. pp. 125–394 Springer-Verlag, Berlin (2011)
11. Fagin, R., Halpern J.Y., Moses Y. and Vardi M.: Reasoning about Knowledge. MIT Press, Cambridge (1995)
12. Asher, N., Lascarides A.: Logics of Conversation. Cambridge Univ. Press, Cambridge (2003)
13. Schlangen, D., Lascarides, A. and Copestake, A.: Resolving Underspecification using Discourse Information. In: Kühnlein, P., Rieser H., Zeevat H. (eds.) Proceedings of the 5th International Workshop on Formal Semantics and Pragmatics of Dialogue. pp. 79–93 Bielefeld (2001)
14. Asher, N., Lascarides A.: Imperatives in dialogue. In: Kühnlein P., Rieser H. and Zeevat H. (eds.) The Semantics and Pragmatics of Dialogue for the New Millenium, New Series 114. pp. 1–24 Benjamins, Amsterdam (2003)
15. Searle, J. R.: Expression and meaning – Studies in the Theory of Speech Acts. Cambridge Univ. Press, Cambridge (1979)
16. Grice, P.: Logic and conversation. In: Cole P., Morgan, J. L. (eds.) Syntax and Semantics 3. pp. 41–58 Academic Press, New York (1975)
17. Alberti, G., Vadász, N., Kleiber, J.: Ideal and Deviant Interlocutors in a Formal Interpretation System. In: Zuczkowski, A. et al. (eds.) Communicating Certainty and Uncertainty in Medical, Supportive and Scientific Contexts. Dialogue Studies 25. pp. 59–78 Benjamins, Amsterdam (2014)
18. Groenendijk, J., Stokhof, M., Veltman, F.: Coreference and Modality. In: Lappin, S. (ed.) The Handbook of Contemporary Semantic Theory. pp. 179–213 Blackwell, Oxford (1996)
19. Károly, M., Alberti G., Szeteli, A.: On the Prosodic Structure of Imperative and Real Hungarian Questions. Talk at the Linguistics Beyond and Within Conference. Lublin, 18–19 October (2017)
20. Szeteli, A.: *Szerint(em).* A kifejezés funkcióinak vizsgálata formális interpretációs rendszerben. In Böhm G., Czeferner D., Fedeles T. (szerk.) Szemelvények 4. Válogatás a PTE BTK XXXIII. OTDK-n I–III. helyezést elért hallgatóinak pályaműveiből. pp. 88–107. PTE BTK TDKT, Pécs (2017)
21. Farkas, J., Ohnmacht, M. Aspect and Eventuality Structure in a Representational Dynamic Semantics. In Alberti, G., Farkas, J., Kleiber, J. (eds.) Vonzásban és változásban pp. 353–379 Doctoral School of Linguistics at Univ. of Pécs, Hungary (2012)
22. Szeteli, A., Alberti, G., Kleiber, J., Dóla, M.: The World is Built with our Words to Each Other – Basic and Fine-Tuned Intensional Profiles in Hungarian. Jelen kötet.

Szerzői index, névmutató

Ács Evelin, 295
Ács Judit, 146
Alberti Gábor, 78

Beke András, 231
Berend Gábor, 272
Boros István, ifj., 109

Csapó Tamás Gábor, 196
Csuma-Kovács Ádám, 405

Dobi Jan Sándor, 423
Dóla Mónika, 78
Dömötör Andrea, 25, 305

Farkas Richárd, 405, 450
Ferenczi Zsanett, 260

Galántai Júlia, 317
Gosztolya Gábor, 171, 196, 219
Grósz Tamás, 171, 196
Gulyás Attila, 109, 317

Hamp Gábor, 393
Havadi Gergő, 109
Hoffmann Ildikó, 219
Horváth Csilla, 331, 413

Ilyés Virág, 88
Indig Balázs, 3, 25, 125

Jászai Henrietta, 206

Kalivoda Ágnes, 3
Kálmán János, 219
Kemény Boldizsár, 437
Kiss Margit, 423
Kleiber Judit, 78
Kovács Viktória, 381

Labádi Gergely, 450
Laki László János, 281
Lázár Bernadett, 340
Ligeti-Nagy Noémi, 25, 52

Makrai Márton, 352
Markó Alexandra, 196
Markovich Réka, 393
Mészáros Tamás, 423
Mittelholcz Iván, 260

Nagy Ágoston, 413
Nagy T. István, 405
Nagy Roland, 450
Naszódi Mátyás, 38, 51
Németh Gergely Dániel, 146
Novák Attila, 67, 159
Novák Borbála, 67, 159
Nőthig László, 456

Nyíri Zsófi, 88

Pákáski Magdolna, 219
Péter Róbert, 450
Pólya Tibor, 360, 371
Puskás Éva, 360
Puskás László, 371

Recski Gábor, 295, 437

Sass Bálint, 16, 352
Simkó Katalin Ilona, 381
Simon Eszter, 135, 260
Sliz-Nagy Alex, 405

Szabó Martina Katalin, 88, 109, 317, 340
Szakadát István, 393
Szántó Zsolt, 405
Szaszák György, 183
Szebeni Zea, 317
Szeteli Anna, 78, 456
Szilágyi Norbert, 413

Tarján Balázs, 183
Tóth László, 171, 196, 219
Tulics Miklós Gábriel, 206
Tündik Máté Ákos, 183

Ugray Gábor, 247

Vadász Noémi, 3, 25, 135
Vicsi Klára, 206
Vincze Veronika, 100, 219, 340, 381, 405, 413

Yang Zijian Győző, 281, 305