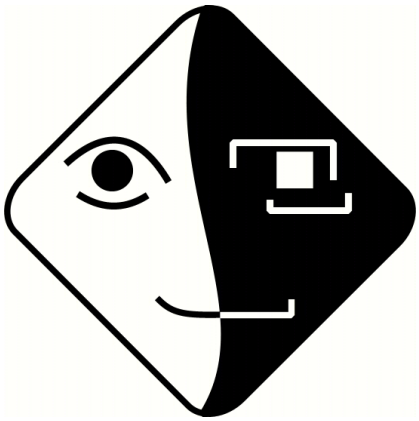
V. Magyar Számítógépes Nyelvészeti Konferencia



MSZNY 2007

Szeged, 2007. december 6-7.
http://www.inf.u-szeged.hu/mszny2007

ISBN: 978-963-482-848-8

Szerkesztette: Tanács Attila és Csendes Dóra
{tanacs, dcsendes}@inf.u-szeged.hu

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

Nyomtatta: Juhász Nyomda
6771 Szeged, Makai út 4.

Szeged, 2007. november

Előszó

2007. december 6-7-én ötödik alkalommal kerül megrendezésre a Magyar Számítógépes Nyelvészeti Konferencia. Nagy örömömre szolgál, hogy a rendezvény évről évre változatlan érdeklődést vált ki az ország különböző tájairól. A konferencia fő célja továbbra is a nyelvtechnológia területén elvégzett vagy folyamatban lévő kutatások és fejlesztések legaktuálisabb eredményeinek bemutatása. Az idei konferencián nagyobb hangsúllyal szerepelnek a beszédtechnológiai vonatkozású előadások, ami azt is mutatja, hogy az elmúlt év során nagy lendülettel haladtak előre az ilyen vonatkozású kutatások. Tanúi lehetünk továbbá a beszéd- és tartalom-feldolgozást ötvöző kutatások első eredményeinek, amelyek jelentős előrelépés irányába mutatnak a nyelvtechnológia területén.

Az idei felhívásra beérkezett tudományos értekezések közül a programbizottság 30-at fogadott el előadás megtartására, és további 8-at poszter-, illetve laptopos bemutató megtartására.

Külön öröm, hogy az idei konferenciára É. Kiss Katalin is elfogadta meghívásunkat, így a szakmai program az Ő plenáris előadásával is gazdagabb lesz. A tavalyi alkalomhoz hasonlóan idén is tervezzük a „Legjobb Ifjú Kutatói Munka" díj odaítélését, amellyel a fiatalabb generáció tagjait kívánjuk ösztönözni arra, hogy a nyelvtechnológiai kutatások terén kiemelkedőt alkossanak.

Szeretnék köszönetet mondani a programbizottságnak: Vámos Tibor programbizottsági elnöknek, valamint Alberti Gábor, Gordos Géza, László János, Prószéky Gábor és Váradi Tamás programbizottsági tagoknak. Szeretném továbbá megköszönni a rendezőbizottság: Csendes Dóra és Tanács Attila munkáját.

Csirik János, a rendezőbizottság elnöke
Szeged, 2007. november

Tartalomjegyzék

I. Beszédszintézis

II. Beszédfelismerés

III. Morfo-fonológia a beszédfeldolgozásban

IV. Ontológia

V. Gépi tanulás

VI. Fordítás és korpusz

VII. Pszichológiai vonatkozású fejlesztések

VIII. Poszter– és laptopos bemutatók

I. Beszédszintézis

Promptgenerátor – Ügyfélszolgálati hangos üzenetek automatikus gépi előállítása egy adott bemondó hangjára

Németh Géza, Zainkó Csaba, Fék Márk, Olaszy Gábor, Bartalis Mátyás

Budapesti Műszaki és Gazdaságtudományi Egyetem,
1117 Budapest Magyar Tudósok körútja 2., Magyarország
{nemeth,zainko,fek,olaszy,bartalis}@tmit.bme.hu

Kivonat: Az egyre szélesedő kommunikációs lehetőségekkel rohamosan nő a a telefonos ügyfélszolgálatok terhelése. A tájékoztatás automatizálásához egyre több hangos üzenetet kell elkészíteni, általában ugyanazzal a bemondóval. Ezt a felolvasó személy véges terhelhetősége korlátozza. A cikkben olyan gépi megoldás lehetőségéről számolunk be, amelyik leveszi a munka nagy részét a bemondó válláról, csak ellenőriznie kell a generált üzenet hangzását. A promptgenerátor olyan új beszédtechnológiai megoldás, amilyent még nem készítettek Magyarországon. Tervezése és fejlesztése mind számítógépes nyelvészeti, mind fonetikai és informatikai szempontból új megoldásokat eredményezett. A rendszer, optimális esetben olyan természetes hangminőséget szolgáltat, hogy a hallgató nem veszi észre, hogy gép beszél.

1 Bevezetés

Modern korunkban az információk beszéddel való közlése információs rendszerekben egyre szélesebb körben terjed el. A cikkben egy szűk témakörre vonatkozó célmegoldás kísérletéről számolunk be, nevezetesen a mobiltelefonok használatával és forgalmazásával kapcsolatos ügyfélszolgálati szövegek automatikus meghangosításával. A kutatás-fejlesztés újszerűsége abban áll, hogy a szintetikus beszédet teljesen természetes hangzási minőségben és egy adott bemondó hangszínezetével kell előállítani. A cél, hogy az üzenetet hallgató ne vegye észre, hogy gépileg előállított hangot hall. A célkitűzés megvalósításához korpusz alapú, elemkiválasztásos beszédszintetizálási módszer tűnt a legoptimálisabbnak. Ezt az általános módszert úgy alkalmaztuk a célfeladat megoldására, hogy a szintézis alapelemének a szó nagyságú elemeket választottuk.

Általánosságban elmondhatjuk, hogy ennél a technológiánál nagy, olykor több 10 órányi hangfelvételből választják ki a szintézis során a megfelelő hangrészleteket, melyek lehetnek hangok, hangkapcsolatok vagy akár szavak is. Az ilyen szintetizátorokat jellemzően egy-egy jól meghatározott témakörben lehet hatásosan elkészíteni. Ilyen terület például az ügyfélszolgálati üzenetek generálása is. Ezek a szövegek elég kötött nyelvezettel és viszonylag kötött szókinccsel rendelkeznek. Lássunk néhány promptszöveget:

Kérjük, válasszon a következő menüpontokból. Tudakozó, 1-es gomb. Hibabejelentés, 2-es gomb. Üzleti ügyintézés, 3-as gomb.
Samsung X510, X160, X660, E370, 1-es gomb. Samsung E900, E870, D900, 2-es gomb. Samsung Z400, P300, Z560, 3-as gomb.
A 3G a jelenlegi leggyorsabb adatátviteli megoldás, amely magában foglalja a jelen és a jövő szolgáltatásait. A 3G hálózat használható egyszerű telefonbeszélgetések lebonyolítására és multimédiás alkalmazásokra is, akár vonalkapcsolt, akár GPRS adatátvitel formájában.

A beszédszintetizátor fejlesztésének alaplépései a következők:

- szövegkorpusz adaptálása a szintézishez (szövegtisztítás és normalizálás),
- a szövegkorpuszból készített hangfelvétel feldolgozása (fonemikus átírás, címkék elhelyezése),
- a szintézis alapegységének kiválasztása,
- a kiválasztási és összefűzési költségfüggvények megtervezése,
- hibakeresési eljárások kidolgozása,
- a rendszer optimális működésének beállítása.

2 A beszédszintetizátor felépítése

A korpusz alapú beszédszintetizátor két fő modult tartalmaz, egy nagyméretű, megfelelően preparált szöveg és beszédkorpuszt (adatbázis), továbbá egy elemkiválasztó kereső algoritmust, amelyik a bemeneti, szintetizálandó szöveget elemezve keresést végez az adatbázisban és kiválasztja a legoptimálisabbnak tartott hullámforma elemeket és összefűzi azokat.

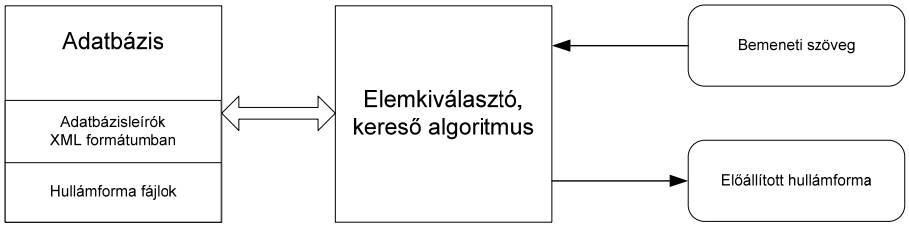


Fig. 1. A promptgenerátor beszédszintetizátor blokksémája

Az adatbázis két nagy részből áll. Egyrészt a hullámformákat tartalmazó hangfájlok gyűjteményéből, melyben minden mondat külön fájlban van tárolva, valamint az adatbázis szöveges leíróinak gyűjteményéből. Ez utóbbiakhoz tartozik a mondatok szövege, a mondatok szövegeinek fonemikus átirata, a hang- és szóhatárok pozíció jelzései a mondatokon belül, alapfrekvencia-adatok. Ezeket az információkat XML struktúrában tároljuk a rendszerben, ezzel is egységesítjük és gyorsítjuk a feldolgozást. Az elemkiválasztó algoritmus az adatbázis szöveges leírói alapján határozza

meg, hogy a bemeneti szövegnek megfelelő hangzó formát mely elemek egymás után illesztésével lehet a természetes hangzáshoz közeli minőségben előállítani.

2.1 A szintézis adatbázisa

Az adatbázis elkészítésénél több fontos szempontot vettünk figyelembe. Elsősorban a kiválasztott témakörhöz tartozó szövegtípusok, és a szintetizálandó mondat összhangját kell biztosítani. Ez esetünkben azt jelenti, hogy az ügyfélszolgálati üzenetek témakörébe tartozó szavak, mondatok tartalmát kell az adatbázisnak jól lefedni [3]. Fontos továbbá a prozódiai modellezés is (a gyakran előforduló kifejezések minden lehetséges prozódiai helyzetben előforduljanak). A cél az, hogy a szintézis során minél hosszabb beszédrészleteket találjunk meg a szintetizálandó mondat szövegéből az adatbázis szövegében és hullámformájában. A szintézishez elkészített adatbázis 3747 mondatot tartalmaz (összesen 69057 szó) szöveges és hangzó formában. Ez az állomány az elmúlt években készült, ugyanazon bemondó által felolvasott prompt-üzenetek hullámformáit tartalmazó fájlokból állt össze. Ebből készítettük el a szintézishez használható állományt, amely a megfelelő címkézéseket is tartalmazta. A címkézések a belső szinkronozást biztosítják. Az adatbázis a szövegen és a hullámformán felül tartalmazza a kettő összekapcsolását végző köztes ábrázolást, a szöveg fonemikus átiratát is. A címkéket és a fonemikus átiratot az adatbázisleíró fájl tartalmazza. A hullámformának, a fonemikus átiratnak és a címkéknek egymással szoros hangszinkronban kell lenni. Ez fontos követelmény. A szövegben nem lehetnek betűhibák, elütések, a hangban sem fordulhat elő, hogy a bemondó nem azt mondja, ami a felolvasásra készített szövegben van.

Az adatbázis elkészítésének lépései a következők:

(A1) Mondatszintű struktúra kialakítása. A szöveget és a hanganyagot mondatokra bontjuk, majd mondatonként hangszinkronba hozzuk őket egymással (a szövegben feloldjuk a rövidítéseket, a betűszavakat, a számokat stb.). Eltérés esetén (amikor a bemondó mást olvasott, mint ami a szövegben volt) a szöveget igazítjuk a hanghoz, mivel a hanganyag már nem módosítható. A számok és rövidítések feloldását előfeldolgozóval végezzük (Profivox [7] szövegfelolvasó átíró modulja) automatikus módon a manuális munka előtt. A szinkronba hozás részben manuális munkával történik, ennek eredménye, hogy a szövegben lévő betűhibák is javításra kerülnek. Ennek a munkafázisnak az eredménye, hogy olyan szövegforma áll elő, amelyben a rövidítések és számok fel vannak oldva. Ebből készül el a fonemikus átirat.

(A2) A fonemikus átirat elkészítése. A leírt szöveg nem definiálja egyértelműen a megvalósuló hangsorozatot. Figyelembe kell venni a hasonulásokat és más kiejtési sajátosságokat (hangkiesés, hangbetoldás). A fonemikus átírást a beszédfelismerőhöz tartózó átíróval készítjük (ez a felismerő jelöli ki később a hanghatárokat is).

(A3) Hullámforma előkészítés. Az adatbázisban minden mondatnak van egy hullámforma reprezentációja. A hullámformát címkékkel kell ellátni. Kétfajta alapjelölést alkalmazunk. Az egyik a hanghatár jelzése, vagyis a hang fonemikus szimbóluma és a hozzá tartozó hullámforma részlet összekapcsolása. A másik a szóhatár jelölése, vagyis a szövegben szereplő szó kezdőpontjának meghatározása. Ez a két jelölés a legfontosabb, mivel ezek alapján fogjuk kiválasztani a szintézis során az összekap-

csolandó hullámforma részleteket. Mindezeken felül címkézzük még a gerjesztési formákat (zöngés/zöngétlen), valamint a zöngés hangperiódusok kezdőpontjait (zöngejelzők). A zöngeperiódusok jelölése az alapfrekvencia pillanatnyi értékének a meghatározásához, illetve az alapfrekvencia-menet esetleges módosításához szükséges. Az alapfrekvencia pillanatnyi értékét az elemkiválasztás folyamán használjuk fel.

Az eszközrendszer a fenti címkézésekhez a következő. A zöngeperiódus-határok bejelöléséhez a Praat 4.6 fonetikai és beszéd-analizátor szoftverben implementált ablakfüggvénnyel korrigált autókorrelációs módszeren alapuló alapfrekvencia-detektálót használjuk [1]. A hanghatárok jelölése a beszédjel szintjén történik, azaz megadjuk, hogy hányadik mintán kezdődnek az egyes hangok. Ezt a feladatot egy a BME-TMIT-en kifejlesztett magyar nyelvű beszédfelismerő [2] segítségével oldjuk meg. A felismerőt kényszerített módban használjuk, ami azt jelenti, hogy a beszédet tartalmazó hangfájl mellett bemenetként megadjuk annak fonemikus változatát is, ami segít abban, hogy milyen hangsorozatokat keressen a felismerő. A felismerő 30 ms-os keretekkel és 10 ms-os kereteltolással dolgozik, azaz a hanghatárokat elvileg is csak 10 ms-os pontossággal határozza meg. A gyakorlatban a hanghatárjelölésekre 20 ms-os átlagos hiba a jellemző. A szóhatárok jelölését a beszédfelismerő által visszaadott, a beszédjelhez legjobban illeszkedő fonémasorozaton végezzük. A szavakon átívelő hangegybeolvadások (például: „ideig ködös”) miatt előfordulhat, hogy egy hang egyszerre két szóhoz is tartozik. Ennek kezelésére külön jelölést alkalmazunk a szavak kezdetére és végére. Így lehetővé válik a szavak közötti szóhatár-átfedést kezelése (például: „<idei<k>ödös>”, ahol a „<” jel a szó kezdetét, a „>” jel a szó végét jelöli). A szóhatárok jelölését teljesen automatikusan végezzük, olymódon, hogy a beszédfelismerő által visszaadott fonémasorozatot illesztjük a szintén a beszédfelismerő által előállított, (a szóhatároknál elágazó) összes lehetséges fonetikus átírást megadó irányított gráfhoz [4]. Az illesztést egy állapotgéppel végezzük, amely a fonémasorozat és a gráf alapján követi az elágazásokat és ennek megfelelően helyezi el a szóhatárt jelölő címkét.

(A4) Építőelemek. A szintetizátorban kétféle építőelem-típust definiáltunk, ezek a szó és a beszédhang. Az adatbázis belső címkézése illeszkedik a szintézis alapegységeihez. A szó szintű összeállítás során szavakból illetve szókapcsolatokból állítjuk össze a mondatot. Ez jó hangminőséget szolgáltat, ha elég nagy az adatbázis, és a keresési algoritmus is jól működik (figyelemmel kell lenni a prozódia helyes megvalósítására is). A szó szint továbbá gyors keresést tesz lehetővé. A beszédhang képviseli a tartalékot, vagyis azt az esetet, amikor az adatbázisban nem találjuk meg a szintetizálandó szövegrésznek megfelelő szót, szókapcsolatot. Ilyenkor hangokból, hangkapcsolatokból állítja elő az adott szót a rendszer, természetesen ennek a hangminősége gyengébb lesz, mint a szó szintű összeállításé.

(A5) Tesztelés, hibakeresés, utólagos hibajavítás emberi és gépi erővel. Az adatbázis fejlesztése során több lépcsős tesztelésre van szükség, mivel a jelölések és a hullámforma között sok esetben nincs szoros összhang. Az első ilyen szükséges teszt, amikor azt ellenőrizzük, hogy a hanghatárok közötti időtartamok mennyire jellemzőek a jelöléshez tartozó fonemikus hangra. A hanghatár-jelölés ellenőrzéséhez minden hangra egy hanghossz-eloszlás hisztogrammot készítettünk. Ennek segítségével meghatároztuk azokat a hangokat, amelyek hossza jelentősen eltért a velük azonos hangok átlagolt hosszaitól. Az ilyen hangokat tartalmazó mondatokat külön-külön manuálisan

megvizsgáltuk és javítottuk. A tesztelés következő fázisában a spektrális tartalom és az adott hang összevetését végeztük el automatikusan. Az így feltárt hibákból kiderült, hogy sok esetben a szövegben olyan rövidítések maradtak, amelyeket nem tudott megfelelően feloldani az előfeldolgozó program. Az elemzések során feltárt hibákat manuálisan javította egy fonetikai szakképzettségű informatikus. A munka az egész adatbázisra vonatkozóan 3 hónapot vett igénybe.

2.2 A vágási pontok meghatározása

A beszédszintetizátor optimális működése szempontjából fontos, hogy olyan ponton vágjuk el a hullámformát (vegyük ki az adatbázisból), amely a legkevesebb torzítással jár a későbbi összeillesztésnél. A döntést két tényező befolyásolja: milyen hangkapcsolódások vannak az adott ponton és, hogy milyen a prozódiai szerkezet. A fizikai vágási pont kialakításához ismerni kell a beszédhangok artikulációs és spektrális belső szerkezetét, valamint tisztában kell lenni a hangkapcsolódások megvalósulásakor létrejövő hangszerkezeti és spektrális módosulások fajtáival. A vágást akkor végezhetjük sikeresen, ha tudjuk, hogy a beszédhangoknak milyen az egymásra hatása, a belső akusztikai szerkezete, hol milyen változás zajlik le a hang frekvencia-, illetve intenzitásszerkezetében a folyamatos artikuláció következtében, melyek azok a hangrészek, amelyek esetleg egymással megegyeznek, illetve nagyon hasonlóak egymáshoz. Úgy kell kiválasztani a kivágandó elemet, hogy ne sértsük meg a spektrális folyamatosság elvét. Az optimális vágási pontok a következők: a hangsor minden olyan pontja, ahol gerjesztés váltás megy végbe (tiszta zöngés szakaszt tiszta zöngétlen követ és fordítva, itt ugyanis a jelben intenzitás minimum keletkezik), továbbá a hangok belsejében lévő néma fázisok, fojtott zönge szakaszok (ez a zár- és zár-rés hangok sajátja). Az optimális vágási pont kijelöléséhez 5-10 ms pontosságú helymeghatározásra, általában zöngeszinkron jelölésre van szükség. A hangsor összeállításánál ezek után a kivágott beszédrészletek egymáshoz való illesztését hangszerkezeti és artikulációs fonetikai szabályok alkalmazásával tehetjük torzításmentessé. Az illesztés akkor lesz sikeres, ha a beszédjelen nincs hallható akusztikai torzulás a beavatkozás után [6].

2.3 A prozódia modellezése

Az adatbázisban szereplő mondatok prozódiája adott. A szintetizálandó mondat prozódiáját meg kell jósolni. Az elemkiválasztó algoritmusnak figyelembe kell venni a jósolt prozódiát és annak megfelelő szót, szókapcsolatot kell keresni az adatbázisban. Az adatbázisban a prozódia modellezése bonyolult feladat, hiszen sok paraméter (hangsúly, ritmus, dallam, tempó stb.) jelölésére, kezelésére van szükség. A bemeneti mondat esetében pedig a prozódia jóslását kell elvégezni. Ez bonyolult nyelvi elemzéssel érhető el. Jelenleg egyik megoldásra sincsenek eszközök, ezért más utat kerestünk. Modellünk kialakítását egy mondatszerkezeti vizsgálatra épülő szópozíció meghatározásra alapoztuk. Másik kiindulási pontunk az volt, hogy a magyar kijelentő mondatokban a hangsúlyozás többnyire a frázis első tartalmas szaván van [5]. A modell lényege a következő. Az elemkiválasztás előtt a bemeneti szöveget prozódiai

egységek szerint tagoljuk. A prozódiai egység a szintetizátor jelenlegi megvalósításában egy írásjelekkel határolt, tagmondat-jellegű szövegrészt jelent. Minden egyes prozódiai egységet megcímkéztünk aszerint, hogy a mondaton belül milyen pozícióban van (Me első, Mk közbenső, Mu utolsó). Ugyanilyen címkézéssel láttuk el a prozódiai egységek szavait is (Sze első, Szk közbenső, Szu utolsó). Így egy kétszintű ábrázoláshoz kötöttük a prozódiát. Példaként lássunk egy prozódiai címkékkel ellátott mondatot.

✓Me,(Sze)A (Szk)3G (Szk)hálózat (Szk)használható (Szk)egyszerű (Szk)telefonbeszélgetések (Szu)lebonyolítására Me☛✓Mk1,(Sze)és (Szk)multimédiás (Szk)alkalmazásokra (Szu)is Mk1☛,✓Mk2,(Sze)akár (Szu)vonalkapcsolt Mk2☛, ✓Mu,(Sze)akár (Szk)GPRS (Szk)adatátvitel (Szu)formájában Mu☛.

Az elemkiválasztás folyamán megpróbálunk a bemeneti szövegben szereplő szavakhoz hasonló pozíciójú szavakat kiválasztani. Természetesen a pozíció jellegű információ nem határozza meg egyértelműen sem a hangsúlyokat, sem a hangidőtartamokat, azonban a percepciós tesztek szerint közelíti a természetes prozódiát. A módszer előnye az egyszerűsége és gyorsasága, hátránya, hogy vannak esetek, amikor nem biztosít megfelelő prozódiát. A kísérleti rendszer jelenlegi implementációja nem tartalmaz utólagos jelfeldolgozást a prozódia simítására.

2.4 Elemkiválasztás és összefűzés

A prozódia meghatározása után, a szintézis következő lépése az elemkiválasztás. Az elemkiválasztás alapelve, hogy a rendszer a bemeneti szöveget (elviekben) az összes lehetséges módon összerakja a beszédkorpusz elemeiből, és azok közül a legtermészetesebben hangzót választja ki. A természetesség automatikus megállapításához kétféle költséget vezettünk be: Az **egyezési költség** megadja, hogy egy adott elem mennyire felel meg a szintetizálandó beszédszakasznak. A jelenlegi megvalósításban a beszédszakaszt az annak betűsorozatként megadott szöveges tartalma, illetve a hangsorozatként megadott fonetikus átírása határozza meg. Ehhez járulnak még a szintetizálandó szövegből meghatározott prozódiai címkék. Az **összefűzési költség** azt adja meg, hogy a leendő szomszédos elemek hangilleszkedési szempontból mennyire folytonosan illeszkednének egymáshoz. Az elemkiválasztás folyamata azt a mondatot választja ki az összes lehetséges közül, amelyre az egyezési és összefűzési költségek összege a legkisebb. A rendszer a költségfüggvények alapján automatikusan osztályozza is a mondatok minőségét egy ötfokozatú skálán (jó minőségűnek tekinthető a 4-es osztályzat feletti). Az elemkiválasztás hierarchikusan történik. Először csak szószintű elemeket keres az elemkiválasztó. Ha a bemeneti szövegnek vannak olyan szavai, amit nem sikerült szóalapon megtalálni, akkor a hiányzó szavakat beszédhangokból rakjuk össze. A szószinten már megtalált elemeket csak abban az esetben próbálja meg a rendszer kisebb egységekből felépíteni, ha annak az egyezési költsége nagyobb, mint egy előre meghatározott érték. Az elemkiválasztást mondatonként végezzük. A keresés folyamán az adott mondatban szereplő egy-egy szóhoz, vagy hanghoz többféle lehetséges jelöltet is kiválasztunk. Egy-egy elemhez implementációs és hatékonysági okokból maximáltuk a lehetséges jelöltek számát nyolcvanban. Ha a kiválasztás folyamán egy adott elemhez tartozó jelöltek száma elérte a

nyolcvanat, akkor minden egyes további jelölt hozzávétele után a legmagasabb egyezési költségű elemet eldobjuk. Az összefűzés során elvileg minden elem esetében tetszőleges jelöltet kiválaszthatunk. Ebből következőleg a különböző előállítható lehetséges mondatok számát a jelöltek számának szorzata adja meg. Az optimális mondat kiválasztását a dinamikus programozáson alapuló Viterbi-algoritmus segítségével végezzük. Az algoritmus minden egyes lehetséges útra (mondatra) meghatározza az egyezési és összefűzési költségek összegét, és a minimális költségű utat (szavakat illetve szófüzéreket) választja ki.

Az **egyezési költség** kialakításának paraméterei

1. A jelöltet (szó vagy beszédhang) megelőző és követő fonéma egyezése az előírt célelemet megelőző és követő beszédhanggal. A legoptimálisabb keresési eredmény a teljes egyezés. Ehhez a legkisebb az egyezési költség értéke (nulla). Arra az esetre, ha nem biztosítható a teljes egyezés, definiáltunk egymással helyettesíthető hangkategóriákat is. Az azonos kategóriába eső hangok egyezési költsége is kicsi. Abban az esetben, ha egyik korábbi eset sem áll fenn az fennmaradó elemekből választ a rendszer. Ez adja a legnagyobb költséget. A fennmaradó elemek esetére az egyezési költséget egy költségmátrix definiálja.

2. Prozódiai egység mondaton belüli pozíciójának egyezése. Ezt csak szavak esetében vesszük figyelembe.

3. Prozódiai egységen belül előírt pozíciótól való eltérés. Ezt csak szavak esetében vesszük figyelembe.

Az **összefűzési költség** a következő paraméterek szerint alakul ki:

1. Ha a vizsgált két jelölt a beszédkorpuszban egymás után következett, akkor az összefűzési költség mindig 0. Ez a legoptimálisabb eset.

2. Ha a vizsgált két jelölt a beszédkorpuszban azonos mondatból származott, akkor kisebb összefűzési költséget rendelünk hozzá, mintha eltérő mondatokból származott volna.

3. Alapfrekvencia-menet folytonossági költsége, amit két elem összekapcsolásakor az első elem végső és a második elem kezdő zöngés hangjából számított átlagos alapfrekvencia eltéréséből arányosan származtatunk.

Az egyes költségek értékeit számos hangminta meghallgatása során, tapasztalati úton állítottuk be. Ezek további optimalizálása még jelentősen javíthatja az előállított beszéd minőségét.

3 A prompt generátor tesztelése, az optimális működés beállítása

A fejlesztés során többször hajtottunk végre meghallgatásos tesztet. A géppel generált hangüzeneteket 1-5 közötti MOS skálán (Mean Opinion Score – átlagos szubjektív osztályzat) osztályoztuk. Az ilyen belső tesztekben 3 fő vett részt (a fejlesztői gárdából). Ezen értékelések adatait felhasználtuk a szintetizátor belső költségfüggvényeinek javításához. A fejlesztés során háromhavonta szubjektív tesztelésen, szövegesen is értékeltük a hibákat, majd hibacsoportokat állítottunk fel a következők szerint: hangerő problémák; zaj hallatszik; a hangsúlyozás nem megfelelő; kevés a szünet a tagolási helyeken; túlságosan tagoltan beszél; túl gyorsan mondja a szöveget; adatbá-

zis hiba (hanghatár rossz helyen van, rossz a hangdefiníció). A hibacsoportok elemzése alapján folyamatosan optimalizáltuk a költségfüggvények paramétereit.

A rendszer stabilitása érdekében terheléses teszteket végeztünk. Vizsgáltuk a program futási idejét, a memóriahasználatot, a futási stabilitást. A hibák és felmerülő problémák folyamatos javításával elértük, hogy a rendszer stabilitása fokozatosan növekedett. Jelen állapotában megbízhatóan működik folyamatos felhasználói tesztelésre alkalmas.

A rendszerben a komponensek folyamatosan naplózzák az egyes eseményeket. Ilyen naplóbejegyzések pl.:

- A szintetizálandó prompt szövege.
- A szintetizálandó prompt szövegének fonetikus átirata.
- A szintetizálandó prompt szövegének megfelelő korpuszelemek, melyek közül a program válogat.
- A végleges korpuszelemek, melyből a prompt hangzó formája előáll.
- A program működésével kapcsolatos bejegyzések (indítás, leállítás, beállítások…).

A részletes naplófájlok használata a fejlesztési szakaszban és utána is megkönnyíti a hibák keresését, illetve azok okának kiderítését, ezáltal javításukat is.

A rendszer kész állapotú hangminőségének ellenőrzésére 120 prompt (312 mondat) automatikus generálását végeztük el. A költségfüggvények alapján automatikusan adott osztályzatok átlaga 4,12. A meghallgatások során 3 személy értékelte ugyanezeket a mondatokat, ítéleteik átlaga 4,02 volt. Ez azt mutatja, hogy a költségfüggvények alapján adott gépi értékelés jól közelíti a szintetizált mondat hangminőségének jellemzését.

4 Összegzés

A beszédszintézissel szemben támasztott új követelmények eredménye a bemutatott beszédtechnológiai megoldás. A kutatás-fejlesztés újszerűsége abban áll, hogy a szintetikus beszédet teljesen természetes hangzási minőségben és egy adott bemondó hangszínén kell előállítani adott témájú, ügyfélszolgálati üzenetek szövegeiből kiindulva. A célkitűzés megvalósításához a korpusz alapú, elemkiválasztásos beszédszintetizálási módszert választottuk. Magyarországon ez az első ilyen beszédelőállító rendszer. Az eredmények azt mutatják, hogy ilyen technológiával, optimális esetben el lehet érni a követelményként megadott hangminőséget. A rendszer legérzékenyebb pontja az elemkiválasztás költségfüggvényének a behangolása. Valószínűsítjük, hogy az optimálisabb működéshez több-szintű költségfüggvény kombinációt kell majd alkalmazni.

Bibliográfia

1. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, IFA Proceedings 17: 97–110
2. Mihajlik, P., Révész, T., Tatai, P.:Phonetic Transcription in Automatic Speech Recognition, Acta Linguistica Hungarica, Vol. 49 (3-4), (2002) 407–425
3. Nagy, A., Pesti, P., Németh, G., Bőhm, T.: Korpusz-alapú beszédszintézis rendszerek megvalósítási kérdései, Híradástechnika, (2005/1) 18–24
4. Németh, G. – Olaszy, G. – Fék, M.: Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei. Beszédkutatás - 2006. MTA Nyelvtudományi Intézet, (2006) 183–196
5. Olaszy, G.: A Korpusz alapú beszédszintézis nyelvi, fonetikai kérdései, Híradástechnika (2006/3) 43–50
6. Olaszy, G.: Az artikuláció akusztikai vetülete, a hangsebészet elmélete és gyakorlata. Kif-LAF 2003. Szerk.: Hunyadi László. Debreceni Egyetem, (2003) 241–254
7. Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Gordos, G.: PROFIVOX – A Hungarian Professional TTS System for Telecommunications Applications, International Journal of Speech Technology, Vol. 3, Numbers 3/4, (2000) 201–216

Ezt a kutatás-fejlesztést az NKFP 2. programja (szerződésszám: 2/034/2004) támogatta.

Fonetikai algoritmus a hanghatárok gépi meghatározásának javítására nagyméretű beszédadatbázisokban

Olaszy Gábor

BME Távközlési és Médiainformatikai Tanszék
olaszy@tmit.bme.hu

Kivonat: A beszédtechnológiai kutatásokhoz és gyakorlati alkalmazásokhoz egyre nagyobb méretű beszédadatbázisokat terveznek. Ezek egyik fajtája, amikor előre meghatározott szöveg felolvasásával hozzák létre a több órányi beszédanyagot. Ilyen adatbázis például a BME TMIT időjárás jelentéseket tartalmazó beszédkorpusza, amelyik 5400 mondatot tartalmaz. Ahhoz, hogy gépi módszerekkel lehessen ezt a hanganyagot a későbbiekben feldolgozni (szavakat, hangkapcsolatokat keresni stb.), a szöveget át kell írni hangszintű szimbólumsorozattá, majd jelölni kell a hanghullámon a hangokat, azok határát, a szavak kezdetét, a szüneteket, valamint szinkronba kell hozni a szöveget a hangzó formával automatikus gépi felismerés segítségével. Az így jelölt hangok és hanghatárok csak mintegy 95%-os pontossággal adnak kellően helyes eredményt. Ez azonban nem elegendő a jó minőségű, korpusz alapú beszédszintézishez. Ebben a tanulmányban egy olyan utófeldolgozó algoritmust ismertetünk, amelyikkel növelni lehet a pontosságot, ezzel a szintézis minőségét.

1 Bevezetés

A beszédtechnológiai kutatásokhoz és gyakorlati alkalmazásokhoz egyre nagyobb méretű beszédadatbázisokat terveznek. Ezek egyik fajtája, amikor előre meghatározott szöveg felolvasásával hozzák létre a több órányi beszédanyagot. Ilyen adatbázis például a BME TMIT időjárás jelentéseket tartalmazó beszédkorpusza [2], amelyik 5400 mondatot (95678 szó) tartalmaz. Ahhoz, hogy gépi módszerekkel lehessen ezt a hanganyagot a későbbiekben feldolgozni (szavakat, hangkapcsolatokat keresni stb.), a szöveget át kell írni hangszintű szimbólumsorozattá, majd jelölni kell a hanghullámon a hangokat, azok határát, a szavak határát, valamint a szüneteket. Ezeknek a jelöléseknek elméletileg szinkronban kell lenni az eredeti szöveggel. Ennek elérése gépi feldolgozással csak meghatározott hibaszázalékkal lehetséges. Ebben a tanulmányban a hanghatárok gépi kijelölésével kapcsolatos problémákkal foglalkozunk. Célunk az elméletileg lehetséges legoptimálisabb hanghatár-jelölés közelítése.

A gépi hanghatár-kijelölés egyik módszere a beszédfelismerésre kidolgozott technikát is alkalmazza, amikor a hangok sorozatát alapul véve (fonemikus átirat) irányított felismerést végeznek és bejelölik a hanghullámon a hanghatárokat, valamint a

szüneteket [3]. A módszer helyes működésének kritériuma, hogy a fonemikus átirat és a hanghullám között szoros egyezőség legyen, aminek a teljesülését feltételezzük, mint kiindulási alapot. A felismerés eredménye a mondat kezdeti pontjának a megjelölés, továbbá pedig annyi hanghatár bejelölése a hangsorba, ahány hang van a fonemikus átiratban. A BME TMIT beszédkutató laboratóriumában is ilyen módszert alkalmaznak [3]. A beszédfelismerő alapvetően HMM alapú algoritmus. A jelölés eredménye a körülbelüli hanghatár minden megadott hangra. Méréseink szerint az így jelölt hanghatár a humán jelöléshez viszonyítva akár 20ms-os eltérést is mutathat. Ezekre a jelölésekre támaszkodnak az adatbázist különböző célokra felhasználó további algoritmusok (például a korpusz alapú beszédszintézis válogató modulja). A beszédfelismerés gyakorlati alkalmazásainál ez a pontosság kielégítő, azonban a beszédszintézises alkalmazás során kiderült, hogy a hanghatár helyének jelölését tovább kell finomítani. Például női hang esetén a 20ms-nyi idő körülbelül 4 zönge periódust fedhet le. Sok esetben ezek a hanghatár-elcsúszások fonetikailag irreleváns helyzeteket hoznak létre (például egy zöngétlen hangban zöngés periódusok lesznek a jelölés szerint), illetve hangelcsúszás is előfordulhat, ami például egy beszédszintetizátor hangzásában zavaró. Mindezen hibák kiküszöbölésére felmerült a gondolat, hogy egy utófeldolgozó, finomító algoritmusra van szükség.

2 Anyag és módszer

A gépi címkézéshez és fonemikus hangjelöléshez HMM alapú beszédfelismerőt [3] alkalmaztunk, irányított felismerési technikával. Ennek lényege, hogy segítjük a felismerőt a döntéshozatalban a fonetikai átirat szerinti hangsor megadásával. A beszédhullámban tehát annak kezdetét jelöli meg az algoritmus, majd pontosan annyi hangot jelöl, ahány hang van az fonetikai átirat szerint a mondatban. Azt azonban a program nem tudja garantálni, hogy minden jelöléshez korrekt hang tartozik. Ez a kényszer – ritkán – helytelen hangazonosításhoz is vezethet. Ha például nem talál olyan akusztikai tartalomnak megfelelő részt a hullámformában, mint amilyen hang az átiratban szerepel, akkor kényszerből betesz egy hanghatárt a saját döntése alapján, csak azért, hogy a hanghatárok száma megegyezzen az előre megadott számmal. A szintézis során csak a bejelölt hanghatár jelekre támaszkodunk a szóválasztásnál, azok alapján veszünk ki a hullámforma részeket a hanghullámból. Rossz helyzetű hanghatár rossz hangot eredményez.

A szöveg fonemikus átírását a beszédfelismerő végezte (a hangjelölések az ábrákon a számítógépes program belső kódjai szerintiek, ezeket az adott helyen magyarázzuk).

A hibafeltáráshoz a teljes szöveg- és hangkorpusz 5400 mondatából válogattunk ki 110 címkékkel ellátott mondatot, és manuálisan elemeztük azokat. Egyrészről a hangjelölések és az akusztikai tartalom megegyezését vizsgáltuk, másrészről pedig ellenőriztük a gépileg bejelölt hanghatárok pozíciójának helyzetét a hanghullámhoz viszonyítva. A hanghatárok humán megállapítása is nehézségekkel teli, különböző pontossággal tudjuk meghozni a döntést [4]. Viszonylag pontos meghatározás lehetséges, ha gerjesztésváltás van a két hang kapcsolódási pontjában. A hangintenzitás viszonyok hirtelen változása is jó támpont (például magánhangzó és zöngés zárhang találkozá-

sánál). A legnehezebb a hanghatár kijelölése az olyan hangkapcsolódási pontokon, ahol a hangok természetéből adódóan nincs jelentős akusztikai szerkezeti változás (például magánhangzó kapcsolatokban). Korrekt manuális hanghatár-jelölésekre példákat találunk a http://fonetika.nytud.hu/cccc honlapon.

Elemzéseinkhez az eredeti mondat hangzó változatát és az azzal szinkronban lévő gépi hangjelöléseket használtuk. Az analízis során a Praat 4.6 fonetikai programmal [1] egymás alatt megjelenítettük a mondat rezgésképét, spektrogramját, valamint a hozzá elkészített hangjelöléseket és hanghatárokat. Hangonként elemeztük a jelölés és az akusztikai tartalom helyességét, illetve hibás voltát. Az elemzésekhez felhasználtuk az 1. táblázat szerinti hangosztályozást. Négy olyan jellemző paraméter szerint osztottuk fel a hangsor elemeit, amelyek a hibajavító algoritmusban egyértelműen elkülöníthetők és felhasználhatók. Ezek: a gerjesztés formája, a képzési üreg, a hang intenzitása és végül a hang szerkezeti felépítése. A gerjesztés formája a fonemikus jelölés alapján is és az akusztikai elemzésből is rendelkezésre áll (a zöngés periódusok meg vannak jelölve a hanghullámban). Az üreget a jelölt hang alapján definiáljuk. A hang intenzitását pontosan tudjuk mérni, ennek a kategorizáláshoz három szintet használunk (intenzív, közepes, gyenge). Az intenzitás-szinteket a hangok akusztikai felépítéséből ismert adatok alapján rendeljük hozzá a vizsgált elemhez. A szerkezeti felépítést a fonemikus jelölés alapján két kategória szerint osztottuk meg: egyszerű és összetett. A beszédhangokat besoroltuk ezekbe a paraméter-csoportokba és 11 osztályt különböztettünk meg a kategorizáláshoz. Minden osztályon belül ugyanazok a paraméterek jellemzik a beszédhangokat azok artikulációs meghatározásától függetlenül.

1. Táblázat: a hangsor elemeinek kategorizálása 11 hangosztály, 4 paramétercsoport és ezeken belül összesen 10 paraméter szerint. A hangokat a betűjelükkel jelöltük. V=magánhangzó, h*=zöngés h hang.

	OSZTÁLY	1	2	3	4	5	6	7	8	9	10	11
	HANG	V	b d g gy	ny	p t k ty	m n	j h* v l	h f	z zs	sz s	c cs	r
PARAM. CSOPORT	PARAMÉTER											
1	zöngés	x	x	x		x	x					x
	zöngétlen				x			x		x	x	
	vegyes								x			
2	nazális			x		x						
	orális	x	x		x		x	x	x	x	x	x
3	intenzív	x										
	közepes			x		x	x		x	x	x	x
	gyenge		x		x			x				
4	egyszerű szerk.	x				x	x	x	x	x		
.	összetett szerk.		x	x	x						x	x

3. Az analízis eredményei

Az elemzés során 4300 hanghatárt ellenőriztünk, ezekből 232 esetben kellett módosítást elvégezni. A talált hibákat két fő kategória szerint osztályoztuk: helytelen

hangjelölés, vagyis a jelölt hang teljes egészében nem felel meg a hangsorbeli akusztikai tartalomnak (H1). Ebbe a kategóriába tartozik a beszédhangokon kívül a szünet (sil) helyének vizsgálata is. A második kategóriába (H2) tartoznak azok a hibák, amikor a jelölt hanghatár nem a tényleges kapcsolódási ponton van, hanem attól lényegesen (több mint 10ms-mal) eltér. Ez utóbbi esetben a kapcsolódó hangok mindegyikéből szerepel egy-egy rész a két jelölt hanghatár közötti hangsorrészben.

A (H1) hiba kétféle okból keletkezhet: egyszerű szöveghibából, illetve beszédfelismerési tévesztésből. Szöveghiba, amikor például a felolvasandó szövegből egy betű kimaradt (1. ábra). A bemondó az ilyen betűhibákat nem érzékeli, korrektül olvassa fel a szöveget. Ha a hibát nem veszi észre az első ellenőrzést végző személy sem, akkor a hullámforma és a szöveg fonemikus átírása között nem lesz hangszinkron egyezés. A hullámformában eggyel több a hang lesz, mint a fonemikus átiratban. Ilyen esetben a kényszerített beszédfelismerési algoritmusnak el kell dönteni, hogy a hangsor mely pontján „spórol meg” egy hangot.

Példa: *Az eleinte jellemző lehűlést az év utolsó napján a magasban* ***melegedé váltja fel****.*

A példából látható, hogy a felismerő úgy döntött, hogy a réshangot szünetként jelöli. Ilyen hiba detektálása esetén manuálisan ki kell javítani az adott mondat szöveges és fonemikus formáját. A szinkronitás így helyreáll. Egyébként, ha a program az így jelölt szünetet választja ki a szintézis során, akkor egy erős réshang fog hallatszani a szünet helyett. A szünet-jelzést a beszédfelismerő olyan helyeken is beiktathatja, amikor a beszélő glottalizál a hangkapcsolódás folyamán (2. ábra).

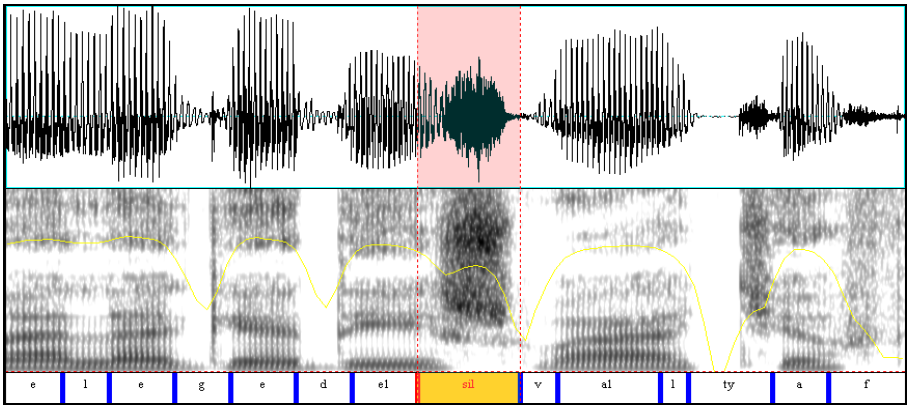


Fig 1. A fenti példamondat végén a *melegedés* szóban a réshang helyén szünet (sil) jel szerepel (szürke mező), mivel az eredeti szövegből kimaradt az s betű. (e1 = é; a1 = á). Az ábrán példát látunk a (2) hibakategóriára is, mivel a szünet hanghatára belelóg az előző magánhangzó végébe

Ezek valójában nem igazi szünetek, ezért nem szabad őket szünet funkcióra felhasználni (tehát ezeket is jelöléssel kell ellátni).

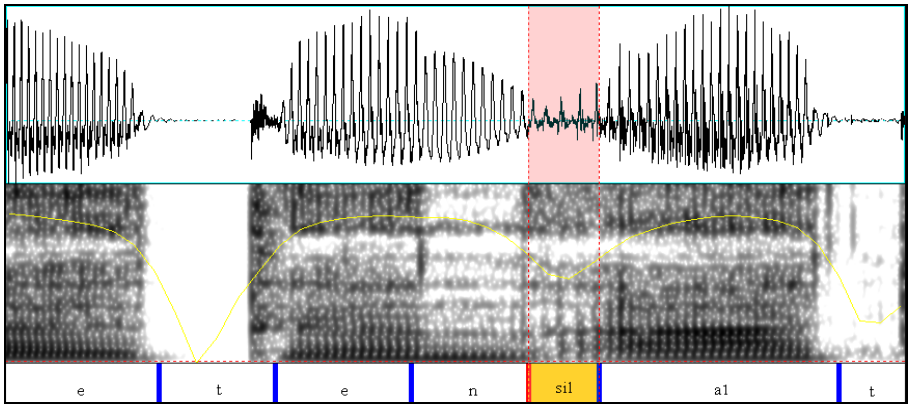


Fig. 2 A *Keleten átmeneti*... kezdetű mondat szóhatárán a bemondó glottalizált. Ezt a szakasz (szürke rész) a beszédfelismerő szünetnek jelölte (a1=á)

A (H1) típusú hibát okozhatja még bizonyos hangkapcsolódásokra jellemző hangszerkezeti változás. Erre példa az *ismét* szó réshang-nazális hang kapcsolatának hibás feldolgozása (3. ábra).

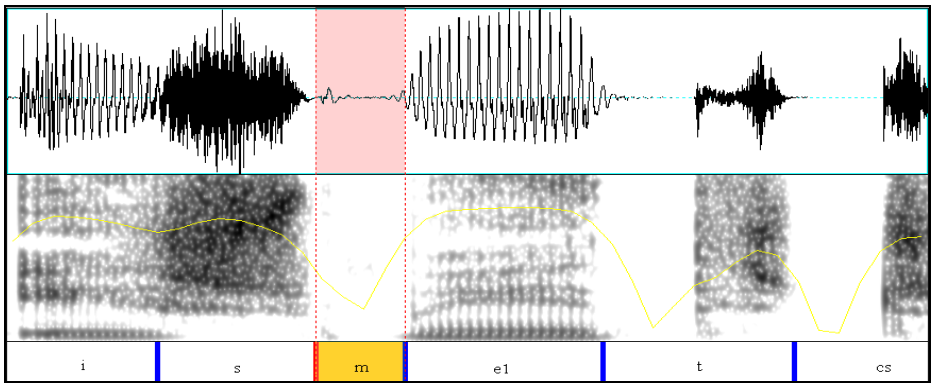


Fig 3. A koartikulációs néma fázis megzavarja a felismerést, ezért a gép erre a szakaszra sorolja be a teljes nazális hangot (szürke jelölés)

A felismerő a réshang végén kialakuló koartikulációs néma fázishoz [5] azonosította a nazális mássalhangzót, így ez a hang – a jelölés szerint – a réshang belsejébe került. A hiba következménye az is, hogy a hangsor szerinti nazális hang és az őt követő magánhangzó egyetlen hangként lesz jelölve. További példa a (H1) hibatípusra, amikor a beszélő egy svá töltelék elemet is ejt egy hang végén (4. ábra). Ez utóbbi a színészi ejtésnél gyakori (*csütörtökönə, pénteken*). Ekkor a felismerő a svára is tehet szünet jelölést (mást nem tehet rá, mivel a hangok száma kötött). Ezt a szünetet sem szabad felhasználni, mivel egy zöngés hang szól alatta.

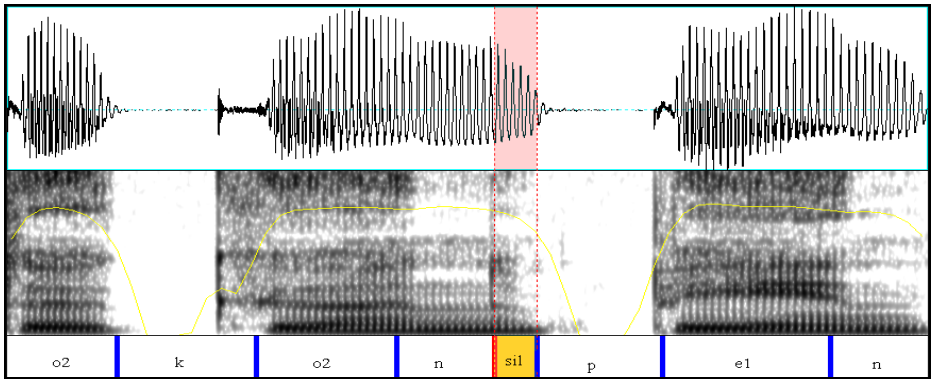


Fig 4. A nazális hang végén ejtett svá elemet a felismerő szünetnek jelölte a példában (szürke sáv), (o2=ö, e1=é)

A (H2) hibatípus a hanghatár eltolódását jelenti a hangon belül, illetve kívül. Ilyen hiba mind zöngés-, illetve zöngétlen hangok kapcsolódásánál, mind pedig zöngétlenek és zöngések találkozásánál előfordul. Nézzünk néhány példát. Az 5. ábra több hanghatár elcsúszását mutatja be zöngés hangok között.

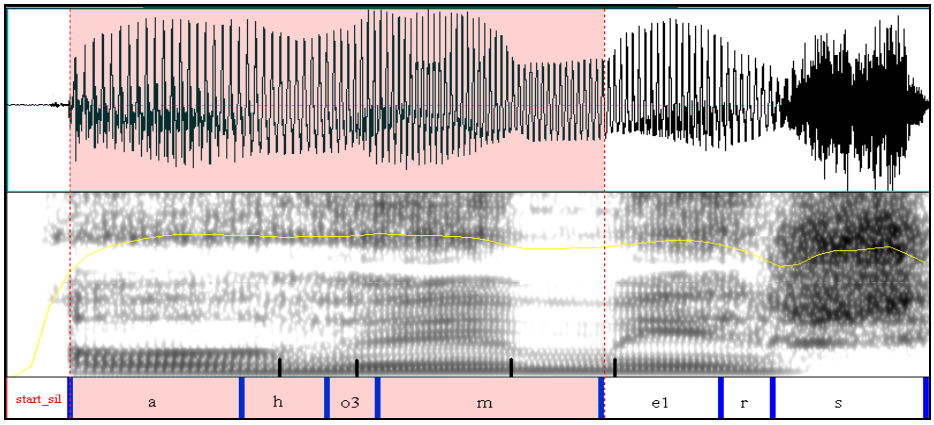


Fig 5. Példa a hanghatárok elcsúszására a *...hőmérséklet...* szó négy első hangjában. A helyes hanghatárokat a fekete jelölésekkel érzékeltetjük (o3=ő, e1=é)

A zöngés hangokban tapasztalt hangelcsúszást algoritmikusan nehéz kijavítani, detektálni azonban egyes esetekben lehet (lásd a példában is, ahol az első hosszú magánhangzó időtartama feltűnően rövid, mindössze 32 ms, így a hangidőtartamok ellenőrzésével felismerhető a hiba). Automatikusan lehet azonban javítani a zöngés-zöngétlen hangok találkozásánál tapasztalt hanghatár elcsúszást. A 6. ábrán egy tipikus példát mutatunk be erre a hibára is.

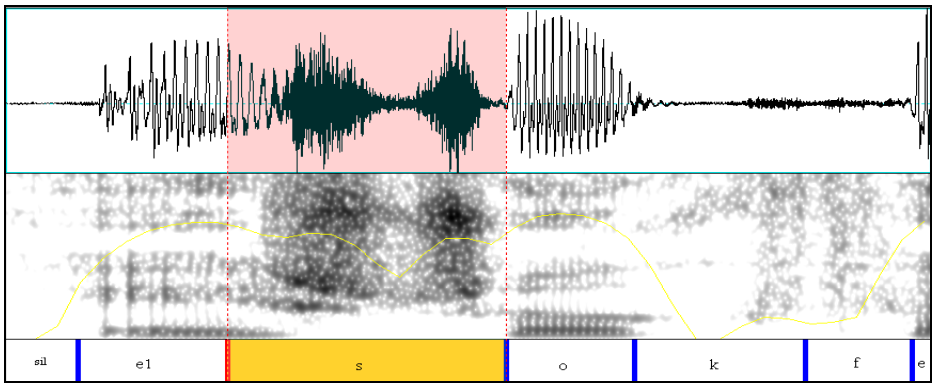


Fig 6. Hibás hanghatár jelölés az ...***és*** *sokfelé*... hangsor első két hangján

4 A fonetikai javító algoritmus

A javító algoritmus nem vizsgálja a spektrális részleteket, más paraméterek alapján végzi az elemzéseket. Az algoritmus leírásában a következő elnevezéseket használjuk:

fonemikus hang, vagyis a gépi hangátíró által a szövegből kikövetkeztetett hang;

jelölt hang, vagyis a beszédfelismerő által meghatározott két hanghatár között elhelyezkedő kikövetkeztetett hang;

beszédhang, vagyis a hanghullámban lévő, ugyanezen pozíciójú hang.

Hibátlan jelölés esetén a három elemnek szinkronban kell lennie az időtengelyen. Az algoritmusnak három feladata van: (F1) hangazonosítás, (F2) annak eldöntése, hogy a jelölt hanghatár jó helyen van-e, továbbá (F3) annak jelzése, ha egy hang a hangkörnyezetéből fakadóan annyira megváltoztatja a szerkezetét, hogy önálló hangként nem használható fel a szintézis során. Az algoritmus a vizsgálatot külön lépésekben végzi el. Mindhárom esetben a mondat elejétől indulva, a fonemikus hangszimbólumokra támaszkodva megvizsgálja az adott hangkapcsolathoz tartozó jelölt hanghatár helyét és összehasonlítsa a fonetikailag optimálisnak tartott beszédhanggal (a szünet jelzés is hangsorépítő elem, tehát erre is ugyanolyan döntéseket hozunk, mint a beszédhangokra). Az F1 vizsgálatnál döntést hoz arról, hogy egyáltalán az a beszédhang van-e a hangsorban, mint ami a fonetikus átírásban szerepel. Az F2 vizsgálatnál az a kérdés, hogy a hanghatár pozíciója helyes-e. Ha nem tartja helyesnek, akkor döntést hoz arról is, hogy automatikusan javítható-e vagy nem, majd ha javítható, akkor kijavítja. Az F3 vizsgálatnál azt dönti el, hogy a beszédhang felhasználható-e a szintézis során önálló hangként. Az algoritmus a következő döntések egyikét hozza:

(D1) a beszédhang és a jelölt hang megegyezik a fonemikus hanggal (jó jelölés);

(D2) a beszédhang jelölése rossz, javítása manuális vizsgálatot kíván;

(D3) a beszédhang jelölése jó, és a bal oldali hanghatár is jó helyen van;

(D4) a beszédhang fonemikus jelölése jó, de a bal oldali hanghatár el van csúszva, javítása automatikusan megtörténhet;

(D5) a beszédhang jelölése jó, de a bal oldali hanghatár el van csúszva, javítása csak manuálisan javasolt;

(D6) a beszédhang jelölése jó, de a hangot önálló elemként nem szabad használni (ezért ezt jelölni kell).
A hibajavítási folyamat várható eredménye az, hogy a pontosabb hanghatár eltünteti a szintetizált hangsorokban korábban észlelt hangzási hibák nagy részét.

Az elemzéshez egyrészről az 1. táblázat paramétereit használjuk, másrészről figyelembe vesszük a hangokra és a hangkapcsolódásokra jellemző szerkezeti jellemzőket is (a hang várható időtartama, a zárszakasz esetleges rövidülése, svá képződés lehetősége stb.). A vizsgálati paramétereket és a hozzájuk tartozó információkat a 2. táblázatban mutatjuk be. A javító algoritmus először a hangazonosítást (F1) végzi el. A döntéshez az 1. és a 2. táblázat paramétereit használjuk.
A vizsgálati lépések az (F1) típusú elemzéshez és javításhoz:
1. A hang hangsorkezdő (hangsor belseji) pozícióban van.
2. A hang fonemikus azonosítása és besorolása az 1. táblázat 11 osztályába, illetve a szünet osztályba.
3. Hosszú magánhangzó esetén a jelölt időtartam mérése.
4. Összevetés a jellemző időtartammal (2. táblázat 14. paraméter)
5. Ha az időtartam nincs a jellemző tartományban, a D2 hibajelzés elhelyezése manuális vizsgálathoz.
6. Egyezés esetén a további paraméterek vizsgálata (1. táblázat függőleges oszlop)
7. Eltérés esetén hibajelzés elhelyezése.
8. Egyezés esetén tovább lépés a következő hangra.
9. Ez szünet, vagy hangsorvégi szünet?
10. Ha nem a 2. ponttól új vizsgálat indul.

Vegyük példának az 1. ábrán látható hibás hangjelölés vizsgálatát. A hangjelölés szünetet (sil) állapít meg, ezért az algoritmusunk a 2. táblázat 11, 12, 13 paraméterét vizsgálja. Az időtartam 110 ms, ez megfelelő, beleesik a megadott sávba, akár levegővételi szünet is lehetne. A 12. paraméter vizsgálatánál kiderül, hogy ebben a szünetben az intenzitásszint sokkal magasabb, mint a megadott jellemző szint, ez azt sugallja, hogy a szünet jelölés mögötti akusztikai tartalom nem szünet, hanem valamilyen beszédhang. A 13. paraméter vizsgálata ezt még jobban alátámasztja, mivel a vizsgált szünetben találunk zöngésségi periódusjelzéseket is, ez is hiba. Ennek megfelelően a D2-es hibajelet kell elhelyezni. A 2. ábrán bemutatott szünet elem vizsgálata már a 11. paraméterrel való összehasonlításnál sejteti, hogy nem valódi szünetről van szó, mivel a jelölt határok közötti időtartam mindössze 33 ms. Ez alatta van a megadott minimális értéknek. A 3. ábrán látható szünet jelzés ugyanezzel a méréssel mondható hibásnak, de a 12. és 13. paraméter vizsgálati eredménye is hibára utal. Ebben az esetben tehát háromszoros megerősítéssel mondhatjuk ki, hogy a jelzés hibás.

2. Táblázat: a beszédhangok és a jelölt hanghatárok vizsgálatánál használt paraméterek és azok jellemző adatai

	PARAMÉTER	JELLEMZÉS
1.	zöngés a hang	a Praat program zöngeszinkron jeleket helyez el a hangsor zöngés részein
2.	zöngétlen a hang	a Praat program nem helyez el zöngeszinkron jeleket a zöngétlen részeken
3.	vegyes gerjesztés	a fonemikus hangjelből döntjük el. (z, zs, dz, dzs)
4.	nazális	a fonemikus hangjelből döntjük el. (m, n, ny)
5.	Orális	a fonemikus hangjelből döntjük el. minden hang, kivéve a 4. sort
6.	intenzív	a hangintenzitás mérésével döntjük el. Ez a maximális szint
7.	közepesen intenzív	a hangintenzitás mérésével döntjük el. Jellemző: 6-10 dB-lel alacsonyabb, mint a 6. sor
8.	gyenge	a hangintenzitás mérésével döntjük el. Jellemző: 12-20 dB-lel alacsonyabb, mint a 6. sor
9.	összetett szerkezetű	a fonemikus hangjelből döntjük el. (p, t, ty, k, c, cs)
10.	egyszerű szerkezetű	a fonemikus hangjelből döntjük el. Minden hang, kivéve a 9. sort
11.	Szünet paraméter: időtartam	rövid szünet, felesleges szünet: min. 40 ms, max. 60 ms levegővételi szünet: a rövid szünetnél hosszabb, max. 200 ms
12.	Szünet paraméter: intenzitás szint	a szünet elvárható intenzitás szintje a 6. sor maximális értékétől -20 db-nyi a jellemző. Amennyiben nem így van, akkor a jelölt szünet intenzitás szerkezetét vizsgálni kell. Hibának számít, ha a szünetben erős intenzitás ingadozás van (glottalizációra utal), illetve ha magas az intenzitás.
13.	Szünet paraméter: zöngétlen gerjesztés	a szünetben a gerjesztés nem lehet zöngés
14.	jellemző hangidőtartam	hosszú mgh. esetén hangsor belsejében az elfogadható érték min. 70 ms, max 120 ms között van
15.	rövidre módosult zárszakasz a zárhangban	a hangkapcsolat: mb, md, nc, ncs, nd, ng, nk a zárhang jellemző hossza 20-30 ms.
16.	koartikulációs néma fázis	CC kapcsolatoknál, a hangkapcsolat: s,sz,c,cs + m,n,ny A hangkapcsolat első hangjának végén jön létre a koartikulációs néma fázis

A vizsgálat második szakaszában az (F2) javító algoritmus lép működésbe és a hangok bal oldali hanghatárának estleges elcsúszását vizsgálja. Ezek a hibák többnyire automatikusan javíthatók is. Ennél a lépésnél feltételezzük, hogy az (F1) elemzéssel megtalált hibákat már kijavítottuk, minden hang jelölése megfelel a hangsor adott hangjának.

Az algoritmus vizsgálati lépései a következők:

1. A hang hangsorkezdő (hangsor belseji) pozícióban van.

2. A jelölt hang fonemikus azonosítása és besorolása az 1. táblázat 12 osztályába, illetve a szünet osztályba.

3. A jelölt hangot balról határoló szomszédos hang azonosítása és besorolása az 1. táblázat 12 osztályába (bal szomszéd). Ilyen elemnek tekintjük a szünetet is.

4. A jelölt hang bal oldali hanghatárának a vizsgálata és hiba esetén javítása

4.1 Ha zöngés beszédhangról van szó, akkor összevetjük a hanghatár helyét a periódusjelzőkkel (2. táblázat 1. paraméter), valamint a bal szomszéd gerjesztési besorolásával.

4.1.1 Ha a bal szomszéd zöngés hang, akkor a hanghatárt elfogadjuk jónak (D3), tovább lépünk jobbra a következő hangra. A vizsgálat újból kezdődik a 2. ponttól.

4.1.2 Ha a bal szomszéd zöngétlen gerjesztésű és a vizsgált hanghatár zöngétlennek jelzett hangrészen van, akkor ez azt jelenti, hogy a zöngés beszédhang bal oldali jelzett hanghatára kicsúszott a hangból és tartalmaz egy részt az előző zöngétlen hangból (esetleg szünetből) is.

JAVÍTÁS- A hanghatárt jobbra mozgatjuk addig, amíg zöngés periódusjelzést nem találunk. A hibát ezzel kijavítottuk (D4), léphetünk a következő hangra.

4.1.3 Ha a bal szomszéd zöngétlen gerjesztésű és a vizsgált hanghatár balról az első zöngésnek jelzett hangperióduson, vagy annak 10ms-os körzetében van, akkor a jelölés nem hibás (D3), léphetünk a következő hangra.

4.1.4 Ha a bal szomszéd zöngétlen gerjesztésű és a vizsgált hanghatár balról nem az első zöngésnek jelzett hangperióduson van, hanem attól jobbra, akkor a hibás a jelölés, mivel a hanghatár nem a zöngés hang kezdetén van, hanem annak a belsejében.

JAVÍTÁS- A hanghatárt balra mozgatjuk az első zönge periódus kezdetéig (D4). A hibát kijavítottuk, léphetünk a következő hangra.

4.2 Ha zöngétlen beszédhangról van szó (2. táblázat 2. paraméter), akkor összevetjük a jelölt hanghatár helyét az időtengelyen a periódusjelzőkkel, valamint a bal szomszéd gerjesztési besorolásával.

4.2.1 Ha a bal szomszéd zöngétlen beszédhang és a vizsgált hanghatár zöngétlen részen van, akkor tovább lépünk jobbra a következő hangra, a hanghatárt elfogadjuk jónak (D3).

4.2.2 Ha bal szomszéd zöngés hang és a vizsgált hanghatár zöngés területre esik a hanghatár, akkor ez azt jelenti, hogy a zöngétlen beszédhang bal oldali jelzett hanghatára kicsúszott a hangból és tartalmaz egy részt az előző zöngés hangból.

JAVÍTÁS- A hanghatárt jobbra mozgatjuk, annyira, hogy a zöngés-zöngétlen váltóponttól balra 10ms-ra kerüljön (azért nem tesszük pontosan a váltópontra, mert a zöngésből zöngétlenbe való váltás során a gerjesztések mintegy 10ms-nyi részen átfedik egymást, tehát az előző zöngés hang egyre csökkenő nagyságú periódusaira szuperponálódik már az induló zörej). Ezzel a hibát kijavítottuk (D4). Léphetünk a következő hangra.

4.2.3 Ha bal szomszéd zöngés hang és a vizsgált hanghatár zöngétlen területre esik, de 10ms-nál nem messzebb a bal oldali utolsó zöngés periódus jelzőtől, akkor a hanghatár jó (D3), léphetünk a következő hangra.

4.2.4 Ha bal szomszéd zöngés hang és a vizsgált hanghatár zöngétlen területre esik, de 10ms-nál messzebb (jobbra) a bal oldali utolsó zöngés periódus jelzőtől, akkor a hanghatár jelölése hibás, az a zöngétlen hang belsejében van.

JAVÍTÁS- A hanghatárt balra kell mozgatni az bal oldali zöngés hang utolsó zöngés periódus jelzésére (D4). Ezzel helyére toltuk a hanghatárt, léphetünk a következő hangra.

A hanghatárt vizsgáló és hibajavító algoritmus esetenként az 1. táblázatban 11 osztályának más paramétereit is figyelembe veszi, azonban erre itt részleteiben nem tudunk kitérni. A fenti algoritmus működését egy példán is bemutatjuk. A 6. ábrán látható hangsorrészben a réshang vizsgálatakor a következőket állapítja meg az algoritmus. A hang zöngétlen gerjesztésű (4.2), a bal szomszédja zöngés (3). A zöngétlen hang bal oldali hanghatára a 4.2.2 pont szerint zöngés területre esik, ez hibás, javítást kell végezni.

Az (F3) jelzésű vizsgálat során a módosult szerkezetű hangokat keressük és jelöljük meg, hogy a későbbi szintézis során a válogató modul ezeket ne használja fel önálló hangként. Ezzel a hibás hangzásokat kerüljük el. Ennél az elemzésnél feltételezzük, hogy az (F1) és (F2) elemzések hibáit már kijavítottuk. Az elemzéshez a 2. táblázat 15. és 16. sorát is felhasználjuk. Amennyiben a vizsgált hang és baloldali szomszédja a 15. sorban definiált hangkapcsolatokba tartozik, akkor a vizsgált hangra az „önállóan ne használd” jelölést tesszük. Amennyiben a vizsgált hang és baloldali szomszédja a 16. sorban definiált hangkapcsolatokba tartozik, akkor a bal szomszédra az „önállóan ne használd” jelölést tesszük.

5 Összefoglalás

A nagyméretű beszédadatbázisok címkézése csak gépi megoldásokkal végezhető el, ami sok esetben téves. A bemutatott hibadetektáló és javító algoritmus fonetikai és akusztikai jellemzők ötvözéséből kialakított kritérium rendszert használ fel a hangsorba már gépi úton bejelölt hangok és hanghatárok ellenőrzésére és javítására. A manuális vizsgálati eredmények szerint a gépi hang- és hanghatár-jelölés 95%-ban ad korrekt eredményt. A meghallgatásos vizsgálatok szerint ez a magasnak tűnő szám nem elégséges, mert a szintézis során számos esetben a hibás hangazonosítás, illetve a hibás hanghatár miatt torzulások kerülnek a szintetizált hullámformába, ami rontja a hangminőséget. Reméljük, hogy az ismertetett algoritmussal gépi úton lehet majd az eddigi 95%-os hatásfokot egy-két százalékkal emelni. A kombinált módszert más, a jövőben készítendő ilyen beszédadatbázisokra is lehet alkalmazni. További eredmény az is, hogy a helyes hanghatárok megnyitják az utat másfajta mérések (hangidőtartamok, időszerkezeti elemek, szóhosszúságok stb.) tömeges vizsgálatához is.

Bibliográfia

1. Boersma P., Weenink D.: Doing Phonetics by Computer. [Computer software], www.praat.org
2. Fék, M., Pesti, P., Németh, G., Zainkó, Cs., Olaszy, G.: Corpus-Based Unit Selection TTS for Hungarian. In: Sojka P., Kopcek I., Pala K. (eds.) Text, Speech and Dialogue TDS 2006, Springer, Brno (2006). 367–374
3. Mihajlik, P., Révész, T., Tatai, P.:Phonetic Transcription in Automatic Speech Recognition, Acta Linguistica Hungarica, Vol. 49 (3-4), (2002) 407–425
4. Olaszy Gábor: Hangidőtartamok és időszerkezeti elemek a magyar beszédben. Nyelvtudományi Értekezések 155. Akadémiai Kiadó. (2006)
5. Olaszy Gábor: Mássalhangzó-kapcsolódások a magyar beszédben. Tinta Könyvkiadó. (2007)

Ezt a kutatást az NKFP 2. programja (szerződésszám: 2/034/2004) támogatta.

Számítógépes összehasonlító szövegelemzés ügyfélszolgálati tájékoztatók legfontosabb prozódiai elemeinek a meghatározására

Abari Kálmán[1], Tamm Anne[2], Gábor Kata[3], Olaszy Gábor[4]
[1] Debreceni Egyetem, Pszichológia Intézet és Matematikai és Számítástudományi Doktori Iskola
abarik@delfin.unideb.hu
[2] Észt Nyelvtudomány Intézet, Tallinn és Firenzei Egyetem, Finnugor Szektor,
anne.tamm@eki.ee, anne.tamm@unifi.it
[3] MTA Nyelvtudományi Intézet
gkata@nytud.hu
[4] BME Távközlési és Médiainformatikai Tanszék
olaszy@tmit.bme.hu

Kivonat: A szövegelemzéssel történő hangsúlykijelölés bonyolult feladat. Jelenleg nincs olyan elemző algoritmus, amelyik gépi úton képes a magyar mondatokban a hangsúlyok kijelölésére. Alapvető célunk, hogy az eddig elért és hozzáférhető elméleti nyelvészeti eredményeket, valamint kész mondatelemző algoritmusokat egyetlen, jól körülhatárolható struktúrált számítógépes rendszerré fejlesszük tovább automatikus hangsúlykijelölési kísérletek végzése céljából. Ebben a tanulmányban munkánk végeredményét, egy futtatható programrendszert (elemző) mutatjuk be. Az elemző bemenetére a mondat szöveges formája kerül, majd a feldolgozás során a mondat minden szavát egy hangsúlycímkével látja el. A feldolgozást két szinten végezzük: (i) tagmondatokra bontás, (ii) a hangsúly kijelölése tagmondatonként. Öt hangsúlykategóriát definiáltunk az elemzéshez: (F)=erős hangsúly, (E)=kiemelt, (W)=normál, (N)=hangsúlytalan és (-)= erősen hangsúlytalan (redukált) címkék. Az elemző 12 modulból épül fel, melyben mindegyik modul azonos koncepcióra épül. Az elemzőt egy szűk témakört leíró szöveges állományra fejlesztettük. A hangsúlykijelölés hatásfoka összességében 85%.

1 Bevezetés

A szövegelemzéssel történő hangsúlykijelölés bonyolult feladat. Nincs is olyan elemző algoritmus a magyarra, amelyik gépi úton képes a feladat megoldására. Alapvető célunk, hogy az eddig elért és hozzáférhető elméleti nyelvészeti eredményeket egyetlen, jól körülhatárolható struktúrált számítógépes rendszerré fejlesszük további elemzési kísérletek végzése céljából. A munkánk köztes eredményéről korábban már beszámoltunk [7], jelen tanulmányban az algoritmus fejlesztésének eredményét, egy futtatható programrendszert (elemző) mutatjuk be. Az elemzőt egy szűk témakört leíró szöveges állományra fejlesztettük.

Az elmúlt két évtized mondattani, fonológiai és fonetikai eredményeire alapozva [2], [3], [5], [6], [8] állítottuk össze azt az algoritmust, amelyikkel jó eséllyel meg tudjuk jósolni, milyen lesz - vagy lehet - a hangsúlyok eloszlása a mondatban. Az elmélet szerint a magyar mondat topik részre és predikátum részre oszlik. A topik nulla, egy vagy több igebővítményt és szabad határozót tartalmaz. Bizonyos típusú összetevők (pl. a határozók *szerencsére, valószínűleg, látszólag* típusú mondathatározók) csak a topik részben állhatnak. A topik rész összetevői mind gyenge hangsúlyt viselnek. A mondat legerősebb hangsúlya a predikátumrész első fő összetevőjére esik. A predikátumrész tetszés szerinti és számú (nulla, egy vagy több) disztributív kvantorral (azaz *mindenki, senki, minden előfizető, a posta is* típusú összetevővel) kezdődik. Ezek mindegyike főhangsúlyos. Őket követi a szintén főhangsúlyos, közvetlenül az ige előtti összetevő, mely akár fókusz (*A* ***postás*** *csengetett be*), akár igekötő (***be****csengetett*), akár névelőtlen főnév (***levelet*** *hozott*) lehet. Az ezt követő ige hangsúlytalan. Bizonyos mondattípusokban az ige előtti pozíciók üresen maradnak és maga az ige a predikátumrész kezdete: ilyen esetben az ige főhangsúlyos. A főhangsúlyos elemek hangsúlyának erőssége balról jobbra csökken. Az ige utáni fő összetevők attól függően hangsúlyosak, hogy ismert vagy új információt közölnek-e és hogy van-e fókusz a mondatban. Az ige utáni disztributív kvantorok akár hangsúlyosak, akár hangsúlytalanok lehetnek.

2 Az elemző felépítése

Az elemző bemenetére a mondat szöveges formája kerül, majd a feldolgozás során a mondat minden szavát egy hangsúlycímkével látja el a program. A feldolgozást két szinten végezzük: (i) tagmondatokra bontás, (ii) a hangsúly kijelölése tagmondatonként. Az elemzőt a következő hangsúlykategóriák meghatározására terveztük: erős hangsúly (F címke), közepesen erős hangsúly (E), normál hangsúly (W), hangsúlytalan a szó (N), erősen hangsúlytalan a szó (-). A megvalósított algoritmus alapvetően elemző modulokra, szabályokra és listákra támaszkodik. Elemző modulok a következők: szófajmeghatározó és NP elemző. Ez utóbbi egy korábban fejlesztett általános algoritmus, amelynek a magyarított és az adott feladatra adaptált változatát használtuk (Gábor 2007). Ez, mint futtatható önálló modul áll rendelkezésre. Az algoritmus összes többi eleme saját fejlesztés. Az elemző 12 modulból épül fel, melyben mindegyik modul azonos koncepcióra épül. Az elemzési algoritmus szabálymoduljait és a listákat alább ismertetjük az algoritmus folyamatának bemutatásával (1. ábra).

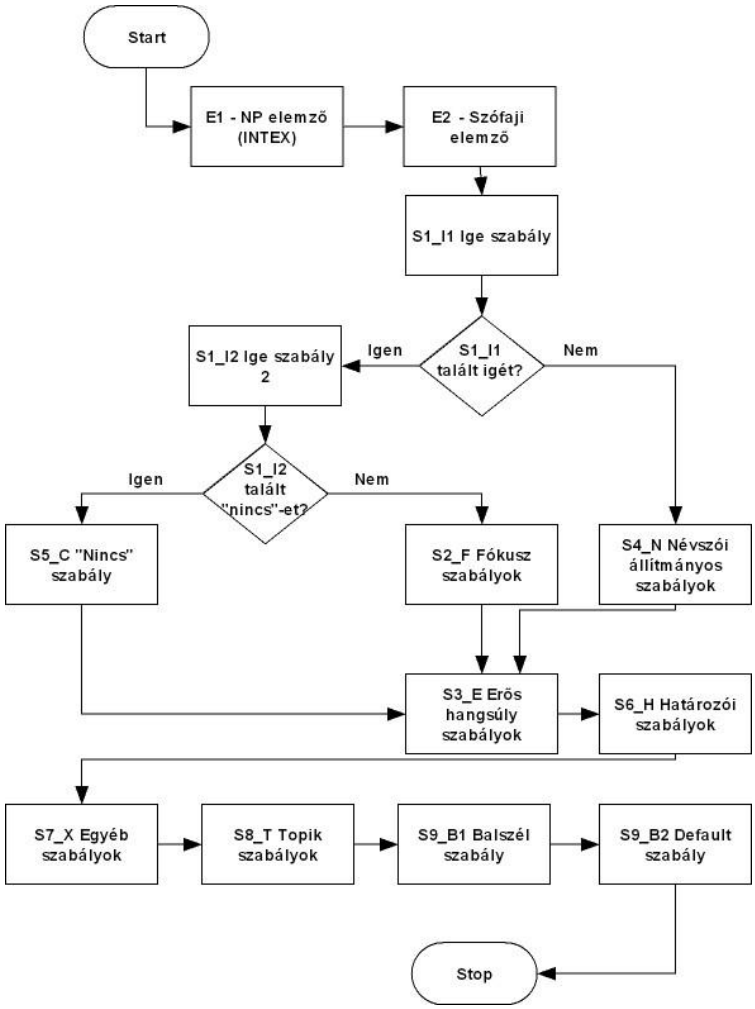


Fig 1. A szöveg alapján működő hangsúlykijelölő elemző moduljai és működési folyamatábrája

E1 – a frázisok meghatározása. Ezt a feladatot az MTA Nyelvtudományi Intézetben magyarított INTEX elemző parancssorból futtatható változata végzi [7]. Ez a modul tartalmaz tokenizálót (az Intexbe beépített funkcióként, ám magyar-specifikus és szöveg-specifikus szabályokkal kiegészítve), szótárat a szöveg lexikai elemzéséhez, valamint a feladathoz szükséges szintaktikai elemző nyelvtanokat. A parancssori alkalmazás egyrészt megjelöli a mondatban található NP-k (főnévi csoportok), melléknévi csoportok (AdjP) és névutós kifejezések kezdetét és végét (beágyazottakét is), másrészt megjelöli a mondat- és tagmondathatárokat. Az elemző kimeneteként egy köztes szövegállomány jelenik meg amelyben {S}, <NP> és <AdjP> címkék jelölik az előbbi szerkezethatárokat. A főnévi csoport határait azért releváns megkeresni,

mert az elmélet szerint [2] a frázisra adott erősebb hangsúly egy főnévi csoportban csak az első „tartalmas” szóra esik. A többi szó az NP-ben semleges hangsúlyt kap.

Az NP elemzés végeredménye erősen befolyásolja a további modulok helyes döntéseit. Az NP keresés számos esetben bizonytalan lehet. Például függhet a szöveg tartalmától és szerkezetétől. Az általános írott szövegek feldolgozására készült NP elemzőt több ponton adaptálni kellett az ügyfélszolgálati szövegek kezelésére, egyfelől a távközlési témájú szövegek speciális szókincse, másfelől a szöveg beszélt nyelvihez közelítő szerkezete miatt. A szókincs megfelelő kezeléséhez szükség volt a szótár kibővítésére. Ezen felül a szövegre jellemző, hogy mondatai nagy arányban tartalmaznak számos, nyílt tokenosztályba tartozó entitást (pl. telefonszámok, egy- vagy többszavas márkanevek, kódok), melyeket nem lehet a szótárban kezelni, így helyes címkézésükről reguláris nyelvtanokkal kell gondoskodni. A reguláris nyelvtanok egy részét az NP-elemzés előtt, a tokenizálást végző modul részeként alkalmazzuk a szövegre (pl. telefonszámok felismerése). A többszavas márkaneveket a szótárban kezeljük, így a feldolgozás későbbi lépéseiben ezek is külön tokenként kezelhetők. A nyílt tokenosztályba tartozó entitások kezelésekor a felismerésen túl az NP-nyelvtant adaptálni kellett a speciális viselkedést mutató, de a névszói kategóriába sorolt entitásokhoz (pl. idegen szó mint főnévi fej esetén gondoskodni kell a fejhez kötőjellel csatolt esetrag felismeréséről is).

A szókincs és különösen a nyílt tokenosztályba tartozó elemek mellett a másik nehézséget a beszélt nyelvihez hasonlító jellegzetességek jelentik, melyek leginkább gondolatjeles közbevetések, valamint hiányos mondatok és tagmondatok formájában mutatkoznak meg. Az ilyen szerkezetek egyfelől megnehezítik a tagmondatra bontást, másfelől újabb hibafaktort jelentenek az NP-k felismerésében is a szerkezeti homonímia miatt.

E2 - szófaji elemző.

Az NP keresés számos esetben bizonytalan lehet. Például függhet a szöveg tartalmától és szerkezetétől. Esetünkben az ügyfélszolgálati szövegek némely pontjának helyes feldolgozására be kellett tanítani az elemzőt. Felsorolunk néhány ilyent. A feldolgozott témakör mondatai nagy arányban tartalmaznak gondolatjeles közbevetéseket, valamint számos, nyílt tokenosztályba tartozó entitást (pl. telefonszámok, egy- vagy többszavas márkanevek, kódok), melyeket nem lehet a szótárban kezelni, így helyes címkézésükről reguláris nyelvtanokkal kell gondoskodni. A nyílt tokenosztályba tartozó entitások kezelésekor a felismerésen túl az NP-nyelvtant adaptálni kell azokhoz (pl. idegen szó mint főnévi fej esetén gondoskodni kell a fejhez kötőjellel csatolt esetrag felismeréséről is). Az NP elemzés végeredménye erősen befolyásolja a további modulok helyes döntéseit.

E2 - szófaji elemző. Ez a modul [4] végzi az ige, a létige, az igekötő, a határozószó, a kötőszó és néhány esetben a szótő felismerését. A szófaji elemző címkékkel látja el a mondat szavait.

S1 – igeazonosítás. A mondatban az ige képezi az első döntési pontot. Két igeszabály végzi a lehetséges esetek teljes feldolgozását. Az első (S1-I1) eldönti, hogy van-e ige a mondatban és ennek megfelelően ágazik el. A második (S1-I2) a talált igét

osztja két kategóriába (tagadás a *nincs* szóval, illetve más ige), e szerint válik ketté a további feldolgozás menete.

S2 – fókusz. A fókusz szabályok két részre oszlanak, a hangsúlyadó és hangsúlytörlő szabályokra. A hangsúlyadó szabály akkor lesz aktív, ha van fókusz a mondatban. A fókusz keresésére a következő négy szabályt alakítottuk ki. Ha névelőtlen főnév, igekötő vagy azzal azonos státusú igerész közvetlen az ige után helyezkedik el, akkor az a frázis, ami az ige előtt helyezkedik el, fókusz (S2-F1). Ha a létige bővítménye közvetlen az ige után helyezkedik el, akkor az a frázis, ami az ige előtt helyezkedik el, fókusz (S2-F2). Ha negatívan minősítő határozószó (Lista-10) áll közvetlenül az ige előtt, akkor ez a határozószó fókusz (S2-F3). Ha van egy frázis a kvantorok (Lista-17) és az ige között, akkor ez a frázis fókusz (S2-F4). A hangsúlytörlő szabály (S2-F5) azt mondja ki, hogy a fókusz után levő ige a fókusz szerkezetétől függően hangsúlytalan (N), illetve erősen hangsúlytalan (-) minősítést kap, a mondatban az utána levő frázisokon (végig) az (N) minősítést kell alkalmazni. Itt két részletszabállyal finomítjuk a végleges döntést. Ha többtagú a fókusz, akkor utána az ige (N) jelzést kap (S2_F5.1), ha egytagú a fókusz, akkor az ige (-) jelzést kap (S2_F5.2.).

S3 – nem fókuszos főhangsúly. Ez a hangsúlyfajta a főhangsúly enyhébb változata, abban különbözik a fókusztól, hogy az utána következő szövegrészben a törlő szabályok másképp működnek. Hat szabály határozza meg az (E) jelű hangsúlyokat. A frázis (E) hangsúlyt kap, ha a névelőtlen főnév (NP), igekötő vagy azzal azonos státusú igerész közvetlen az ige előtt vagy az ige részeként helyezkedik el (S3_E1), ha egy disztributív kvantor (Lista-17) található a mondatban, ha egy frázis disztributív kvantor, vagyis egy *is* szót tartalmaz (S3_E3). Ilyen hangsúlyt kap a tagadószó (S3_E4), illetve a létige ige előtti bővítménye (S3-E5). Ha nincs fókusz vagy más E-frázis, akkor az ige kapja az (E) hangsúlyt (S3_E6). A törlő szabályok (S3-E7) a nem fókuszos főhangsúly után a következők. Az utána levő névelőtlen főnév esetén az ige (N) jelzést kap, igekötő esetén pedig az ige (-) minősítésű, a mondat további részében az utána levő frázisokon végig (W) lesz a hangsúly.

S4 – predikatív mondatok szabály. A névszói állítmányos mondatokban nincs ige. Ezért problémás a hangsúlyok helyes kiosztása. Ennek a szabálynak a kidolgozása folyamatban van, a döntést ilyen esetekben jelenleg az S9 balszél szabály veszi át. Ha van (E) hangsúly a mondatban, akkor minden frázis, a balszél-szabály szerint kap hangsúlyt.

S5 – „nincs" igei szabály. A tagadószó „nincs" (F) főhangsúlyt kap.

S6 – egyéb kategóriák.

Ez a modul a hangsúlykiosztás maradék részeit próbálja megoldani a mondatban. Tíz szabályból áll, legtöbbjüknek lokális, lista-alapú jellege van. A szabályok a következők. Ha a szó kötőszó, akkor törlődik az esetleges hangsúly és (N) jelölésre változik (S6-X1). Az *egy* szó hangsúlyos lesz (W), ha utána vagy számnév a 2. lista szerinti szó, vagy mértékegység jön (S6-X2). Ha van kis fokozatot jelölő összetevő (lásd 14. lista), akkor törlődik a korábbi hangsúly és (N) lesz (S6-X3). A szemantikailag kiüresedett bővítmények (15. lista) esetében törlődik a korábbi hangsúly és (N) lesz (S6-X4). Ha címek és rangok vannak a tulajdonnevek előtt (16. lista), akkor törlődik a korábbi hangsúly és (N) lesz (S6_X5). Mértékegységek esetén (2. lista) törlődik a korábbi hangsúly (S6-X6). A hangsúlyszabályokat befolyásoló konstrukciók (18. lista), páros kötőszók) esetén alkalmazható a (W) hangsúly (S6-X7).

Az S6-X8 szabályok a szövegtípusból adódó hangsúlykiosztást írják le a következők szerint. Hangsúlyos a mondat eleje, a topik (S6-X8.1). Normál hangsúlyt (W) kap az ige, ha a predikátumrész rövidebb a topikrésznél (S6-X8.2). Egy vessző utáni „mondásige", akkor is, ha hátravetett igemódosító áll mögötte + NP a következő hangsúlymintát kapja: ige (-), igemódosító (-), NP balszél-szabály szerinti hangsúly (W- N*) (S6-X8.3). A csillaggal többszöri megjelenést jelölünk, azt, hogy egy (N) egészen az adott frázis végéig alkalmazható. Ha a frázisban található egy listázott hangsúlykerülő (például *kell*), akkor (-) hangsúlyt kell alkalmazni (S6-X9). Ha a frázisban van egy tulajdonnév, akkor arra (W) hangsúlyt kell tenni (S6-X10). Ha nincs fókusz vagy E-elem a mondatban, akkor a balszél-szabály szerinti hangsúly (W-N*) alkalmazható a topik és a határozók után ha van topik vagy határozó (S6-X11).

S7 – határozófrázis szabályok. Fő feladatuk a szövegben rejlő határozók azonosítása, címkézése, hangsúlystruktúrájuk azonosítása és címkézése. Mondat- és módhatározókat különböztetünk meg. Két szabály írja le az ezekkel kapcsolatos elemzés menetét. A mondathatározókhoz két külön szemantikai-prozódiai leírással rendelkező listát használunk (Lista-12, 13). Ha a listákban szereplő mondathatározót találunk, akkor az topikrészben van és topikhangsúlyt kap (S7-H1). Ha módhatározó a 11. lista szerinti, azaz, ha pozitív értelmű a határozószó: fok, mód, gyakoriság (például. *nagyon, eléggé, sokszor, állandóan*) és a fókusz előtti pozícióban van, akkor (W) hangsúlyt kap. Ha mondathatározóként és módhatározóként van listázva a határozószó (S7_H3), akkor a viselkedése változó, elemzési szabály erre még nincs kidolgozva.

S8 – topik szabályok Ha „topikos" a mondat, akkor a frázis vagy a frázisok topikhangsúlyt kap(nak), vagyis a (W) jelzés alkalmazható a mondat elején az első tartalmas szón, utána az (N) a predikátumrész kezdetéig. Meg kell jelölni a topikrész végét ahhoz, hogy a topikhangsúlyt adó szabályok és néhány egyéb mondatprozódiai szabály müködni tudjon. Ezt a feladatot a topikszabályok látják el. A topik szűkebb értelemben egy olyan vonzat, amely a mondatban az ige előtt áll. Az itt alkalmazott "topikrész" alatt viszont azt a részt értjük a mondatban, amely több frázist is tartalmazhat. A topikrész alatt a predikátumrész előtti részt, technikailag az első (E)-jelű szó előtt vagy a fókusz előtt levő szövegrészt értjük. Tehát, az első (E) és (F) jel előtti szövegrész a mondatrészkezdet-jelzésig topikrész, beleértve a határozókat is. A topikrészhez tartozik minden olyan NP és határozó, amely az ige előtt áll és nem egy (E) jelű elem, és nem egy (F) jelű elem. Ha topikos a mondat, akkor a frázis vagy a frázisok topikhangsúlyt kap(nak), vagyis a (W) jelzés alkalmazható a mondat elején az első tartalmas szón, utána az (N) a predikátumrész kezdetéig. Több frázist tartalmazó, hosszabb mondatokban, olyan mondatokban, amelyekben van fókusz, de az algoritmusnak nincs egyetlen formai „kapaszkodója" sem" (pl. a hátravetett igekötő vagy a "*csak*"-szó) a fókusz megtalálásához is releváns megjelölni a topikrészt. Például a mondatban nagy valószínűséggel van fókusz akkor, ha a mondat predikátumrésze lényegesen „nehezebb" a topikrésznél, azaz több NP-ből és határozófrázisból áll.

S9 – balszél szabályok. A frázishoz hozzárendelt hangsúlytípus (F, E vagy W) csak a frázis bal szélén marad meg. A következő szavakon a frázis végéig (N) hangsúly lesz (S9-B1).

Listák

L-1 klitikumok [8] szerint, azaz egy szótagú funkciószók: *a, az, egy, és, de, vagy, is, ha, én, ő, ez, már, még, csak*)

L-2 mértékegységek: *fok, másodperc, perc, óra, nap, hét, hónap, év, évtized, évszázad, évezred, milliméter, centiméter, méter-sorozat, kilométer, láb, mérföld, gramm-sorozat, deka, stb.*

L-3 számnevek: *egy, kettő* ..stb.

L-4 hangsúlykerülő, beférkőző igék: *akar, érint, fog, folyik, talál, kell, szabad, szeretnék…*

L-5 névmások: *én*...

L-6 névutók: *mellett, helyett, után..*

L-7 determinánsok: *a, az, e, ez, egy, eme, ama, ezen, azon, ez a, az a, ..*

L-8 névmásszerű főnevek, „üres szavak": *ország, ügy, kormány, (M/m)agyarország(i)*

L-9 vonzatszótár (opcionális)

L-10 negatív hatarozószók: fok, mód, gyakoriság: *csúnyán, rosszul, ritkán, kevésen, alig*

L-11 pozitív értelmű határozószók: fok, mód, gyakoriság: *nagyon, eléggé, sokszor, állandóan*

L-12 mondathatározók (N-jellel) *esetleg, állítólag*

L-13 mondathatározók (W-jellel) *okvetlenül, feltétlenül, tényleg*

L-14 kis fokozatot jelölő összetevők [8] szerint, azaz *néhány, némi, egy kicsi, néha, néhol, egyelőre, enyhén, kissé, némileg, valaki, valahol, valahogyan, valamennyi*

L-15 szemantikailag kiüresedett bővítmények [8] szerint, azaz, balszélre kerülő melléknevek: *bizonyos, valóságos, szegény, kis*

L-16 címek és rangok a tulajdonnevek előtt vagy után: *néni, bácsi, út, köz, utca, doktor, stb*

L-17 disztributív (univerzális) kvantorok: *mind*-sorozat *(mind, minden, mindenki, mindegyik, valamennyi, az összes, minden alkalommal, mindig ...)*

L-18 hangsúlyszabályokat befolyásoló konstrukciók, például. a páros kötőszók: *nem… hanem, akár... akár, vagy ... vagy, mind ... mind*

3 Az elemző tesztelése

Az elkészült hangsúlyelemző tesztelését 580 mondaton végeztük manuális módszerrel. A mondatok összesen 6974 szót tartalmaztak. Tehát ennyi hangsúlyjelzést ellenőriztünk. Felkészültünk arra, hogy az algoritmus sok hibát fog elkövetni. Már a fejlesztés során kiderült, hogy az ilyen elemzők működése erősen függ a szöveg felépítésétől, tartalmától, tehát az ilyen elemzőket hozzá kell igazítani az elemzett szöveghez (általános nyelvi elemző készítése tehát egyelőre irreális célkitűzés). Példaként bemutatunk néhány elemzendő és elemzett mondatot. A hangsúly jeleket a számítógépes feldolgozás szögletes zárójelek közé teszi, kettősponttal kezdődően.

A következő WAP oldalon a tranzakció részleteinek megadásával, annak jóváhagyására kérjük, amit beállításai szerint vásárlási kódjával, vagy anélkül tehet meg.

[:-]A [:W]következő [:N]WAP [:W]oldalon [:-]a [:W]tranzakció [:N]részleteinek [:N]megadásával,
[:N]annak [:E]jóváhagyására [:N]kérjük,
[:N]amit [:W]beállításai [:N]szerint [:W]vásárlási [:N]kódjával,
[:N]vagy [:W]anélkül [:N]tehet [:N]meg.

A forgalmi díj – az 50 Mbyte feletti forgalom esetén – minden időszakban mindössze 1 Ft minden megkezdett 10 kbyte után.

[:-]A [:W]forgalmi [:N]díj [:-]az [:W]1000 [:N]Mbyte [:N]feletti [:N]forgalom [:N]esetén
[:E]minden [:N]időszakban [:W]mindössze [:W]1 [:N]Ft [:E]minden [:N]megkezdett [:N]10
[:N]kbyte [:N]után.

Átírás cégek közötti jogutódlás a Számlafizető elhalálozása esetén lehetséges, minden egyéb esetben Számlafizető módosítás történik.

[:W]Átírás [:W]cégek [:N]közötti [:N]jogutódlás
[:N]és [:-]a [:W]Számlafizetõ [:N]elhalálozása [:N]esetén [:W]lehetséges,
[:N]minden [:N]egyéb [:W]esetben [:W]Számlafizetõ [:E]módosítás [:N]történik.

A Budapest Bank mobilbank menüjében lekérdezheti számlatörténetét és bankszámlájának egyenlegét, módosíthatja bankkártyája limitösszegeit, letilthatja, vagy aktiválhatja azt, átutalásokat végezhet, árfolyamokhoz juthat, vagy egyenlege változásairól értesítést állíthat be.

[:-]A [:W]Budapest [:N]Bank [:F]mobilbank [:N]menüjében [:N]lekérdezheti
[:N]számlatörténetét [:N]és [:N]bankszámlájának [:N]egyenlegét,
[:N]módosíthatja [:W]bankkártyája [:N]limitösszegeit,
[:N]letilthatja,
[:N]vagy [:N]aktiválhatja [:N]azt,
[:E]átutalásokat [:N]végezhet,
[:E]árfolyamokhoz [:N]juthat,
[:W]vagy [:W]egyenlege [:N]változásairól [:F]értesítést [:-]állíthat [:N]be.

4 A tesztelés eredményei

A hangsúlyjelzések ellenőrzésének eredményei szerint az 580 mondatból 98-ban nem volt hibás jelölés. A 6974 szóból összesen 1060 szón találtunk hibás jelölést. Hibák típus szerinti eloszlását az 1. táblázat mutatja. A négy hibatípus közül azokat számítjuk nagyobb hibának, amikor az adott szóra F, E, W kerül annak ellenére, hogy a szó hangsúlytalan, tehát az 1, 2, 3 kategóriákat. Ezekből összesen 302 esetet találtunk. Kevésbé zavaró hibának számítjuk a 4. kategóriát, mivel a hiányzó hangsúly kevésbé zavarja meg a hangzási képet, mint feleslegesen hangsúlyozott szó.

1. Táblázat: a hangsúly jelölési hibák eloszlása a vizsgált 580 mondatban

	A hiba típusa	A hibákszáma
1.	N kell F helyett	8
2.	N kell E helyett	64
3.	N kell W helyett	230
4.	W kell N helyett	228

Az eredmények tehát azt mutatják, hogy a súlyosabb hibából több van, mint a kevésbé zavaróból. A vizsgálatokból világossá vált, hogy a mondatelemzés legjelentősebb része a főnévi csoportok azonosítója. Egy ilyen mondatelemzési hiba több hangsúlyhibához is vezethet, mivel a hangsúlystruktúra a mondatszerkezetre épül. Végeredményben azt mondhatjuk, hogy az elemzési eredmények javítására egyrészről a főnévi csoportok azonosítóját kell pontosítani, más részről, az empirikus kutatást kell kiterjeszteni minél több mondatra. További vizsgálatokat kell végezni a szabályok sorrendjével kapcsolatban is, valamint azokat az eseteket is vizsgálni kell, ha két szabály esetleg ütközik egymással. Végül megjegyezzük, hogy a vizsgált szövegek nyelvi szerkezete sem mindig támogatja a sikeresebb elemzést.

7 Összefoglalás

Bemutattuk az első olyan mondatelemző algoritmust, amelyikkel magyar nyelvre végezhető hangsúly meghatározás. Legfőbb eredménynek azt tartjuk, hogy sikerült az eddigi, a témába vágó elméleti eredményeket algoritmikus formába önteni és működő programmá fejleszteni. Az automatikus elemzés hibázási aránya elég magas, a vizsgált szavak 15%-át hibás jelöléssel látja el. Ennek a hibaaránynak a csökkentése szerepel további céljaink között, valamint az, hogy kipróbáljuk az elemzőt különböző szövegtípusokon és meghatározzuk, hogy mennyi adaptációra van szükség, ha nem ügyfélszolgálati szövegeket kell elemezni.

Bibliográfia

1. Gábor, K.: Syntactic Parsing and Named Entity Recognition for Hungarian with Intex. In: Silberztein, Koeva, Maurel (eds.): Formaliser les langues avec l'ordinateur. Presses Universitaires de Franche-Comté, Besancon (2007) 353–366
2. É. Kiss, K.: The Syntax of Hungarian. Cambridge Syntax Guides. Cambridge: Cambridge University Press (2002)
3. É. Kiss, K., Kiefer, F., Siptár, P.: Új magyar nyelvtan. Budapest: Osiris (1998)
4. Kiss, G., Németh, G.: Tisztán statisztikai alapú szófaji címkéző használata a Szeged Korpuszon. IV. Magyar Számítógépes Nyelvészeti Konferencia MSzNy 2006., Szeged, (2006) 52–59

5. Koutny, I., Olaszy, G., Olaszi, P.: Prosody prediction fromt text in Hungarian and its realization in TTS conversion. International Journal of Speech Technology, Volume 3, Numbers 3-4. Kluwer Academic Publishers. (2000) 187–200
6. Olaszy, G.: The most important prosody patterns of Hungarian. Acta Linguistica Hungarica, Vol. 49 (3-4) (2002) 277–306
7. Tamm, A., Olaszy, G.: Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához, in Z. Alexin, D. Csendes (szerk), III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, (2005) 383–393
8.Varga, L.: Intonation and Stress: Evidence from Hungarian. New York: Palgrave Macmillan (2002)

Ezt a kutatás-fejlesztést az NKFP 2. programja (szerződésszám: 2/034/2004) támogatta.

Érzelmes beszéd gépi előállítása érzelem specifikus beszédadatbázisok felhasználásával

Fék Márk, Zainkó Csaba, Németh Géza

Beszédtechnológiai Laboratórium,
Távközlési és Médiainformatikai Tanszék,
Budapesti Műszaki és Gazdaságtudományi Egyetem,
1117 Budapest, Magyar tudósok körútja 2.
{fek, zainko, nemeth}@tmit.bme.hu

Kivonat: Tanulmányunkban megvizsgáljuk hogyan lehet érzelem specifikus beszédadatbázisok felhasználásával gépileg érzelmes beszédet előállítani. Kísérletünket magyar nyelvre végeztük, de a módszer nyelvfüggetlen. Felvettünk egy szemantikailag semleges tartalmú mondatot és 26 logatomot amelyek a mondat szintetizálásához szükséges diádokat és CVC triádokat tartalmazták. A hanganyagot egy profi színésznő mondta fel a hat alapérzelemnek megfelelően, illetve semleges érzelmi változatban. A logatomok felhasználásával 7 érzelemfüggő beszédelem adatbázist hoztunk létre. A 7 beszédelem adatbázist összepárosítva a természetes mondatokból kinyert 7 prozódiai kontúrral 49 szintetizált mondatot állítottunk elő. A logatomokban, illetve a természetes és a szintetizált mondatokban hallható érzelmeket 194 tesztalany értékelte ki. A tesztelők a logatomok 99%-ban, illetve az összes természetes mondatban szignifikánsan a véletlen találgatás szintje felett ismerték fel a színésznő által kifejezett érzelmeket. Az érzelmek azonosítási aránya egyes szintetizált mondatok esetén meghaladta a természetes mondatokét.

1 Bevezetés

A napjainkban elterjedten használt elemösszefűzéses beszédszintetizátorok általában adnak valamilyen lehetőséget a prozódiai paraméterek (alapfrekvencia, hangidőtartamok, intenzitásmenet) vezérlésére. A gépi beszéd hangszínezetét azonban alapvetően a beszédadatbázis határozza meg, amelyből az összefűzendő elemeket kiválasztják. Ismert, hogy a beszédben az érzelmeket a prozódia és a hangszinezet együttesen hordozza [1].

Korábbi kísérletek több nyelv esetében vizsgálták [2,3] az érzelem specifikus beszédadatbázisok használatának lehetőségét elemösszefűzéses beszédszintézis esetében. A kísérletek minegyikében a bemondók felolvastak egy szöveget az előzetesen definiált érzelmeknek megfelelő változatokban. A felvett beszédből minden érzelmehez egy-egy érzelem specifikus adatbázist hoztak létre. Elemkiválasztáson alapuló beszédszintézis segítségével szemantikailag semleges mondatokat állítottak elő az egyes érzelmi adatbázisokból.

A [2] és [3]-ban leírt kísérletekben a szintetizált mondatok prozódiai paramétereit természetes mondatokról másolták, illetve [2] második kíséletében a beszédszintézis algoritmus állította elő a prozódiát. Ezután meghallgatásos teszteket végeztek, hogy meghatározzák a szintetikus mondatokkal kifejezni szándékolt érzelmek felismerési arányát.

Montero és szerzőtársai [2] szemantikailag semleges szöveget vettek fel spanyolul egy professzionális színész segítségével, aki négy érzelemmel és semleges stílusban mondta be a szöveget. A szerzők a a természetes mondatokról átmásolt prozódia segítségével szintetizált mondatokon a véletlen szintnél magasabb azonosítási arányt értek el (semleges 76%, öröm 62%, meglepetés 91%, szomorúság 81%, harag 95%) egy 6 választási lehetőséget megengedő meghallgatásos tesztben (amely a "nem azonosítható érzelem" opciót is tartalmazta). Hasonló eredményeket értek el (kivéve a meglepetésre 53%) az érzelmi adatbázisokon betanított, automatikusan generált prozódia használata esetében is.

Bulut és szerzőtársai [3] örömöt, szomorúságot, haragot és semlegességet közvetítő rövid szövegrészeket vettek fel egy fél-profi színésznő segítségével. Az adatbázisokból előállított mondatokat a véletlen találgatás szintje felett azonosították (harag 86%, szomorúság 89%, öröm 44%, semleges 82%) egy 4 választási lehetőséget tartalmazó meghallgatásos tesztben.

Schröder és Grice [4] egy teljes német diádos elemkészletet vettek fel három különböző vokális erőfeszítés mellett (lágy, modális és hangos). Meglátásuk szerint a különböző hangszalag feszességeknek megfelelő vokális erőfeszítések fontos szerepet játszanak az érzelmek kifejezésében. A diádokat értelem nélküli hordozó szavakba (logatomokba) ágyazták és egy német anyanyelvű bemondó olvasta fel őket, akit megkértek, hogy állandóan tartott hangmagasságon beszéljen. Egy (elemkiválsztás nélküli) elemösszefűzéses szintetizátort használtak a tesztmondatok előállításához. A meghallgatásos teszt folyamán kimutatták, hogy a szintetizált mondatokban a szándékoltnak megfelelően észlelhető a vokális erőfeszítés.

A fent felsorolt munkák ellenére még mindig kérdéses, hogyan lehet olyan érzelmes beszédszintetizátort készíteni, amely a megfelelő hangszínezettel állítja elő az egyes érzelmeket. Tanulmányunkban az érzelem specikfikus diád és triád elemkészlet használatát vizsgáljuk érzelmes beszéd előállítására.

Kísérletünk hasonló a korábbiakhoz a következő lényegi eltérésekkel. Az általunk használt szintézis módszer elemkiválasztás nélküli elemösszefűzésen alapul és [4]-hez hasonlóan szintén értelem nélküli logatomokat vettünk fel, de a bemondott logatomok közvetlenül hordozták az egyes érzelmeket, hasonlóan a [2] és [3]-ban felolvasott szövegekhez. A mi beszédelem-készletünk mind diádokat, mind CVC triádokat tartalmazott növelve az összefűzés hangzásának folytonosságát a csak diádokból álló elemkészlethez képest. Meghallgatásos teszttel ellenőriztük az érzelmek felismerhetőségét az egyes logatomokban. A forrásanyag ilyen jellegű formális ellenőrzése a korábbi munkákban nem történt meg. Ellentétben a [4]-ben leírtakkal, a színésznő szabadon variálhatta a hangmagasságát, amely jelentős változatosságot mutatott a logatomokon belül és azok között, illetve a mondtokon belül is. Emiatt a szintézis folyamán alkalmanként jelentős alapfrekvencia módosításra volt szükség, amely nemkivánt torzulást vitt a szintetizált jelbe. Az elemkiválasztáson alapuló szintézis módszerek is érzékenyek az érzelmes beszédre jellemző jelentős alapfrekvencia

változásokra, ami fokozottan jelentkezhet a [2] és [3]-ban használt kisméretű elemkészletek esetében. Az alapfrekvencia változtatás okozta torzulások feltehetően befolyásolták ezen kísérletek eredményeit. Az idézett szerzőkhöz képest több érzelmet vizsgáltunk, ami több választási lehetőséghez vezetett a meghallgatásos tesztek folyamán rontván a várható érzelemazonosítási arányt. Kísérletünket magyar nyelvre végeztük, de a korábbi kísérletekhez hasonlóan a módszer más nyelvek esetében is alkalmazható.

2 A hanganyag felvétele

A kísérletben a következő szemantikailag semleges mondatot használtuk: „A menüben minden szükséges információ elhangzik”. Meghatároztuk mondat fonetikus átírásának megfelelő, diádokból és CVC triádokból álló szekvenciát. Összeségében 14 CVC triádot, 3 VV és 9 CC diádot használtunk. A diádokat és triádokat jelentés nélküli szavakba (logatomokba) ágyaztuk, amelyek a CC diádok esetében kétszótagosak, a VV diádok és a CVC triádok esetében három szótag hosszúak voltak.

A semleges mondatot és a 26 logatomot 7 változatban vettük fel a hat alapérzelmet (öröm, harag, meglepetés, undor, szomorúság, félelem) és semlegességet kifejezve. A hanganyagot egy professzionális (30 éves) magyar színésznő mondta be, akinek már volt tapasztalata beszédszintézishez szükséges elemkészlet felvételével, illetve korábbi kísérletekhez már mondott fel érzelmes mondatokat. Minden egyes érzelmhez iteratívan több (4-7) változatban kerültek bemondásra a logatomok, amíg a cikk szerzői megfelelőnek nem ítélték őket. A felvétel után informális meghallgatásos tesztet végeztünk, melynek során a cikk két szerzője kiválasztotta az érzelmileg legmeggyőzőbb logatomokat és mondatokat. A logatomok kb. 20%-át nem volt elég kifejező, ezért azokat egy második alkalommal újra bemonddattuk. Az első felvétel folyamán készült legmeggyőzőbb mintákat lejátszottuk a színésznőnek az egyes érzelmek másodszori felvétele előtt. A logatomok több (4-6) változatban is bemondásra kerültek, melyek közül a cikk két szerzője kiválaszotta a legjobban sikerülteket.

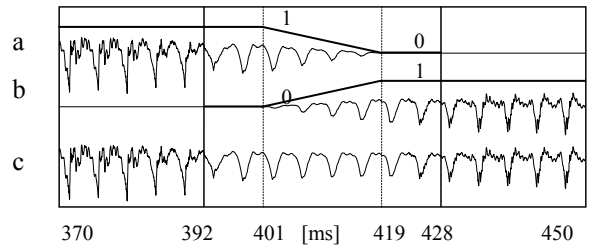
A felvételeket süket szobában készítettük AKG C-414 B-ULS kondenzátor mikrofonnal, a kardioid karakterisztika és a 75 Hz-es aluláteresztő szűrés bekapcsolásával. A mikrofon elé pop szűrőt helyeztünk. A beszédjelet 44.1 kHz-en, mintánként 16 biten digitalizáltuk.

3 Szintetizált minták előállítása

A logatomokból készített hét érzelem specifikus beszédelem adatbázist és a hét változatban bemondott mondatokat az összes lehetséges módon összepárosítottuk. Mind a 49 pár esetén a természetes mondat prozódiáját az érzelem specifikus adatbázisból szintetizált mondatra másoltuk.

A szintézist és a prozódia másolását a következő módon végeztük. A banghatárokat manuálisan bejelöltük a felvett logatomokban és mondatokban. Az alap-

frekvencia menetet a Praat-ban implementált autokorrelación alapuló algoritmus [5] segítségével detektáltuk. A cél lapfrekvencia görbék periódusonként egy frekvencia értéket definiáltak. Az intenzitást 8 ms-onként számítottuk 32 ms széles ablak segítségével. A Praat-ban implementált PSOLA alapú prozódia módosító algoritmust [5] használtuk a hangidőtartamok, az alapfrekvencia és az intenzitások diádok és triádokra másolásához.



1. ábra: Szomszádos triádok összefűzése időtartománybeli átmosással (50%-os átfedéssel).

Fontos megjegyezni, hogy a diádok két egymást követő hangot teljesen lefednek, míg a triádok három egymás követő hang teljes időtartamán át tartanak. Mint az 1. ábrán látható, a prozódia módosítás után a diádokat és a triádokat időtartománybeli átmosással fűztük össze, olymódon, hogy a szomszédos szegmensek első, illetve utolsó hangjai között 50% átfedés legyen. Az ábrán a függőleges folytonos vonalak a hanghatárokat, a függőleges szaggatott vonalak pedig az átfedési tartományt jelölik. A felső két görbe (a és b) mutatja az összefűzött triádokat az átlapoló ablak feltüntetésével. Az alsó görbe (c) mutatja az összefűzés után kapott jelet.

4 Meghallgatásos teszt

Egy webes felületen végzett meghallgatásos teszt során meghatároztuk, hogy a tesztelők milyen érzelmeket azonosítanak a logatomokban, illetve a természetes és a szintetizált mondatokban. 208 magyar anyanyelvű alany vett részt a vizsgálatban. A tesztalanyok többsége motivált informatika szakos hallgató volt.

14 tesztelő eredményét nem vettük figyelembe, mert vagy nem fejezték be a tesztet, vagy véletlenszerű válaszokat adtak. A kizárt tesztelők közül néhányan hanglejátszási problémát jelzetek. A maradék 194 résztvevő közül 159 férfi és 35 nő volt, átlagos koruk 23 év. 83 tesztelő fej- vagy fülhallgatót használt, míg 111 tesztelő hangszóró segítségével hallgatta meg a felvételeket.

A teszt hat részből állt és átlagosan 18 percet vett igénybe. A résztvevők az első részben a logatomokat, a negyedikben pedig a természetes és szintentizált mondatokat értékelték ki. A második és a negyedik rész egy a jelen tanulmánytól független vizsgálathoz tartozott. A harmadik és a hatodik részt azért iktattuk be, hogy ki tudjuk szűrni a véletlenszerűen válaszoló tesztelőket. Ebben a két részben egy 5-fokozatú skálán kellett minősteni egy mondat természetes, illetve három különböző

szintetizátorral előállított változatát. Négy tesztelőt kizártunk, mivel az általuk adott minősítések nem tükrözték a felvételek közötti különbséget.

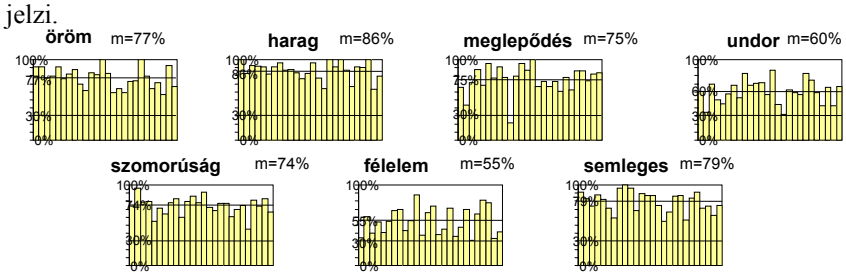
Hogy csökkentsük a tesztelők terhelését, egy résztvevőnek nem kellett az összes mondatot, illetve logatomot meghallgatnia. A 49 szintetizált és 7 természetes mondatot összekevertük és véletlen választással két 28-28 mondatot tartalmazó csoportra bontottuk. 98 tesztelő csak az első, míg 94 csak a második csoportot hallgatta meg. Hasonlóképpen a logatomokat is hét egyforma méretű csoportra osztottuk. Egy ilyen csoportot legalább 23-an hallgattak meg.

A tesztelőknek 7 lehetőség (hat alapérzelem, illetve semlegesség) közül kellett kiválasztani a logatomok, illetve a természetes és a szintetizált mondatok által kifejezett érzelmet. A teszttel kapcsolatos ismertető szöveget ugyanaz a színésznő olvasta fel, mint aki a hangját adta a kísérlethez, így a résztvevők megismerhették érzelmileg semleges beszédstílusát. Azért, hogy ne lehessen ezt a részt átugorni, a továbblépéshez meg kellett adni egy az ismertető végén elhangzó, véletlenszerűen generált kódot. Az egyes elemek között a tesztelők maguk szabályozták a továbblépést. Egy mintát akárhányszor meg lehetett hallgatni, viszont egy korábban már kiértékelt mintára nem volt megengedett a visszatérés. A minták lejátszási sorrendje tesztelőnként véletlenszerűen változott. A tesztben szereplő hangminták honlapunkon elérhetőek: http://speechlab.tmit.bme.hu

5 Eredmények

Annak eldöntésére, hogy az azonosítási arányok szignifikánsan a 14%-os véletlen szint felett vannak-e binomiális próbát ($p<0.05$) használtunk nemegyenlő (1:7) arányokkal. A logatomos teszt esetében (23 tesztelő) a 30% feletti azonosítási eredmény volt szignifikánsan a véletlen szint felett. A mondatok kiértékelésénél (94 tesztelő) a 21% feletti azonosítási arány volt szignifikánsan a véletlen szint felett.

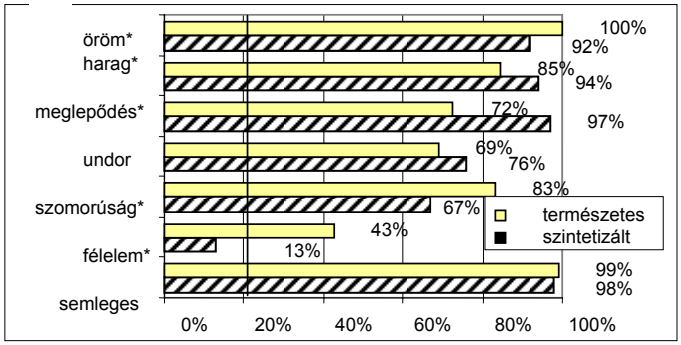
A 2. ábra minden egyes logatomra mutatja a kifejezni szándékolt érzelem felismerési arányát. Az alsó vonal mutatja a szignifikánsan a véletlen szint feletti szintet. A felső vonal az adott érzelemhez tartozó adatbázis átlagos felismerési arányát jelzi.



2. ábra: Érzelmek azonosítási aránya a 26 logatomra.

A kifejezni szándékolt érzelmet a 182 logatomból 180 esetében szignifikánsan a véletlen szint felett azonosították. A harag volt az átlagban legjobban felismert érzelem (86%), míg az undor (60%) és a félelem (55%) sokkal kevésbé volt felismer-

hető. Ugyanakkor több undort és félelmet kifejező logatom 80% felett teljesített, ami mutatja, hogy a tesztelők ezeket az érzelmeket képesek rövid mintákon is felismerni. A többi logatomra kapott alacsonyabb eredmény oka valószínűleg az, hogy a színésznőnek nehéz volt ezen érzelmek konzisztensen az összes mintán kifejezni.



3. ábra: Érzelmek azonosítási aránya a természetes mondatokra, illetve az egyező érzelmi adatbázissal és prozódiával előállított szintetizált mondatokra. A szignifikáns különbséget * jelöli.

A 3. ábra mutatja szándékolt érzelem felismerési arányát természetes mondatokra, illetve az egyező érzelmi adatbázissal és prozódiával rendelkező szintetizált mondatokra. Két-mintás t-próba ($p<0.05$) segítségével ellenőriztük, hogy a természetes és a szintetizált mondatok felismerési arányai között szignifikáns-e az eltérés. A függőleges vonal által jelölt értékénél magasabb felismerési arányok szignifikánsan a véletlen szintje felettiek.

A természetes mondatok közül a legjobb eredményt az öröm (100%) érte el. A haragot (85%) semlegesnek ítélte a tesztelők 9%-a. A meglepetést (72%) örömnek ítélte a tesztelők 28%-a. Ennek az lehetett az oka, hogy a színésznő pozitív meglepetést kifejezendő mosolyogva fejezte be a mondatot. A színésznőnek nehéz volt az undor (69%) kifejezése, amit 12% örömnek, illetve 10% haragnak gondolt. A szomorúságot (83%) semlegesnek ítélte a tesztelők 11%-a. A legkevésbé felismert félelmet (43%) a tesztelők 30%-a semlegesnek, míg 22%-a szomorúnak tartotta. A színésznő a mondat második felében abbahagyta a hangmagasság remegtetését, ami magyarázatot adhat a semleges válaszok magas arányára. Ráadásul a színésznő lágy hangon mondta be a mondatot, ami megmagyarázhatja miért tartották azt sokan szomorúnak. A mondat még a félelmet kifejező logatomok átlagánál is rosszabb eredményt ért el, ami szintén mutatja a kevésbé szerencsés realizációt. A semleges mondatot (99%) könnyen azonosították a tesztelők.

A szintetizált örömöt (92%) könnyen felismerték. A kapott arány jelentősen meghaladja a [2]-ben és [3]-ban közölt korábbi eredményeket. [2] szerzői feltételezték, hogy a mosolyt tartalmazó, illetve nem tartalmazó beszédelemek keveredése csökkenti a szintetizált mondat örömteli hangszínezetét. Az öröm viszonylag alacsony felismerési arányát [3] szerzői részben a kevésbé sikeres színészi teljesítménnyel magyarázták. A természetes örömöt kifejező mondatra, illetve logatomokra kapott magas felismerési arányok mutatják, hogy a fenti problémák egyike sem jelentkezett

az általunk végzett kísérletben. A szintetizált haragot (94%) szintén könnyen felismerték. Érdekes módon a szintetizált harag felismerési aránya szignifikánsan meghaladta a természetes harag felismerését. Megfigyeltük, hogy a prozódia módosító algoritmus érdesebb hangzásúvá tette a beszédet, ami magyarázatot adhat a jobb eredményre. Hasonló megfigyelést [2] is közölt. A szintetizált meglepetést (97%) könnyen felismerték, és a természetes mondattal ellentétben nem tévesztették össze az örömmel. Ez megmagyarázható, ha figyelembe vesszük, hogy a szintetizált mondat hangszínezete (a természetes mondattal ellentétben) nem ment át mosolygásba a mondat végén. A szintetizált undort (76%) a tesztelők 13%-a a haraggal keverte össze. Továbbá a szintetizált felismerési arány meghaladta (de nem szignifikánsan) a természetes mondatét. Mindkét megfigyelés magyarázza a szintetizélt mondat érdes hangszínezete. A szintetizált szomorúságot (67%) a tesztelők 17%-a semlegesnek, míg 7%-uk félelemnek ítélte. Megfigyeltük, hogy a szintetizált szomorú mondat a többi érzelemhez képest torzabban hangzott, ami megmagyarázza a kapott viszonylag alacsony eredményt. A szintetizált mondaton leginkább eltorzult szakaszokat megvizsgálva észrevettük, hogy az azoknak megfelelő szomorú logatomok szabálytalan zönge periódusokat (glottalizációt) tartalmaztak. A szomorú beszédelem adatbázis használatával előállított összes mondatban hasonló torzítást találtunk. Továbbá azt is észrevettük, hogy a természetes szomorú mondat is tartalmaz glottalizációt. A színésznő valószínűleg részben glottalizációval fejezte ki a szomorúságot, de a szabálytalan zönge periódusok megzavarták az alapfrekvencia detektáló és módosító algoritmusok működését, ami torzulást okozott. A szintetizált félelmet (13%) a véletlen szint alatt ismerték csak fel. A mondatot a tesztelők 64%-a haragnak, 16%-a semlegesnek, 6%-a pedig undornak azonosította. Az alacsony felismerési eredmény részben a természetes félelmre kapott alacsony eredménnyel magyarázható. A szintetizált semlegesség (98%) majdnem olyan jó eredményt ért el, mint természetes párja.

Az 1. táblázat mutatja az összes (49) adatbázis és prozódia párosításra kapott felismerési eredméyeket. Az egy-egy adott érzelmet legjobban kifejező párosításhoz tartozó cellákat árnyékolással jelöltük meg. Eltérő adatbázis és a prozódia esetén a függőleges nyíl jelzi, ha a mondatot az adatbázis érzelmi tartalmának megfelelően azonosították. A vízszintes nyíl jelzi, ha a mondatot a célprozódiával megegyezően azonosították. A felismert érzelmet külön kiírtuk, ha az sem az adatnázisnak, sem a prozódiának nem felelt meg. A félkövérrel jelzett számok jelzik az adott párosításhoz tartozó legmagasabb felimerési arányt. A táblázatban csak a véletlen találgatás szintjét szignifikánsan meghaladó értékeket tüntettük fel.

Az öröm és a félelem kivételével minden érzelem esetében az egyező adatbázis és prozódia párosítása adta a legjobb eredményt. A félelmet csak 13% azonosította egyező prozódia és adatbázis esetén. Sokkal jobb eredményt hozott, ha természetes félelem mondat prozódiáját a szomorú (62%), vagy a meglepődés (60%) adatbázisból előállított mondatra másoltuk. Az így kapott felismerési arányok még a természetes félelemre (43%) kapott eredményt is meghaladták. Megfigyeltük, hogy a félelem adatbázisban szereplő logatomok „préselt” hangzással lettek bemondva, ellentétben a szomorúságot illetve félelmet kifejező logatomokkal. Ez megmagyarázhatja, miért lett „lágyabb” az utóbbi két adatbázisból szintetizált mondatok hangzása. A 2. ábra alapján a „préselt” logatomok többé-kevésbé a szándékolt érzelmet fejezték ki

(átlag=55%), de a prozódia módosítás által bevitt érdes hangzás a harag felé (64%) tolta el a mondat felismerését.

1. táblázat: Érzelmek azonosíátsi aránya szintetikus mondatokra az összes prozódia és adatbázis párosítás esetén.

↑adatbázis ←prozódia	öröm	harag	meglepetés	undor	szomorúság	félelem	semleges
öröm	**92%**	← 2%, **↑74%**	**←35%,** ↑14% félelem 33%	← 2%, ↑ 14% **félelem 37%**	← 3%, **↑ 50%** félelem 29%	←11%, **↑35%** szomorú 23%	←1%, **↑ 50%** szomorú 23%
harag	←1%, **↑ 93%**	**94%**	←12%, ↑12% **félelem 41%**	**←63%**, ↑ 22%	← 3%, **↑39%** semleges 27% félelem 26%	**← 60%,** ↑ 2%	←35%, **↑ 39%**
meglepetés	**←68%,** ↑ 31%	**← 56%,** ↑41%	**97%**	**← 82%**, ↑ 5%	**←75%**, ↑ 5%	**← 85%**, ↑ 1%	**← 87%,** ↑ 4%
undor	←13%, **↑ 82%**	← 48%, **↑49%**	←23% , ↑10% **öröm 29%**	**76%**	← 14%, **↑45%**	**← 71%,** ↑ 2%	**← 68%**, ↑ 9%
szomorúság	← 1%, **↑ 78%**	← 0 %, ↑ 33% **semleges 34%** undor 30%	← 11%, ↑ 12% **semleges 37%**	←16%, **↑ 43%**	**67%**	← 20%, ↑12 % **undor 36%** semleges 28%	← 32%, **↑ 52%**
félelem	← 3%, **↑ 74%**	←1%, **↑87%**	**←60%**, 11%	←6%, ↑22 % **harag 58%**	**←62%**, ↑18%	13% **harag 64%**	← 9%, **↑77%**
semleges	←22%, **↑ 75%**	←33%, **↑54%**	**←79%**, ↑ 9%	**←80%**, ↑13%	**←79%**, ↑13%	**←85%**, ↑ 0%	**98%**

Az öröm adatbázis harag prozódiával kombinálva (93%) némileg magasabb felimerési arányt ért el, mint öröm prozódiával (92%). Ez azt jelentheti, hogy egy közös prozódiai modell használható ezen érzelmek szintetizálásához.

Ha az 1. táblázatban megvizsgáljuk a semleges prozódiához tartozó párosításokat, észrevehetjük, hogy a szintetizált harag (54%) és öröm (75%) kifejezésében inkább az adatbázis játszik szerepet. A semleges adatbázishoz tartozó párosítások alapján látható, hogy a meglepetést (87%) és az undort (68%) inkább a prozódiával lehet kifejezni. Hasonló eredményre jutottak [2] és [3] szerzői is, az undort (ami nem szerepelt a korábbi kísérletekben) és a szomorúságot kivéve, mely inkább a prozódiától függpt mindkét korábbi kísérletben. Az utobbi eltérés egy lehetséges magyarázata az, hogy a mi kísérletünkben a színésznő glottalizáció segítségével fejezte ki a szomorúságot. A szintézishez használt módszer nem tudta kezelni a szabálytalan zönge periódusokat, ami akadályozta a szomorú prozódia szintetikus semleges mondatra (13%) másolását. Ezt a feltételezést megerősíti a szomorú adatbázis párosítása öröm

(szomorúság 50%), harag (szomorúság 39%) és undor (szomorúság 45%) prozódiával. Ezen mondatok prozódiáját jobban sikerült a szomorú adatbázisból szintetizált mondatra ültetni, és mindegyik esetben a szomorúság volt a legnagyobb arányban felismert érzelem.

6 Összefoglalás

A kísérlet igazolta, hogy a szintetizálással előállított harag és öröm meggyőzően kifejezhető érzelem specifikus adatbázisok segítségével. Az örömteli beszédelem adatbázis és a haragos prozódia kombinációja meggyőzően fejezte ki az örömöt. Ennek alapján elégséges lehet egy közös prozódiai modul használata ezen érzelmek esetében. A meglepetés és az undor pusztán prozódia segítségével is kifejezhető, de a megfelelő hangszínezet tovább növeli felismerhetőségüket.

Az undort és félelmet kifejező természetes mondatok kevéssbé voltak meggyőzőek, mint a többi mondat. Ez részben megmagyarázza ezek szinetizált változataira kapott rosszabb felismerési arányokat is. A prozódia módosító algoritmus érdessé tette a félelem adatbázisban szereplő erősen „préselt" logatomok hangzását. Az adott technológiai korlátok mellett előnyös lehet kevésbé „préselten" képzett logatomok használata.

A természetes szomorú felvételek szabálytalan zöngeperiódusokat tartalmaztak, amelyeket nem tudott megfelelően kezelni a szintézis algoritmus. A szintetikus mintákban megjelenő járulékos torzítás megmagyarázza a gépi szomorúságra kapott rosszabb felismerési arányt. A szintetizált szomorú kifejező ereje javulhat mesterséges glottalizáció hozzáadásával, ugyanakkor a prozódiamódosítás technológiai korlátai miatt a beszédelem adatbázisokban kerülni kell a glottalizációt.

Az elemösszefűzéses beszédszintézis segítségével lehetséges meggyőzően érzelmeket kifejezni, viszont a különböző érzelmekhez tartozó beszédelem adatbázisok felvétele igen megterhelő. Az egyes érzelmekre jellemző hangszínezet feltehetően inkább marad konzisztens egy rövid logatomon, mint egy teljes mondaton belül. Egy hangszínezet módosító algoritmust egy viszonylag kisméretű, de konzisztens logatom adatbázison betanítva akár hosszabb, elemkiválasztáshoz használt beszédadatbázisokat is érzelmileg kifejezővé lehetne tenni.

7 Köszönetnyilvánítás

A munkát a Nemzeti Kutatási és Technológiai Hivatal támogatta (NKFP 2/034/2004).

Bibliográfia

1. Ladd, D.R., Silverman, K., Tolkmitt, F., Bergmann, G., and Scherer, K.R., "Evidence for the independent function of intonation contour type, voice quality, and f0 range in signalling speaker affect", Journal of the Acoustic Society of America, 78 (2), pp. 435–444, 1985.
2. Montero, J.M., Arriola, G.J., Colas, J., Enriquez, E., and Pardo, J.M., "Analysis and Modeling of Emotional Speech in Spanish", Proc. of ICPhS, pp. 957-960, 1999.
3. Bulut, M., Narayanan, S. S., and Syrdal, A. K.: "Expressive Speech Synthesis Using a Concatenative Synthesizer", In. ICSLP-2002, pp. 1265-1268, 2002.
4. Schröder, M., Grice, M., "Expressing Vocal Effort in Concatenative Synthesis", Proc. of ICPhS, Barcelona, Spain, pp. 2589-2592, 2003.
5. Boersma, P., "Praat, a system for doing phonetics by computer", Glot International, vol. 5:9/10, pp. 341-345, 2001.

II. Beszédfelismerés

Spontán, nagyszótáras, folyamatos beszéd gépi felismerési pontosságának növelése beszélőadaptációval a MALACH projektben

Tüske Zoltán[1], Mihajlik Péter[1], Fegyó Tibor[1,2]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék
{tuske, mihajlik}@tmit.bme.hu

[2]Aitia International
tfegyo@aitia.ai

Kivonat: Cikkünkben bemutatjuk, hogy az MLLR (Maximum Likelihood Linear Regression) alapú beszélőadaptálás során a beszédfelismerési hatékonyság az adott spontán magyar nyelvű adatbázison jelentősen növekszik. Többféle módszert kipróbáltunk mind a felügyelt mind a felügyeletlen adaptálódás esetén is. A globális megoldás mellett regressziós osztályokon alapuló transzformációt is alkalmaztunk; felügyeletlen modellillesztés esetén a többszörös adaptálást is megvizsgáltuk. Továbbá folyamatos, nagyszótáras és spontán automatikus beszédfelismerővel kapott eredményekkel támasztjuk alá, hogy ha a szó alapú nyelvi modell helyett a magyar nyelvet pontosabban leíró morféma alapú modellezést alkalmazzuk, akkor a beszélőadaptálás által okozott javulás még szignifikánsabban jelentkezhet a felismerési hibaarányban.

1 Bevezetés

A nagyszótáras folyamatos beszédfelismerő rendszereket (LVCSR: Large Vocabulary Continuous Speech Recognition) nagyszámú beszélőtől összegyűjtött adatokon tanítják, hogy a tanító halmaz lefedje a különböző dialektusokat és beszédstílusokat. Az így nyert - általánosan használt rejtett Markov-modell (HMM: Hidden Markov-Modell) alapú - beszélőfüggetlen rendszer átlagosan jól teljesít minden beszélő esetén. Hátránya, hogy az átlagos beszédet modellezi, és az egyes beszélőkre nem optimális ez a modell. Kézenfekvő megoldás lehet, hogy a jobb felismerési pontosság elérése érdekében, adott beszélőre tanítsunk be egy teljesen új felismerőt. A baj csak az, hogy adott beszélőtől ehhez szükséges mennyiségű adatot szerezni meglehetősen nehéz feladat. A megoldás, hogy beszélőfüggetlen felismerőt az adott beszélőtől származó, a tanítóhalmazhoz képest kevés adatokkal nem újratanítjuk, hanem adaptáljuk, az egyes beszélőhöz „igazítjuk” az akusztikus modelleket. Ezzel egy köztes felismerőt kapunk, ami jobban teljesít, mint egy beszélőfüggetlen felismerő, de a felismerési hatásfok elmarad attól, amit egy elegendően sok tanítóadattal kapott egy beszélős tanítással érhetnénk el. Az adaptáló adatok mennyiségétől függően az adaptált rendszer valahol a két felismerő között helyezkedik el. Ennek a módszernek egyik

legnagyobb előnye a vele elérhető jobb szóhibaarány mellett, hogy attól a beszélőtől, amelyikhez illeszteni akarjuk a rendszert, nagyságrenddel kevesebb adat kell ahhoz képest, mintha egy teljesen új felismerőt tanítanánk be.

A szakirodalomban különböző módszereket találhatunk a rejtett Markov-modellek adaptálására. A beszédfelismerés területén publikációk sora bizonyítja, hogy a modellparaméterek adaptálóadatokra nézve ML (Maximum Likelihood) értelemben optimális lineáris transzformációja (MLLR) igen hatékony megoldás a modellillesztés megvalósítására.

A módszer áttekintése után, a kísérletekhez alkalmazott adatbázis, majd az összeállított felismerők ismertetése, végül az elért eredmények bemutatása következik.

2 Adaptálás MLLR-val

Lényege, hogy az adaptáló adatok alapján az akusztikus modellek minden egyes Gauss összetevőjének kovariancia mátrixát és várhatóérték vektorát lineáris transzformációval - utóbbit eltolással is - módosítja, így az adaptáló adatokhoz jobban illeszkedő akusztikus modelleket kaphatunk. A transzformált Markov-modell a kiinduló modellhez képest nagyobb valószínűséggel állítja elő az adaptáló adatokat. Minden állapot, minden komponensének várhatóértékére:

$$\underline{\hat{\mu}} = \underline{\underline{W}}\,\underline{\xi}$$

Ahol $\underline{\hat{\mu}}$ az új várhatóértékvektor, $\underline{\xi} = \begin{bmatrix} 1 & \underline{\mu}^T \end{bmatrix}^T$, $\underline{\mu}$ a régi várhatóértékvektor és $\underline{\underline{W}}$ *n*(n+1)* méretű egy állapothoz és annak egy komponenséhez tartozó transzformációs mátrix, *n* a jellemzővektorok dimenziója. Hasonlóan a kovariancia mátrixot is transzformálni kell a jobb illeszkedés érdekében.

$$\underline{\underline{\hat{\Sigma}}} = \underline{\underline{H}}\,\underline{\underline{\Sigma}}\,\underline{\underline{H}}^T$$

Ahol $\underline{\underline{\hat{\Sigma}}}$ az új kovariancia mátrix és $\underline{\underline{H}}$ a transzformációs mátrix.

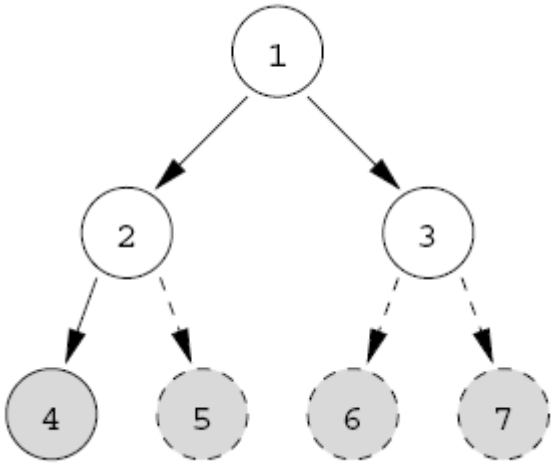
Az ideális transzformációs mátrixok, iteratív úton, EM (Expectation Maximalization) algoritmus segítségével határozódnak meg: először csak a várhatóérték transzformáció rögzített szórásmátrix mellett, majd csak a szórásmátrix transzformáció rögzített várhatóérték mellett. Az eljárás többszöri alkalmazásával egyre nagyobb valószínűséggel, pontosabban likelihood értékkel, fedi le az adaptáló adatokat a transzformált modell.

$$\ell\left(O|\tilde{\lambda}\right) \geq \ell\left(O|\hat{\lambda}\right) \geq \ell\left(O|\lambda\right)$$

Ahol O az adaptáló adatokból származó megfigyelési vektorok, ℓ a teljes likelihood érték adott λ modell paraméter mellett, O megfigyelés esetén. λ a transzformálatlan, $\hat{\lambda}$ a várhatóértékben transzformált, $\tilde{\lambda}$ pedig a két lépésben transzformált modellparamétereket jeleni adott Markov-modell esetén. A transzformációs paraméterek meghatározására vonatkozóan [1, 7] tartalmaz további részleteket.

Az egyes gauss komponensek transzformációs mátrixának meghatározására általában nem áll elegendő adat a rendelkezésre, ezért valamilyen távolság definíció alkalmazásával az egymáshoz közel eső komponensekből csoportokat vagy osztályokat képeznek, és erre a közös halmazra számoljuk a várhatóérték- és a szórás-transzformációs mátrixot. Az akusztikus modell gauss komponenseinek csoportosítására általánosan használt módszer a bináris regressziós fa. Ez a felülről lefele haladó módszer - kiindulásként az összes komponens egy csoportban van - azon komponenseket igyekszik egy osztályba gyűjteni, amelyek az akusztikus térben Euklédeszi távolság értelemben közel vannak egymáshoz, hasonló térrészt jellemeznek. A bemenő adat a beszélőfüggetlen felismerő akusztikus modellje, eredményként a komponensek általunk előírt számú csoportokra bontását adja vissza. A bináris fa továbbépítésének művelete a következő lépésekből áll:

- A felosztandó csoport szórását és várhatóértékét a benne szereplő komponensek alapján kiszámoljuk
- Két gyermek csoportot hozunk létre, a szülő várhatóértékétől ellentétes irányban eltávolodva.
- A szülő osztályt alkotó komponenseket a várhatóértékük alapján a közelebbi gyermek osztályhoz rendeljük.
- Újra számoljuk a gyermek csoportok várhatóértékeit
- A szülő komponenseket újra és újra hozzárendeljük a gyermek osztályokhoz, az újabb és újabb gyermek csoport várhatóértékek alapján, míg nincs már változás a hozzárendelésket illetően.



1. Ábra: Csoportosítás regressziós bináris fával. Adaptáló adatok mennyiségétől függően más-más szinten számoljuk transzformációs mátrixokat.

A regressziós fa segítségével az adaptáló adatok mennyiségétől függő adaptálást alkalmazhatunk (*1. ábra*). Kevés adat esetén csak globális transzformációra van lehetőség, ilyenkor egy mátrixszal transzformálunk minden várhatóértéket, illetve egy másikkal az összes szórást. Ha elegendően sok adatunk van, akkor külön transzformációt határozhatunk meg az egyes csoportokra (az ábrán csak a 4-es). Ha azonban egy csoportba kevés adaptáló adat jut (5, 6, 7), a fán feljebb lépve összevonhatunk több osztályt és így érhetjük el a szükséges mennyiségű adatok meglétét (2, 3).

Az adaptációhoz szükséges szöveges átiratok, amik segítségével az adaptáló adatokat az egyes akusztikus modellekhez rendelhetjük, kétféle forrásból származhatnak. Az adaptálás felügyelt, ha a pontos szöveges tartalommal rendelkezünk. Felügyeletlen adaptálás mellett szöveges átirathoz csak a beszélőfüggetlen felismerőtől kapott felismerési eredményt használhatjuk fel, nem tökéletes az átiratunk, így a szegmentálás sem lesz az. Ilyenkor azonban általában sokkal több adattal adaptálódhatunk, mert az összes felismertetendő felvételt bevonhatjuk a transzformációs mátrixok becslésébe.

3 Az adatbázis

A MALACH (Multilingual Access to Large Spoken Archives) projekt célja, hogy adott esetben holokauszt-túlélők beszámolóihoz könnyebben lehessen hozzáférni, és az adatbázis szöveges tartalom alapján kereshetővé váljon. A projekt 32 nyelvű, amiből a magyar nyelvű adatbázis több mint 2000 órányi felvételt tartalmaz. Eddig mintegy 31 órányi anyagnak készült el a szöveges átirata. A kísérleteket ezen az adatbázisrészleten végeztük. A felvételek 44,1 kHz-es mintavételezés mellett, általános

környezetben (általában a beszélő lakásán) készültek. A beszélők idősek, és esetenként erős akcentussal beszélnek.

A beszélőfüggetlen felismerőt 104 beszélőtől, beszélőnként tizenöt percnyi felvétellel tanítottuk (összesen kb. 26 óra). Tesztelésre további 10 beszélőtől származó, váltakozó hosszúságú, összesen 5 órányi hanganyagot használtunk. Ezt a teszthalmazt további részhalmazokra bontottuk. A beszélőfüggetlen (SI) részhalmaz egyik felét – a 15. perc utáni bemondások – a tanuló halmaz szempontjából ILLESZTETTnek neveztük, míg az első tizenöt perc felvételei az ILLESZTETLEN részhalmazba kerültek. A beszélőfüggetlen eredményeket az *1. táblázat*ban foglaltuk össze. Egy másik részhalmaza a teszthalmaznak a beszélőfüggő (SD) részhalmaz, ami egy férfi és egy női alany 1-1 órányi hanganyagát tartalmazza.

1. táblázat: Beszélőfüggetlen felismerési eredmények az SI teszthalmazokon morféma és szó nyelvi modellekkel

Nyelvi model	Illesztetlen		Illesztett	
	WER	LER	WER	LER
Morféma	55.94	28.17	51.07	24.53
Szó	56.01	28.26	50.90	24.39

4 A felismerési tesztek során alkalmazott akusztikus és nyelvi modellek

4.1 Nyelvi modellezés

A hagyományos szó alapú nyelvi modell mellett, a morféma alapú nyelvi modellel is lefuttattuk az adaptálási teszteket. Utóbbi jobban illeszkedik a morfémákban gazdag nyelvekhez, így a magyarhoz is. Az utóbbi esetben statisztikus morfológia elemző segítségével [6] bontottuk a szavakat morfémákra, a felmerülő problémát és azok egy lehetséges megoldását [2, 3] részletezi. Mindkét esetben trigram nyelvi modellt alkalmaztunk, melyeket a SRILM programmal [4] számoltunk.

4.2.a Beszélőfüggetlen akusztikus modellezés.

Az előző fejezetben bemutatott adatbázisból 26 órányi anyaghoz tartozó szöveges adatbázisból automatikus, szabályalapú konverzióval [5] készült el a szavak valamint a morfémák fonetikus átirata, az idegen eredetű és a hagyományos írásmód körébe tartozó szavak megfelelő fonéma sorozatra való átalakítása kivételszótár alapján történt. A mintegy 3000 általánosított, 3 állapotú balról jobbra haladó HMM-lel modellezett akusztikus trifón modelleket PLP jellemzőkből előállított vektorokkal tanítottuk.

4.2.b Beszélőadaptált akusztikus modellezés

Egy beszélő teljes adathalmazának 1/5-ét használtuk felügyelt adaptálásra, 4/5-öd részével teszteltünk. Felügyeletlen adaptálás esetén a felismertetendő 4/5-öd részen kapott beszédfelismerési eredményeket használtuk fel. Mindkét esetben globális adaptálással - ekkor minden gauss-összetevőn azonos a transzformáció -, és 32 levelű bináris regressziós fa építésével kapott csoportokra bontással is mértük az adaptáció utáni szóhibaarányt. Tapasztalataink szerint a gauss-összetevők 32 regressziós csoportba foglalása hozta a nem globális adaptáláskor a legjobb eredményeket. A felügyeletlen modellillesztés esetén többszörös adaptálással, valamint a globális és regressziós fán alapuló adaptálások kombinálásával igyekeztünk még pontosabbá tenni az akusztikus modellt.

5 Eredmények

A *2.* és *3. táblázat*ban a női beszélő adatain a beszélőfüggetlen és a beszélőfüggő felismerővel elért felismerési eredmények láthatóak. Az adaptálási eredmények mellett a relatív %-os javulást is feltüntettük a beszélőfüggetlen esethez képest. A táblázatból egyértelműen látszik, hogy a morfémaalapú nyelvi modell mellett sokkal hatékonyabb a javulás mind relatív, mind abszolút értékben is. Már a kiindulási eredmények is jobbak a morfémaalapú megközelítésnél (abszolút 5%-os szóhibaarány-különbség), ennek ellenére abszolút értékben is sokkal többet használt az adaptáció, tovább növelve a különbséget a szó- és morfémaalapú modellezés között, adaptálás után közel 10%-os a WER különbség. Felügyelt és felügyelet nélküli adaptálás mellett is relatív 28%-ot javított az adaptálás a betűhibaarányon. Kétszeri felügyelet nélküli adaptálással –először globális majd regressziós fás – sikerült a felügyelten adaptált felismerő felismerési hatásfokát elérni.

A férfi beszélő adatain elért felismerési hatásfok javulása az adaptálás hatására a *4.* és *5. táblázat*ban látható. Ennek a beszélőnek az adatain gyengébbek a morféma alapú felismerési eredmények a beszélőfüggetlen felismerővel (az előző beszélőhöz képest kb. átlag abszolút 2%-kal). A morfémaalapú nyelvtan itt is jobban teljesít, mint a szóalapú, az adaptálásnak köszönhetően ennek a beszélőnek is kb. relatív 25%-kal sikerült javítani a betűhibaarányán. A kétszeres adaptációval elért legjobb eredmény picivel elmarad a felügyelt tanítás mögött. A pontatlanabb felismerésből eredően - a második beszélő szempontjából rosszabb az akusztikus modellezés - adaptáció után a két beszélő közötti hibaarány különbség 2%-ról 4%-ra nőtt.

2. Táblázat: Az adaptálás nélkül és az adaptálással elért szóhibaarányok (női beszélő)

A felismerő fajtája	Adaptálási technika	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
		WER [%]	LER [%]	WER [%]	LER [%]
beszélő független	nincs	51,86	21,70	46,07	18,79
felügyelten adaptált	globális	47,18	18,33	38,93	14,90
	regressziós fás	45,84	16,87	36,87	13,48
1x felügyelet nélkül adaptált	globális	46,66	17,96	38,62	14,70
	regressziós fás	46,23	17,32	37,56	13,84
2x felügyelet nélkül adaptált	2x globális	46,56	18,05	38,88	14,62
	globális majd regressziós fás	45,99	17,06	36,58	13,48
	2x regressziós fás	46,41	17,23	37,44	13,86

3. Táblázat: Adaptálással a szóhibaarányban elért relatív %-os javulás a beszélőfüggetlen esethez képest (női beszélő)

A felismerő fajtája	Adaptálási technika	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
		ΔWER_{rel} [%]	ΔLER_{rel} [%]	ΔWER_{rel} [%]	ΔLER_{rel} [%]
felügyelten adaptált	globális	9,02	15,53	15,50	20,70
	regressziós fás	11,61	22,26	19,97	28,26
1x felügyelet nélkül adaptált	globális	10,03	17,24	16,17	21,77
	regressziós fás	10,86	20,18	18,47	26,34
2x felügyelet nélkül adaptált	2x globális	10,22	16,82	15,61	22,19
	globális majd regressziós fás	11,32	21,38	20,60	28,26
	2x regressziós fás	10,51	20,60	18,73	26,24

4. Táblázat: Az adaptálás nélkül és az adaptálással elért szóhibaarányok (férfi beszélő)

A felismerő fajtája	Adaptálási technika	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
		WER [%]	LER [%]	WER [%]	LER [%]
beszélő független	nincs	49,09	23,73	48,00	23,41
felügyelten adaptált	globális	46,31	21,21	43,96	19,78
	regressziós fás	44,44	19,11	40,76	17,54
1x felügyelet nélkül adaptált	globális	46,92	21,81	44,07	20,07
	regressziós fás	43,94	19,67	41,34	18,36
2x felügyelet nélkül adaptált	2x globális	46,65	21,70	43,82	19,99
	globális majd regressziós fás	44,57	20,13	41,55	18,48
	2x regressziós fás	43,69	19,60	40,73	18,16

5. Táblázat: Adaptálással a szóhibaarányban elért relatív %-os javulás a beszélőfüggetlen esethez képest (férfi beszélő)

A felismerő fajtája	Adaptálási technika	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
		ΔWER_{rel} [%]	ΔLER_{rel} [%]	ΔWER_{rel} [%]	ΔLER_{rel} [%]
felügyelten adaptált	globális	5,66	10,62	8,42	15,51
	regressziós fás	9,47	19,47	15,08	25,07
1x felügyelet nélkül adaptált	globális	4,42	8,09	8,19	14,27
	regressziós fás	10,49	17,11	13,88	21,57
2x felügyelet nélkül adaptált	2x globális	4,97	8,55	8,71	14,61
	globális majd regressziós fás	9,21	15,17	13,44	21,06
	2x regressziós fás	11,00	17,40	15,15	22,43

A *6. táblázat*ban a két beszélőn elért átlagos legjobb felismerési eredményeket foglaltuk össze. A legjobb felismerési eredményeket morféma alapú modellel kaptuk. Elmondhatjuk, hogy ha nem ismert az adaptáló adatok pontos tartalma, akkor érdemesebb az akusztikus modelleket először csak globálisan adaptálni a félreismerésekkel zajosított szöveges átiraton, kiküszöbölendő a hibásan felismert fonémasorozatokból eredő félretranszformálást. Majd az így nyert pontosabb felismeréssel, most már regressziós fával továbbtranszformálva az akusztikus modelleket, tovább növelhető az illeszkedés.

6. Táblázat: Az átlagos szóhibaarány adaptálással és nélküle, relatív javulás a beszélőfüggetlen esethez képest

A felismerő fajtája	Szóalapú nyevli modell		Morfémaalapú nyelvi modell	
	WER [%]	ΔWER_{rel} [%]	WER [%]	ΔWER_{rel} [%]
beszélő független	50,47	-	47,03	-
felügyelten 32 levelű regressziós fával adaptált,	45,14	10,57	38,81	17,48
2x felügyelet nélkül adaptált globális, majd 32 Levelű regressziós fával	45,28	10,29	39,06	16,94

6 Összefoglalás

Elmondhatjuk, hogy beszélőadaptálással sokkal hatékonyabbá tettük a beszédfelismerő-rendszert. Pontosabb felismerési eredmények mellett az adaptálás jobban teljesít, a hagyományos szó nyelvi modell helyett morféma nyelvi modellt alkalmazva kb. másfélszer akkora relatív javulást értünk el a szó-hibaarányban.

Bibliográfia

1. M. J. F. Gales, Maximum Likelihood Linear Transformations for HMM-based Speech Recognition, Computer Speech and Language, Vol. 12, pp. 75-98, 1998
2. Mihajlik, P., Fegyó T., Németh B., Tüske Z., Trón V., Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages – Hungarian ASR for the MALACH Project, TSD 2007, Pilsen, Czech Republik
3. Mihajlik, P., Fegyó, T., Tüske, Z., and Ircing, P., "A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian" – In Interspeech 2007. Antwerp, Belgium, August 27-31, (2007).
4. Stolcke, A., "SRILM – an extensible language modeling toolkit", In Proc. Intl. Conf. On Spoken Language Processing, Denver (2002) 901–904
5. Szarvas M., Fegyó, T., Mihajlik, P., and Tatai, P., "Automatic Recognition of Hungarian: Theory an Practice", International Journal of Speech Technology, 3:277-287, December, 2000.
6. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy. and Varga, D., "Hunmorph: open source word analysis", In Proc. ACL 2005 Software Workshop, (2005) 77–85
7. Young S., Ollason, D., Valtchev, V., and Woodland, P., The HTK Book (for HTK version 3.2.1), Cambridge University Engineering Department, 2002.

Diktálórendszer pontosságának és hatékonyságának vizsgálata a keresési téren alkalmazott vágási technikák függvényében

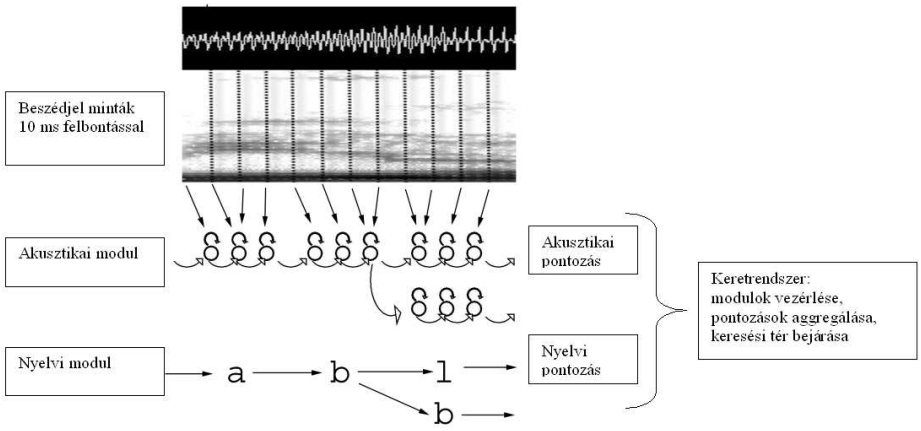
Bánhalmi András, Paczolay Dénes, Tóth László

MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Aradi vértanúk tere 1.
{banhalmi,pdenes,tothl}@inf.u-szeged.hu

Kivonat Folyamatos beszéd felismerése esetén a beszédjelhez illeszthető szósorozatok száma exponenciálisan nő a felvétel hosszával. Ezért a diktálórendszerek hatékonysága szempontjából kulcsszerepe van a különböző, a keresési teret redukáló vágási technikáknak, illetve kiértékelést gyorsító trükköknek. A keresési tér vágásával elért sebességnövekedés könnyen a felismerési pontosság rovására mehet, ezért a módszerek paramétereinek beállításakor meg kell találni a megfelelő egyensúlyt a hatékonyság és a pontosság között. Cikkünkben bemutatjuk, hogy az általunk fejlesztett felismerő hogyan reprezentálja a keresési teret, az azt nagyban meghatározó nyelvi komponenst, továbbá hogy maga a keresés hogyan történik. Ismertetjük, hogy a keresés során milyen vágási technikákat alkalamazunk, majd konkrét felismerési teszteken keresztül megvizsgáljuk, hogy különböző paraméterértékek mellett ezek hogyan befolyásolják a futási időt és a felismerési pontosságot.

1. Motiváció

Folyamatos diktálórendszerek konkrét technikai megvalósításával kapcsolatosan viszonylag kevés szakirodalom hozzáférhető (néhány elterjedtebb módszer leírása itt megtalálható: [1], [2], [3]). Ennek talán az lehet az oka, hogy a keresési teret bejáró algoritmusokon és különféle hatékony számolási technikákon nagyban múlik az, hogy egy diktálórendszer mennyire lesz gyors és mennyire lesz pontos. Így ezek a technikák piaci szempontból nagy jelentőséggel bírnak. Ebben a cikkünkben egy részletes leírást adunk a saját fejlesztésű diktálórend-szerünk felépítéséről, az egyes modulok processzorigényéről, különböző gyorsítási lehetőségekről, és a keresési tér általunk alkalmazott vágási módszereiről. Elemezzük, hogy az egyes vágási technikák hogyan befolyásolják a felismerési pontosságot és a processzorigényt.



1. ábra. Diktálórendszer moduláris szerkezete

2. A diktálórendszer felépítése

Egy diktálórendszer alapvető feladata abban áll, hogy egy mikrofonba bemondott beszédjelet szöveges információvá alakítson át. Ehhez – a megvalósítás szintjén – két, egymástól eltérő szerepű eszköz áll rendelkezésére. Az egyik a nyelvi modul, aminek a feladata az, hogy egy szósorozathoz pontértéket rendeljen aszerint, hogy az adott szósorozat mennyire valószínű. Ennek a feladatnak a megoldására többféle modellt is javasoltak már. Egy nyelvi modell annál jobb, minél inkább „kiszűri" azokat a szósorozatokat, amik csak nem tipikus diktálás mellett fordulhatnak elő, de nem pontozza le azokat a szósorozatokat, amik ha nem is gyakran, de előfordulhatnak a diktálás során. A másik fontos eleme egy diktálórendszernek az akusztikai modul, aminek a feladata a különböző beszédhangok modellezése valamilyen gépi tanuló algoritmus segítségével. Diktáláskor az akusztikai modul értékeli azt, hogy a bediktált jel egy darabja mennyire tartozik bele egy beszédhang-osztályba. Egy valós idejű diktálórendszer az említett két modul pontozását kombinálva rangsorolja a lehetséges beszédhangsorozatokat, ezekből egy keresési teret építve folyamatosan a diktálás ideje alatt. Hogy ez a tér ne növekedjen exponenciálisan, és így a rendszer hatékony tudjon maradni, különböző heurisztikus vágási technikákat kell alkalmazni. A rangsorolást, a keresési tér felépítését és a vágásokat a nyelvi és akusztikai modulokra épülő keretrendszer valósítja meg (ld. 1. ábra). Ebben a fejezetben részletesen írunk a saját rendszerünkbe beépített modulokról.

2.1. Nyelvi modul

Egy nyelvi modul alapfeladata egy olyan függvénynek a megtanulása és hatékony kiértékelése, ami egy szósorozathoz megad egy pontértéket. A nyelvi modul a mi

implementációnkban a pontértékek hozzárendelése mellett az adott területhez (domain) tartozó lehetséges szósorozatok „bejárására” (generálására) is képes, ezzel támogatva a keresést.

Ezen felül - a mi felfogásunkban - a nyelvi modul nem lehetséges szósorozatokat, hanem lehetséges beszédhangsorozatokat generál (ugyanazon szósorozathoz többféle beszédhangsorozat is rendelhető). A mi megvalósításunkban a nyelvi modulnak meg kell adnia azokat a beszédhangokat, amelyek a modul egy adott állapotában következhetnek. Emellett a modul megadja még a lehetséges folytatások bizonyos jellemzőit is, amik arról adnak információt, hogy egy szó végére értünk-e, folytatható-e tovább a nyelvi kiértékelés, valamint megadja az egyes lehetőségek nyelvi pontozását (valószínűségét).

Egy példán szemléltetve: a nyelvi modul állapota legyen ("klinikai adatok", "k l i n i k a j i j a d a t o k _ v á l"), aminek a jelentése az, hogy az eddig bemondott teljes szavak a "klinikai adatok" voltak, a szavak között hasonulást tételeztünk fel a beszédhangsorozat átiratában, és a harmadik, még ismeretlen szóból a "v á l" beszédhangsorozatig jutottunk el. A nyelvi modul ebben az állapotában visszaadja a lehetséges folytatásokat (pl. "t", "l", "o", "a", stb.), a hozzájuk tartozó pontértéket, valamint szóvégre és nyelvtanvégre hamisat ad vissza. Egy ilyen függvény többféleképpen is megvalósítható, és sokféle nyelvi információt felhasználhat a pontérték megadásához. A mi modulunk három nyelvi modellel rendelkezik jelenleg. Ezek a szó N-gram és ennek simított változatai [4], a környezetfüggetlen nyelvtan, és az MSD-kód alapú csoport N-gram (erről részletesebben [5]). Mivel szó N-gram használatakor igen nehézkes megadni számokat, dátumokat, neveket, ezért a szó N-gram megvalósításunkban lehetőség van szavak helyett szavak egy-egy halmazát megadni, és ezek a halmazok leírhatók egyszerű szó-listával, illetve szabályokkal is. A nyelvi modul a következő lehetőségeket tartalmazza:

- Egy szótár megadása kötelező.
- Szó N-gramok adhatók meg.
- Hasonulási szabályok adhatók meg.
- A szótár szavai csoportokba sorolhatók.
- A csoportokra környezetfüggetlen nyelvtan adható meg.
- A csoportokra N-gram adható meg.

A konkrét megvalósítás oldaláról nézve, attól függően, hogy az előbbi lehetőségek közül melyeket használjuk nyelvi modellezésre, különböző struktúrákat épít fel a nyelvi modul. Legegyszerűbb esetben egy prefix-fa épül fel, amelynek a leveleit visszakötjük a fa elejéhez. Hasonulási szabályok használata esetén már egy ennél bonyolultabb gráfot állítunk elő, amelyben az egyes hasonulási lehetőségekhez tartozó ágak később egyesülnek. Környezetfüggetlen nyelvtan használata esetében pedig egy összetett, prefix-fákból álló gráfot építünk fel, ami utána minimalizálásra kerül. A minimalizálásnál az a célfüggvény, hogy azonos szósorozat illetve hozzá tartozó beszédhangsorozat ne jelenjen meg a gráfban egynél több

ágon. Ez alól egy kivétel van, ami akkor jelentkezhet, ha egy szó több csoportba is tartozhat. Ennek a felépítésnek az alapvető oka, illetve célja, hogy az aktuális hipotézislistában (ami beszédhangsorozatokból áll) kétszer ne jelenhessen meg ugyanaz a beszédhangsorozat, azaz a nyelvi modul ne generálhasson azonos hipotéziseket (mert ezek elvehetik a helyet a többi hipotézistől).

A nyelvi modult úgy valósítottuk meg, hogy amilyen hamar csak lehet, megadja a szó- valamint csoportsorozathoz tartozó pontértéket. Ez azt jelenti, hogy amikor már egyértelmű, hogy melyik szó (illetve csoport) tartozik a beszédhangsorozathoz (még nem feltétlenül értük el a szó végét), akkor a nyelvi modul már megadja a megfelelő pontértéket. Ennek az ún. előrehozott kiértékelési módszernek továbbfejlesztéseként olyan technikát is alkalmazunk, ami már az egyértelművé válás előtt egy (az ágakhoz tartozó maximummal) becsült értéket, majd később korrekciós értékeket ad meg pontozáskor. Annak a fontosságát, hogy a nyelvi modul minél hamarabb megfelelően értékelni tudja a beszédhangsorozatokat, egy korábbi cikkünkben elemeztük [5].

2.2. Akusztikai modul

A diktálórendszerünkbe az általánosan elterjedt Rejtett Markov Model (HMM) alapú akusztikus modult építettük be. Minden beszédhangra egy-egy balról-jobbra struktúrájú HMM-et tanítunk.

Diktáláskor egy-egy beszédhangsorozatnak egy-egy folyamatosan felépülő HMM lánc felel meg. Beszédfelismerés során a nyelvi modul lekérdezésével bővítjük a láncokat. Ezzel párhuzamosan az akusztikai modul feladata az, hogy a feldolgozásban soron következő beszédjel-szelet alapján az aktuális (a láncok végén szereplő) HMM-ek valószínűségi adatait megfelelően módosítsa.

A HMM-ek minden állapotához hozzá van rendelve egy valószínűségi eloszlás (GMM, Gaussian Mixture Modell, Gauss eloszlások súlyozott összege). Alapesetben az akusztikus modell a HMM-ek minden elérhető állapotára kiszámolja a következő értéket:

$$p(x) = \sum_{i=1}^{M} w_i \frac{1}{2\pi \left|\Sigma_i\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}, \quad (\sum_{i=1}^{M} w_i = 1),$$

ahol M a Gauss-komponensek száma, w_i a komponensek súlya, Σ_i és μ_i a komponensekhez tartozó kovarianciamátrix és középérték-vektor, és x a jellemzővektor. A gépi számábrázolás korlátozottsága és a számítások egyszerűsítése érdekében az előző értéknek a logaritmusát szokás kiszámítani. Ehhez a számításhoz a logaritmikus aritmetikából a következő, összeadásra vonatkozó képletet használhatjuk, így végig logaritmikus aritmetikában maradva elkerülhetjük a lebegőpontos alulcsordulást:

$$\log(a+b) = \log(e^{\log a} + e^{\log b}) = \log(a) + \log(1 + e^{\log(b) - \log(a)})$$

További egyszerűsítésként az összegnek a maximummal való közelítése ajánlott (a két érték között nincsen lényeges eltérés a gyakorlatban):

$$p(x) \approx \max_i(\log(w_i \frac{1}{2\pi|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}))$$
$$= \max_i(\log(\frac{w_i}{2\pi|\Sigma_i|^{1/2}}) - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))$$

A kifejezés első tagja egy x-től független, előre kiszámítható konstans. A második taggal kapcsolatosan a beszédfelismerésben általánosan elterjedt gyakorlat (a számítások csökkentése érdekében), hogy a négyzetes kovariancia mátrixot diagonálisnak tételezzük fel. Így a következő számítást kell csak elvégeznünk:

$$\frac{1}{2}\sum_j \frac{(x_j - (\mu_i)_j)^2}{(\Sigma_i)_{jj}}$$

Ezt a számítást nem kell teljes egészében minden - az adott állapothoz tartozó - Gauss-komponensre elvégezni: ha az összegzés során az érték az addigi maximális érték alá csökkent, akkor a számítást nem fejezzük be. Ez a módszer még tovább gyorsítható úgy, hogy a Gauss-komponensek kiszámításának sorrendjét előre lerendezzük azok súlya szerinti csökkenő sorrendbe.

Mindezek mellett egy akusztikai alapú büntető, illetve szűrő eljárás használatát is javasoljuk, amellyel még nem találkoztunk a szakirodalomban. Az ún. egyosztályos statisztikai modellekben egy – a pozitív tanuló példákon számított – ***konfidenciaérték*** alapján sorolják be a tesztpéldákat pozitív példának (osztályba tartozónak), illetve negatív példának (nem az osztályba tartozónak). Rejtett Markov Modell alapú beszédfelismerő modellek esetében - a modellek tanításának kiegészítéseként - mi is meghatározunk minden HMM minden Gauss komponenséhez egy-egy konfidencia értéket. Ha a beszédfelismerés során az akusztikus modell által számított pontérték a konfidenciaértéknél kisebb lesz, akkor a számítást befejezzük, és egy megfelelően kicsi ponttal értékeljük a beszédjelet az adott Gauss-komponensre nézve. A konfidenciaértéket a pozitív tanulóhalmaz legkisebb pontértékéhez viszonyítva adjuk meg, ez nálunk egy 1,5-szörös szorzással adódik (logaritmikus aritmetikában számolva).

3. Keretrendszer a keresési tér felépítésére

Leegyszerűsítve, a keretrendszer feladata abban határozható meg, hogy fel kell építenie a hipotéziseknek egy olyan terét, amelyet a nyelvi modul megenged, és az egyes hipotézisekhez pontértéket kell rendelnie a nyelvi és az akusztikai pontozás alapján. A cél a legnagyobb ponttal rendelkező hipotézis(ek) megtalálása. Saját megvalósításunkban - alapjaiban - egy Viterbi N-best típusú keresést használunk [6], ami azt jelenti, hogy a keresési térnek mindig csak a legjobb N hipotézisét terjesztjük ki. Ennél az egyszerű módszernél azonban több ponton bővül a mi megközelítésünk, amit ebben a fejezetben részletesen bemutatunk.

3.1. A hipotézis fogalma

Egy hipotézist a következő adatok határoznak meg: (időpont, csoportsorozat, szósorozat, beszédhangsorozat, aktuális HMM azonosító, az aktuális HMM állapotaihoz tartozó valószínűségi értékek). Bevezetjük a következő jelölést:

$$H(t, [C_1 \dots C_k], [W_1 \dots W_k], [Ph_1 \dots Ph_n], \Theta_{Ph_n}, [p_0, p_1, p_2, p_3, p_4])$$

Itt t az időpontot, $[C_1 \dots C_k]$ a csoportsorozatot, $[W_1 \dots W_k]$ a szósorozatot, $[Ph_1 \dots Ph_n]$ a beszédhangsorozatot, Θ_{Ph_n} az utolsó beszédhanghoz tartozó HMM-et, és $[p_0, p_1, p_2, p_3, p_4]$ az aktuális HMM állapotaihoz tartozó valószínűségi értékeket jelöli. Az egyszerűbb jelölés kedvéért feltettük, hogy a HMM-ek 3 valódi állapottal, és egy kezdő valamint egy végállapottal rendelkeznek.

Két hipotézist szósorozatban ekvivalensnek nevezünk egy időpontban, ha csoportsorozataik és szósorozataik megegyeznek. Két hipotézist egyesíthetőnek nevezünk egy időpontban, ha szósorozatban ekvivalensek, beszédhangsorozataik hasonulástól eltekintve megegyeznek, és az utolsó beszédhangok azonosak (erről a nyelvi modulnak kell tudnia információt adni).

3.2. Hipotézisek kiterjesztései

Hipotéziseket kétféleképpen terjesztünk ki.

1. Hipotézisek akusztikai kiterjesztése. Ekkor a hipotéziseket meghatározó adatok közül csak az időpont, és az aktuális HMM-ek állapotaihoz tartozó valószínűségi értékek változnak.

$$\begin{gathered}\mathrm{H(t, [C_i], [W_i], [Ph_i], \Theta_{Ph_n}, [p_0, p_1, p_2, p_3, p_4])} \\ \Downarrow \\ \mathrm{H'(t+1, [C_i], [W_i], [Ph_i], \Theta_{Ph_n}, [-\infty, p'_1, p'_2, p'_3, p'_4])}\end{gathered}$$

   Itt az új $\mathrm{p'_i}$ értékek a régiekből a szokványos Viterbi algoritmussal [7] számíthatók.
2. Hipotézisek nyelvi kiterjesztése. Ebben az esetben több új hipotézist is kaphatunk, hiszen egy beszédhangsorozat többféleképpen folytatódhat. Az új hipotézisek adatai közül az időpont nem változik, a csoportsorozat, szósorozat, beszédhangsorozat a nyelvi modell által visszaadott módon bővülhet, az aktuális HMM azonosító a nyelvi modell által megadott következő beszédhang szerint változik, és az aktuális HMM állapotaihoz tartozó valószínűségi értékek felveszik kezdeti értéküket.

$$\begin{gathered}\mathrm{H(t, [C_i], [W_i], [Ph_i], \Theta_{Ph_n}, [p_0, p_1, p_2, p_3, p_4])} \\ \Downarrow \\ \left\{ \begin{array}{c} \mathrm{H'(t, [C_i][C'_1], [W_i][W'_1], [Ph_i][Ph'_1], \Theta_{Ph'_1}, [p'_{1_0}, -\infty, -\infty, -\infty, -\infty])} \\ \vdots \\ \mathrm{H'(t, [C_i][C'_N], [W_i][W'_N], [Ph_i][Ph'_N], \Theta_{Ph'_N}, [p'_{N_0}, -\infty, -\infty, -\infty, -\infty])} \end{array} \right\}\end{gathered}$$

Itt $\mathrm{p'}_{k_0} = p_4 + \log(NyelviPont(k))$, és a szósorozat, illetve a csoportsorozat nem feltétlenül bővül. A nem kezdőállapotok valószínűségei felveszik a $-\infty$ kezdeti értékeket (a logaritmikus aritmetika miatt, $log(0)$). Nyelvi kiterjesztés nyilvánvaló feltétele az, hogy $p_4 > -\infty$, tehát az eredeti hipotézis aktuális HMM-jének végállapotához egy nullánál nagyobb valószínűség tartozzon.

3.3. A keresési tér felépítése és vágása

Először megadjuk az általunk használt algoritmus vázát (1. táblázat), majd részletesen kifejtjük az egyes eljárások működését.

Inicializálás(), $t = 0$
H_t=NyelviHipotézisKiterjesztés(KezdőHipotézis)
Ismételd
Jellemzők=BeszédjelJemmezővektorSzámítás()
H_acc=AkusztikaiHipotézisKiterjesztés(H_t, Jellemzők)
H_grm= NyelviHipotézisKiterjesztés(H_acc)
H_{t+1}= HipotézisekEgyesítése(H_acc, H_grm)
$t = t + 1$
Vége

1. táblázat. A hipotézisteret felépítő algoritmus váza

Jellemzővektor-számítás: a beszédfelismerésben szokásos Mel Frequency Cepstral Coefficients (MFCC), energia, Δ, $\Delta\Delta$ jellemzőket tartalmazó 39 dimenziós vektor kiszámítása (alapesetben) 10 ms-onként történik [2].

Akusztikai hipotézis-kiterjesztés: a jellemzővektornak megfelelően, Viterbi kiértékeléssel [7] módosítja a hipotézisek aktuális HMM-jének állapotaihoz tartozó valószínűségi értékeket. A következő algoritmus határozza meg, hogy melyik hipotéziseket terjesztjük ki:

- A hipotézisket rendezzük $max(p_0, \ldots, p_4)$ érték szerint csökkenő sorrendbe.
- Terjesszük ki az első hipotézist, legyen az új hipotézis H1, és az új való-színűségi értékek $p'_0, \ldots, p'_4$. Vegyük az $M = max(p'_0, \ldots, p'_4) - AkusztikaiKüszöb$ értéket, ahol az ***AkusztikaiKüszöb*** egy előre definiált konstans érték.
- Terjesszük ki sorrendben a többi hipotézist is, legfeljebb ***MaxAkusztikai-Kiterjesztés*** számút. Az új hipotézist eldobjuk, ha a hozzá tartozó valószínűségi értékekre: $max(p_0, \ldots, p_4) < M$.

Nyelvi hipotézis-kiterjesztés: lekérdezzük a hipotézisek lehetséges folytatásait, és azokkal új hipotéziseket hozunk létre. Itt a következőképp járunk el:

- A hipotézisket rendezzük a p_4 értékük szerint csökkenő sorrendbe.
- Kiterjesztünk legfeljebb ***MaxNyelviKiterjesztés*** számú hipotézist.
- További feltételként, megállunk a kiterjesztéssel, ha az új hipotézisek száma elérte a ***MaxÚjNyelviHipotézis*** számot.

– Harmadikként, vannak olyan hipotézisek, amelyek teljesen végigmondott szavakhoz tartoznak. Az ilyen hipotézisekből legfeljebb ***MaxSzóvég*** számú hipotézist terjesztünk ki.

A hipotézisek egyesítése: a kétféle kiterjesztéssel kapott hipotézisek között gyakran ekvivalens hipotézispárok jönnek létre. Például, az "a l m" hipotézist kiterjeszthetjük "a l m a"-ra, azonban ez a kiterjesztés már korábban is megtörténhetett, így már létezhet egy "a l m a" hipotézisünk. Az ekvivalens hipotézisek ismétlődését azonban mindenképpen el kell kerülnünk. Mivel a rendszerünkben a HMM-eknél szokásos Viterbi kiértékelést használjuk, ezért az ekvivalens hipotézisek összevonása egzakt módon megtehető a következő, egyszerű módon. Tegyük fel, hogy a nyelvi kiterjesztés után rendelkezünk egy olyan H_n hipotézissel, ami egyesíthető (ld. 3.1. fejezet) egy, az akusztikai kiterjesztés után kapott H_a hipotézissel. Az egyesített hipotézist a következőképpen kapjuk:

$$\begin{gathered}\mathrm{H_n(t,[C_i],[W_i],[Ph_i],}\Theta_{\mathrm{Ph_n}}\mathrm{,[p_0,-\infty,-\infty,-\infty,-\infty])}\\ +\\ \mathrm{H_a(t,[C_i],[W_i],[Ph_i],}\Theta_{\mathrm{Ph_n}}\mathrm{,[-\infty,p_1,p_2,p_3,p_4])}\\ \Downarrow\\ \mathrm{H(t,[C_i],[W_i],[Ph_i],}\Theta_{\mathrm{Ph_n}}\mathrm{,[p_0,p_1,p_2,p_3,p_4])}\end{gathered}$$

Tehát, az egyesített hipotézis – a HMM Viterbi kiértékelésének megfelelően – megegyezik a H_a hipotézissel, de a p_0 értékét a H_n hipotézistől kapja. Az egyesítés során szintén megadunk egy – keresési teret vágó – paramétert, ami az egyesített hipotézisek maximális száma (***HalomMéret***).

Nyelvi hipotézis-kiterjesztés korlátozása: egy további - keresési teret szűkítő - paraméterrel szabályozni tudjuk a nyelvi kiterjesztés gyakoriságát: megadhatjuk, hogy hány iterációnként történjen nyelvi kiterjesztés (***NyelviIterSzám***). Ennek a paraméternek az alapértéke 1, azonban az értékét növelve a be-szédfelismeréshez felhasznált processzoridő nagymértékben csökkenthető (a pontosság némi romlása mellett).

4. Kiértékelés és eredmények

Mielőtt elemeznénk a mérési eredményeket, röviden írunk a rendszerünk implementációjáról algoritmikus szemszögből nézve. A hipotézisek folyamatos kiterjesztése nálunk egy saját fejlesztésű multistack rendszerben valósul meg. Az egyes stack-ek megvalósítása ún. "hybridlist" [8] adatszerkezetre épül, aminek az oka az, hogy olyan változó méretű rendezett adatszerkezetre volt szükségünk, amelynél a keresés, beszúrás és a törlés $O(logN)$ időkomplexitású művelet. Ilyen adatszerkezetek közül - tudomásunk szerint - az említett algoritmus a leggyorsabb. A beszédfelismerő sebessége szempontjából még egy fontos momentum, hogy egy saját memóriakezelőt is fejlesztettünk a minél hatékonyabb memóriahasználat céljából.

4.1. Tanítás és tesztelés

Az akusztikai modul tanításához a Magyar Referencia Beszédadatbázist (MRBA [9]), egy nagyméretű, szegmentált, 332 beszélő által bemondott hanganyagot tartalmazó adatbázis használtuk. A beszédkorpuszban meglévő beszédhangokat 33 csoportra osztottuk fel a tanításhoz. Csoportonként egy-egy három állapotú és állapotonként három Gauss-komponenssel rendelkező balról-jobbra strukturájú HMM-et tanítottunk.

A tesztek során végig szó 3-gram nyelvi modellt használtunk. Ezen modellek létrehozásához egy pajzsmirigy-szcintigráfiás leletekből álló szövegkorpuszt használtunk. Ebből a szövegkorpuszból három, különböző méretű szótárral rendelkező nyelvtant hoztunk létre (500, 1200, 1900 szóalak), mindegyikük részhalmazként tartalmazta a tesztadatbázis bemondásainak szóalakjait (azaz nem volt szótáron kívüli szó teszteléskor).

A tesztelés egy olyan adatbázison történt, amely 100 szcintigráfiás orvosi lelet bemondását tartalmazza (3 nőtől és 2 férfitól), és összesen mintegy 1000 mondatot tesz ki. A tesztek során AMD Athlon 2 GHz processzorral és 2 GB memóriával rendelkező számítógépet használtunk.

A kísérletek során processzoridő-igényt és felismerési pontosságot mértünk. A processzoridő-igényt a táblázatokban másodpercben, valamint Real-Time Factorban (RTF) adjuk meg, amely a felhasznált processzoridő és az összes hanganyag időtartamának (5325 mp.) a hányadosa. A szószintű felismerési pontosságot (accuracy, vagy word recognition rate, WRR) a szokásos eljárással mértük.

4.2. Felismerési pontosság és processzorigény különböző vágásoknál

Elsőként azt vizsgáltuk meg, hogy hogyan függ a felismerési pontosság és a processzorigény a ***MaxÚjNyelviHipotézis*** paraméter értékétől (ami az új hipotézisek maximális számát adja meg nyelvi kiterjesztéskor). A 2. táblázatban foglaltuk össze a tesztek edményeit, melyek során 500 szavas szótárat használtunk 260 küszöbértékű akusztikai vágás mellett. Az eredmények szerint kielégítő eredményt ad (időben és pontosságban is), ha 200-400 új hipotézisben maximáljuk a nyelvi kiterjesztés eredményét. A táblázat alapján – eleinte – ha megduplázzuk a vágási paramétert, akkor hozzávetőlegesen a felére csökken a szófelismerés hibája, azonban később a pontosság nem növelhető tovább. A szótár méretétől függően azt a beállítást célszerű választani, amikor már lényegesen nem változik a hiba, de a rendszer még valós időben működik.

Max. Új Ny. H.	50	100	200	400	600	800	1000
Pontosság	68.8%	84.6%	94.0%	97.0%	97.9%	97.7%	97.7%
Idő (RTF)	0.08	0.14	0.30	0.74	1.20	1.55	1.74

2. táblázat. Az nyelvi kiterjesztés során kapott hipotézisek számára adott korlát hatása a pontosságra és a processzoridő-igényre

A ***NyelviIterSzám*** paramétert – amellyel azt szabályozhatjuk, hogy menynyi iterációnként történjen nyelvi kiterjesztés – a ***MaxÚjNyelviHipotézis*** paraméterrel együtt vizsgáltuk. A 3. táblázat szerint a felhasznált processzoridő folyamatosan csökken, ahogy a nyelvi kiterjesztés (és ezzel együtt a hipotézisek egyesítésének művelete) ritkábban fut le. Azonban az is látszik, hogy a 200/1 -es eset időben jobb, mint a 400/2-es, azaz átlagosan jobban megéri inkább gyakran kevesebb hipotézist kiterjeszteni, mint ritkábban, de többet. Ez a kijelentés viszont csak az öt ember bemondásait tartalmazó tesztadathalmazon számolt átlagos eredményekre igaz. A mi tapasztalataink szerint sokszor megéri a nyelvi kiterjesztést ritkábban elvégezni, például olyan esetekben, amikor a beszélő jól artikuláltan és normál, vagy lassú tempóban beszél (két megfelelő tesztsze-mélyre ld. 4. táblázat). Egy másik eset, amikor jól alkalmazható ez a fajta vágás (ezt most nem vizsgáljuk), amikor a beszédjelből a 10 ms-os alapértéknél sűrűbben nyerünk ki jellemzővektorokat. Ebben az esetben a nyelvi kiterjesztést nem célszerű azonos gyakorisággal végrehajtani.

# Új Ny. H.	200			400			600		
Kit. Iter.	1	2	3	1	2	3	1	2	3
Pontosság	94.0%	88.8%	78.0%	96.4%	93.6%	84.8%	96.3%	94.0%	85.8%
Idő (RTF)	0.30	0.18	0.15	0.64	0.35	0.28	0.75	0.41	0.32

3. táblázat. Eredmények a nyelvi kiterjesztés gyakoriságának és az új hipotézisek maximális számának különböző értékei mellett.

Bemondó	Személy 1.		Személy 2.	
# Max. Új Ny. H.	200		200	
Nyelvi Kit. Iter.	1	2	1	2
Pontosság	98.2%	96.2%	97.4%	97.0%
Idő (RTF)	0.29	0.17	0.30	0.19

4. táblázat. Tesztesetek, amikor a ritkább nyelvi kiterjesztés hatékonynak bizonyult

Egy fontos kérdés lehet az is, hogy mekkorának célszerű választani a ***MaxÚjNyelviHipotézis***, illetve a ***HalomMéret*** paramétereket különböző méretű szótárral rendelkező nyelvtanok esetében. A 5. táblázat tartalmazza az ezzel kapcsolatos méréseinket a (HalomMéret=2· MaxÚjNyelviHipotézis beállítás mellett). Látható, hogy a szótár méretének növelésekor célszerű a halom méretét és a kiterjesztések számát növelni, de a nem feltételez egyenes arányosságot, hiszen az 500-as hipotézisszám korlátozás megfelelő felismerési pontosságot ad az 1900 szavas szótár mellett is. Sajnos az is kitűnik, hogy egy 1000 méretű halommal

és 500 eleműre korlátozott nyelvi kiterjesztéssel már kilépünk a valósidejűség tartományából. Ennek az okáról később részletesebben írunk.

Szótár méret	500		1200		1900	
# Max. Ny. H.	po.	idő	po.	idő	po.	idő
250	95.3%	0.39	90.9%	0.48	85.8%	0.52
500	97.5%	1.13	95.9%	1.26	93.5%	1.33
1000	98.3%	3.62	97.4%	4.22	96.0%	4.49

5. táblázat. A nyelvi kiterjesztést korlátozó paraméter vizsgálata különböző méretű szótárak mellett

A következő vágási paraméter, amelyet vizsgáltunk, az ***AkusztikaiKüszöb***, amely egy alsó korlátot ad meg a hipotézisek pontértékére vonatkozóan. A teszteket 500 szavas szótáron, 250 számúra korlátozott nyelvi kiterjesztés mellett végeztük. A 6. táblázatban közölt eredmények szerint egy 200-250 közötti paraméterérték processzorhasználat és pontosság szemponjából is kielégítő.

4.3. A részműveletek processzoridő-igénye

Ebben az alfejezetben azt hasonlítjuk össze, hogy hogyan változik a 1. tábálázatban megadott alapmetódusok processzoridő-igénye, ha változtatjuk az ***AkusztikaiKüszöb*** (AK), illetve a ***MaxÚjNyelviHipotézis*** (MUH) paramétereket (ld. 7. táblázat). Az itt vizsgált műveletek tehát: jemmezővektor-számítás (JSZ), akusztikai hipotézis kiterjesztése (AHK), nyelvi hipotézis kiterjesztése (NYHK), és a hipotézisek egyesítése (HE).

Akusztikai Küszöb	100	150	200	250	300
Pontosság	7.5%	56.4%	93.7%	95.3%	95.2%
Idő (RTF)	0.03	0.13	0.26	0.38	0.46

6. táblázat. Az akusztikai küszöb hatása

A következőket állapíthatjuk meg a táblázat alapján. A jellemzővektor-számítás időigénye viszonylag stabil, és elhanyagolható mértékű. Az akusztikus küszöb növelésével mind a nyelvi kiterjesztés mind az akusztikai kiterjesztés időigénye megnő (ez érthető, hiszen több hipotézis marad, ami több nyelvi kiterjesztési lehetőséget jelent). A nyelvi kiterjesztéskorláttal egyenes arányban szintén nő mind a nyelvi kiterjesztés mind az akusztikai kiterjesztés időigénye. A hipotézisek egyesítésének időigénye viszont a kiterjesztések időigényénel lényegesen jobban nő. Ennek az a magyarázata, hogy ha mindkét egyesítendő hipotézishalmaz elemszáma megduplázódik, akkor az egyesítéskor egy négyszeres műveletigény fog jelentkezni.

AK	MUH	teljes idő (mp.)					relatív idő			
		Össz.	JSZ	AHK	NYHK	HE	JSZ	AHK	NYHK	HE
200	250	1410	74	408	203	724	5%	29%	15%	51%
250	250	2017	73	526	243	1173	4%	26%	12%	58%
300	250	2433	76	620	266	1471	3%	25%	11%	61%
300	500	6022	78	1021	533	4390	1%	17%	9%	73%
300	1000	19286	88	2100	1176	15922	0%	11%	6%	83%

7. táblázat. A részműveletek időigénye különböző vágási paraméterezéseknél

4.4. A nyelvi modul relatív súlyának és a szótár méretének szerepe

A nyelvi modul (a mi esetünkben szó 3-gram) általában mindenféle szósorozat bemondását megengedi (tehát a szótár bármelyik szava után bármelyik következhet). Azonban azok a szósorozatok, amelyek az adott modell szerint nem lehetségesek (például szó N-gram esetén egy szó N-es egyszer sem fordult elő a tanító szövegkorpuszban) egy megfelelően kicsi nyelvi pontértéket kapnak (jelölje ezt ϵ). Az, hogy ez az érték mennyire kicsi, meghatározza, hogy mekkora lesz a nyelvi modul súlya az akusztikai modulhoz viszonyítva, a nyelvi és akusztikai pontozás kombinálásakor. Egy beszédjel-keret átlagos akusztikai súlya (helyesen felismert bemondásokon mérve) kb. $10^{-20} \dots 10^{-22}$, ehhez viszonyíthatjuk a nyelvi modellünk pontozását. Nyilvánvaló, hogy minél kisebb az ϵ pontérték, annál inkább igazodni fog a felismert szósorozat a nyelvi modellhez (egyszerűen azért, mert amelyik hipotézis nem felel meg a nyelvi modellnek, az egyre inkább lejjebb kerül a rangsorban). A 8. táblázatban foglaltuk össze az eredményeinket. Azt vehetjük észre, hogy minél kisebb az ϵ értéke, annál gyorsabban történik a beszédfelismerés. Ennek az oka az, hogy a nyelvi modell pontozása elnyomja az akusztikai pontozást, és a küszöb szerinti vágás után a hipotézisekből egyre kevesebb marad. Mindemellett a felismerési pontosság egy érték alatt (kb. 1e-40) csökkenni kezd, ami szintén az előbbi magyarázatnak tudható be.

Szótár méret	500		1200		1900	
nyelvi súly	po.	idő	po.	idő	po.	idő
1e-10	87.3%	0.64	74.4%	0.72	65.8%	0.77
1e-20	94.2%	0.53	86.9%	0.62	82.0%	0.68
1e-30	95.5%	0.45	90.7%	0.53	85.7%	0.58
1e-40	95.3%	0.39	90.9%	0.48	85.8%	0.52
1e-50	94.8%	0.36	90.5%	0.42	85.1%	0.47
1e-60	93.9%	0.33	89.1%	0.39	83.6%	0.44

8. táblázat. A nyelvi modul súlyának szerepe különböző méretű szótárak esetén

5. Konklúzió

Ebben a cikkben először leírtuk az általunk fejlesztett folyamatos diktálórendszer leglényegesebb technikai megoldásait. Ezután megvizsgáltuk, hogy a legfon-

tosabbnak vélt, keresési teret vágó paramétereknek mi a szerepük, és hogyan befolyásolják a processzoridő-igényt, valamint a felismerési pontosságot. Arra a következtetésre jutottunk, hogy a nyelvi modell súlyának, illetve az akusztikai küszöb értékének egy elég stabil optimális értéke van, függetlenül a szótár méretétől. Ezzel szemben nagyobb szótár használatakor a nyelvi kiterjesztést korlátozó, illetve a halomméretet megadó paraméter értékének a növelésével tudunk csak megfelelő felismerési pontosságot elérni. Szerencsére az ezzel kapcso-latos kisérletek azt mutatták, hogy nem szükséges ezek lineáris növelése. Az eredmények arra is rávilágítottak, hogy ha nagyszótáras rendszer felé akarunk később továbblépni, akkor a jelenlegi, hipotézisek egyesítését végző algoritmusunkat le kell cserélnünk egy hatékonyabbra, mert időigény szempontjából ez a kritikus része a diktálórendszerünknek.

Hivatkozások

1. Chou, W., Juang, B.H., eds.: Pattern Recognition in Speech and Language Processing. CRC Press, Inc., Boca Raton, FL, USA (2002)
2. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1993)
3. Ravishankar, M.: Efficient algorithms for speech recognition (1996)
4. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing. Prentice Hall (2001)
5. Bánhalmi, A., Kocsor, A., Paczolay, D.: Magyar nyelvű diktáló rendszer támogatása újszerű nyelvi modellek segítségével. In: MSZNY. (2006) 337–347
6. Forney, G.D.: The viterbi algorithm. Proceedings of The IEEE **61** (1973) 268–278
7. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. (1990) 267–296
8. Kozub, D.: http://www.dankozub.com/cpp/hlist.htm (1999)
9. Vicsi, K., Kocsor, A., Teleki, C., Tóth, L.: Beszédadatbázis irodai számítógép-felhasználói környezetben. In: MSZNY. (2004) 315–318

Prozódiai információ használata az automatikus beszédfelismerésben; mondat modalitás felismerése

Vicsi Klára, Szaszák György és Németh Zsolt

Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium, 1111 Budapest, Sztoczek u. 2.
{vicsi, szaszak}@tmit.bme.hu

Kivonat: A mai, statisztikai elvi alapokra épülő folyamatos gépi beszédfelismerők kimenetén szóláncok sorozata jelenik meg, tehát a beszédfelismerés több szintű feldolgozási folyamatából a szószintig jutott el a mai beszédfelismerési technológia. Robusztus beszédfelismerés eléréséhez azonban további – például szemantikai – szintek bevonása szükséges.

A beszéd szupraszegmentális (prozódiai) paramétereinek bevonásával egy olyan prozódiai felismerőt hoztunk létre, amely a mondatok és tagmondatok fajtáit, azaz modalitását, illetve a mondatok határait ismeri föl, és ezzel hozzájárulhat a szemantikai szintű nyelvi felismerés biztosabb döntéseihez. Ez az ún. modalitás felismerő statisztikai elven működik, a mondatok, tagmondatok intonációs struktúráját leíró Rejtett Markov modellekből, és egy igen egyszerű, a mondatok kapcsolódására vonatkozó modellből épül fel.

A felismerő tesztelési eredményei azt mutatták, hogy azoknál a modalitás típusoknál, amelyekre a statisztikai betanításhoz elegendő minta állt rendelkezésre, a helyesen felismert modalitás aránya 75 és 95% között változott az adott mondat modalitásától függően.

1 Bevezetés

A beszédfelismerési folyamatnak számos szintje létezik: akusztikai, fonetikai-fonológiai, szintaktikai, szemantikai, pragmatikai szint (Ainsworth 1976). Ezek közül a szintek közül minél többet tudunk a gépi beszédfelismerési folyamatba bevonni, annál biztosabb lesz a felismerés.

A gépi beszédfelismerésnél akusztikai szinten működik az akusztikai előfeldolgozó egység, amely a beszédjel elemzését, a lényegkiemelést, tömörítést végzi el. Kimenetén jelennek meg az időkeretenkénti lényegi paraméterek (jellemző vektorok), amelyek szegmentális akusztikai szintű előfeldolgozásnál jellemzően 10 ms időkeretekben, 25-50 ms időablakban mért színképi paraméterek, leggyakrabban MFC (Mel-frekvencia kepsztrális) együtthatók (ld. 1. ábra 'a' feldolgozási ága) (Young 2005). A hagyományos beszédfelimerési folyamatban a következő szinten, a fonetikai-fonológiai szinten történik a beszéd szegmentális feldolgozása, vagyis az akusztikai szinten kapott lényegi paraméterek segítségével végezzük el a beszéd-

hangok modelljeinek a megalkotását. Felismeréskor az akusztikai szinten kapott jellemző vektorsorozatot hasonlítjuk össze a beszédhangok modelljeivel. Ez az a feldolgozási szint, ahol a kimeneten fonémasorozatot kapunk. Amennyiben szintaktikai szintű nyelvtant is bekapcsolunk a felismerésbe – például a legelterjedtebben használt statisztikai alapú N-gram[1] nyelvi modelleket –, akkor a kimeneten szósorozatot kapunk (Becchetti–Ricotti 1999), amint az az 1. ábra 'a' ágában látható. A kereskedelemben manapság elérhető beszédfelismerők így működnek.

2 Célkitűzés

A Beszédakusztikai Laboratóriumban olyan vizsgálatokat végeztünk, amelyekben arra kerestük a választ, hogy az akusztikai előfeldolgozással hogyan lehet hozzájárulni a magasabb feldolgozási szintek, a szintaktikai, valamint szemantikai szintű nyelvi feldolgozás eredményesebbé tételéhez. Az akusztikai előfeldolgozásra ekkor már nem célszerű a fent említett szegmentális tartományban végzett lényegkiemelés jellemző vektorait használni. Más, szupraszegmentális (prozódiai) jellemzőkön alapuló lényegkiemelésre van szükség, amely tükrözi bizonyos beszédtartalmak megkülönböztetését, az értelmi tagolást és akár az érzelmeket is. Ennek megfelelően a beszéd fizikai paramétereit a szupraszegmentális tartományban, jellemzően durvább frekvencia- és időfelbontásban célszerű vizsgálnunk, mint amikor a szegmentális jellemzés volt a cél. A szupraszegmentális jellemzők vizsgálatánál figyelembe kell vennünk néhány tényt, amelyek nehezítik e tartományban a jellemző vektorok kinyerését. Ezek közül néhányat alább közlünk.

– A szupraszegmentális paraméterek jelentős mértékben variálódnak a beszédstílus, a beszélő, a tartalom, a környezet, stb. függvényében.

– A nyelv rétegződését a nyelv sok esetben a szupraszegmentális információval is érzékelteti. Az üzenet nyelvi rétegződési szintjei és a szupraszegmentális információ között azonban nagymértékű az egymásra hatás, jellemző a szintek közötti kapcsolatok bonyolultsága (Langlais, 1993). Egy adott szinten jellemző szupraszegmentális jellemzőket nehéz kinyerni, mivel a felette lévő szint erősen befolyásolja az alsóbb szintek alakulását. Közismert például, hogy a szóhangsúlyokat a mondathangsúly erősen befolyásolja.

– A szupraszegmentális jellemzésre használt fizikai paraméterek közül az alapfrekvencia pontos mérésének nehézségei, illetve rendszerbe illesztésük közismert.

Amikor a szupraszegmentális paramétereket (illetve a belőlük előállított jellemző vektorokat) használjuk a beszédfelismerés segítésére, ezt több szinten tehetjük meg, amint ez az 1. ábrán a 'b' és a 'c' ágban látható. A magasabb nyelvi szintek felé haladva, lépésről lépésre járulunk hozzá a beszédfelismerés biztonságosabbá tételéhez. Szintaktikai szinten, a hagyományos szólánc kimenetű felismerők teljesítményéhez

[1] A beszédfelismerésben nyelvi modell alatt egy adott nyelv lehetséges szókapcsolatainak leírását értjük. Statisztikai nyelvi modell esetén az „N-gram" elnevezés arra utal, hogy a nyelvi modell adott, N darab szóból felépülő szósorozatok előfordulási valószínűségeit tárolja paraméterként.

képest lényegesen javíthatjuk a felismerés biztonságát kötött hangsúlyozású nyelveknél (pl. magyar vagy finn nyelvekre) a szóhatárok automatikus bejelölésével. Ez a

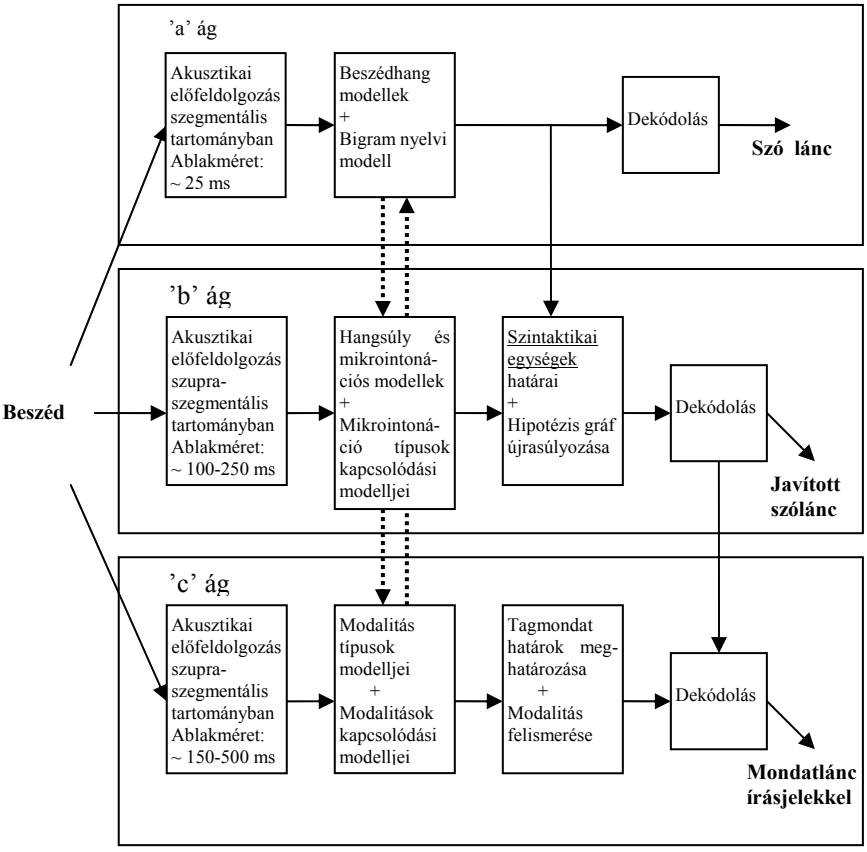


Fig. 1: A kibővített több szintű beszédfelismerő tömbvázlata (az ábra részletes magyarázatát lásd a fenti szövegben

szintaktikai szintű feldolgozás az 1. ábra 'b' feldolgozási ágában követhető végig. A módszer lényege, hogy olyan kötött hangsúlyozású nyelveknél ahol a szóhangsúly az első szótagon van, szupraszegmentális paraméterek segítségével mikrointonációs egységeket tudunk automatikusan a beszédfolyamba bejelölni, amelyek mutatják a szóhatárt. A magyar nyelvre beszédpercepciós vizsgálatok is megerősítik e módszer helyességét, hiszen az emberi beszédfeldolgozó rendszerünk is használja ezeket a szupraszegmentális információkat a szóhatárok megtalálásához (Tóth 2007). A szóhatárok automatikus bejelölésének módszeréről, a felismerésbe történt beépítés hatékonyságáról már korábban beszámoltunk (Vicsi–Szaszák 2004, 2005).
Az 1. ábra 'c' feldolgozási ágában szemantikai szintű feldolgozás látható. Jelen cikkünk tárgyát éppen ez a szupraszegmentális paraméterekre épülő szemantikai szintű

feldolgozás képezi, amely során az esetleges tagmondatok határainak lokalizációját és a mondatok modalitásának felismerését hajtjuk végre. A modalitás felismerésnél az dönthető el, hogy például a feltételezetten kimondott „*alma van a fa alatt*” szósor állítás, kérdés vagy felkiáltás formájában hangzott el. A mondat és tagmondat határok lokalizációja révén támpontot is kapunk a tekintetben, hogy melyek a szósorozatban a (tag)mondatok kezdő- és végidőpontjai.

3 Szematnikai szintű modalitás felimserő elvi alapjai, a fejlesztés módszere

A mondat-, ill. tagmondatfajták felismerését statisztikai elven, a Rejtett Markov Modell (HMM) módszer alapján végeztük el, amelyre a HTK fejlesztői rendszert (Young 2005) használtuk. A felismerő betanításához a modalitásfajták szerint feldolgozott beszédadatbázist használtuk. Megjegyezzük, hogy a tagmondatokat is el kellett különítenünk modalitás szempontjából aszerint, hogy az a mondat, amelybe beágyazódnak, milyen modalitású. A továbbiakban tehát – némi pongyolasággal – tagmondatok modalitásáról (ill. fajtáiról) is írunk majd, ez alatt természetesen mindig az értendő, hogy az adott tagmondat milyen modalitású mondat része. Az egyes (tag)mondattípusokra HMM modelleket építettünk fel, amelyek segítségével a tagmondatfajtákat felismertük. A felismeréshez felhasználtuk a tagmondatok egymás utáni sorrendjét figyelembe vevő szöveg szintű prozódiai modellt is, mely tulajdonképpen egy egyszerű nyelvi szabálymodell. Alapvető feladatnak tekintettük a különböző HMM tagmondatfajták modelljeinek optimális beállítását.

3.1 A betanító anyag elkészítése

A BABEL (Vicsi et al. 1998) és az MRBA (Vicsi et al. 2004) beszédadatbázisainkból kigyűjtöttük a különböző mondatfajtákhoz tartozó mondatokat. Összesen 10 alapvető mondat, ill. tagmondatfajtát (típust, modalitást) különböztettünk meg (vö. Olaszy 2002) az 1. Táblázat szerint.A kiválasztott hangfájlokat lehallgatás és az alapfrekvencia, illetve energiaszint mérése alapján tagmondatok szerint szegmentáltuk és címkéztük a (tag)mondatok „modalitása” szerint, vagyis a hanganyagba bejelöltük a mondatok, tagmondatok határait, valamint a tagmondatok típusainak (modalitásainak) megfelelő szimbólumokat. Szegmentálásra, címkézésre a 2. ábrán mutatunk példát. A 2. ábra első sorában a hanganyag hullámformája, a második sorban az alapfrekvencia és az intenzitás görbéi láthatók. A harmadik sorban van a kézzel elvégzett szegmentálás és címkézés. A mondat és tagmondat típusokon, illetve ezek határain kívül a tagmondatok és mondatok közötti szünetrészt külön bejelöltük (’U’ szimbólummal). A szünetrész bejelölésére a tagmondatok között a kb. 400 ms-nál, míg a mondathatároknál a kb. 500 ms-nál nagyobb szüneteknél került sor – lásd például a 2. ábrán a két kijelentő (’S’-sel jelölt mondat) közti kiemelt részt. Ezen értékeknél kisebb szüneteknél általában csak határt jelöltünk, és a határ feléhez tettük be az „elválasztást” mint a 2. ábrán az ’E’-vel jelölt eldöntendő kérdés, és a ’T’-vel jelölt tagmondat között.

1. Táblazat: A szegmentálás és címkézés statisztikája

Egyszerű mondat modalitás és összetett mondat modalitása tagmondatonként[2]	**Jelölés (cimke)**	**Összes előfordulás (db)**
Kijelentő mondat, Kijelentést záró tagmondat	S	445
Kijelentő tagmondat, a záró tagmondat nélkül	T	287
Kiegészítendő kérdés	K	40
Kiegészítendő kérdés tagmondata	KT	13
Eldöntendő kérdés	E	35
Felszólító és felkiáltó mondatok	FF	52
Felszólító és felkiáltó mondatok tagmondata	FFT	24
Óhajtó mondat	O	2
Felsorolás	F	41
Semleges	N	125
Összesen:		**1029**

Az ilyen módon feldolgozott adatbázissal végeztük el a prozódiai felismerőnk betanítását. Mivel az óhajtó mondatra csak 2 mintánk volt az adatbázisban, ezt a tanításnál kihagytuk. Így 9 tagmondat és egy szünet HMM modellt hoztunk létre. A szegmentálás és címkézés statisztikáját az 1. táblázat mutatja.

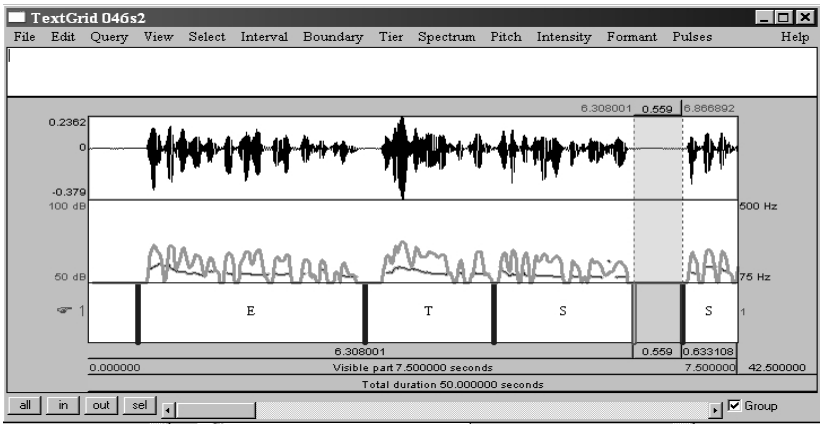


Fig. 2: Szegmentálás és címkézés a Praat programban.

[2] Tagmondat modalitáson jelen cikkünkben, a beszédtechnológiai kezelhetőség érdekében, most azt értjük, hogy egy adott tagmondat milyen modalitású összetett mondat része.

Egyszerű és összetett mondatokat vegyesen használtunk fel, és a táblázat statisztikájából látható, hogy több mint 1000 címkét helyeztünk el szegmentálás közben.

Az eredeti energia- (e_i) és alapfrekvencia-értékeket (f_{oi}) 25 ms időablakban, 10 ms-os időkeretenként mértük. Az alapfrekvencia értékét gördülő átlagos magnitudókülönbség-függvény (Short-time Average Magnitude Difference Function – AMDF) segítségével határoztuk meg (Gordos et al 1983).

Az előfeldolgozás utolsó lépéseként a mért energia és alapfrekvencia értékeket különböző időablakban átlagoljuk. Ezek az 5, 10, 20, 26, 30, 36, 40 és 50 keretszámok, szorozva a 10 ms keretidővel. A bizonyos intervallumban átlagolt érték lesz a középső elem értéke. A különbözőképpen betanított modellek közül teszteléskor választjuk ki az optimálisat.

Az alapfrekvencia értékeinek a feldolgozásakor oktávszűrést hajtunk végre, mivel a felhangszerkezetben téveszthet az alapfrekvenciát detektáló algoritmus: oktávot ugorhat.

Az e_i és f_{oi} értékek mellett három intervallum nagyság alapján 3-3 első és második deriváltat számítottunk ki mind az alapfrekvenciához, mind az intenzitáshoz. Az intervallum nagyságát a deriváltak számítására szolgáló regressziós képlet szerint vettük figyelembe:

$$d_t = \frac{\sum_{i=1}^{T} i\left(c_{t+i} - c_{t-i}\right)}{2\sum_{i=1}^{T} i^2}. \qquad (1)$$

A (1) képlet a t időpillanathoz számít derivált értéket, a c_t a t-hez tartózó együtthatót jelenti. A T az intervallum nagyságát jelentő változó, amelynek értéke 10, 20 és 40 lesz.

Így keletkezik az összesen 14 elemű jellemzővektor:
$V_{jell}=\{f_{oi}, e_i, df_{oi}^{10}, d^2f_{oi}^{10}, df_{oi}^{20}, d^2f_{oi}^{20}, df_{oi}^{40}, d^2f_{oi}^{40}, de_i^{10}, d^2e_i^{10}, de_i^{20}, d^2e_i^{20}, d^2e_i^{40}, d^2e_i^{40}\}$.
A d, d^2 az első és a második deriváltat, míg a deriváltak utáni indexben lévő szám az intervallum nagyságát jelenti.

3.2 Betanítás a különböző paraméterekkel

A feldolgozott beszédadatbázist két részre bontottuk: az egyik résszel a betanítást, míg a másikkal a tesztelést végeztük. A mondatok nagyrészt véletlenszerűen lettek kiválasztva, de arra odafigyeltünk, hogy minden felismerendő címke szerepeljen mind a betanított, mind a tesztelésre szánt anyagban.

A betanítás során az adatbázis hangfájljaiból az előfeldolgozással nyert szupraszegmentális jellemző vektorokat, valamint az adatbázis szegmentálási és címkézési adatait használjuk fel a prozódiai modellek felépítéséhez.

3.3 Tesztelés

A szemantikai szintű prozódiai felismerő tesztelése két fő folyamatból áll: a felismerésből és az összehasonlítás utáni értékelésből. A folyamatot a 3. ábra szemlélteti. A szupraszegmentális előfeldolgozás után, a betanítás során kialakított tagmondattípus modelleket használtuk a felismeréshez, valamint a korábban már említett, mondatok kapcsolódását leíró szabályokat. Ebben olyan nyelvtani szabályokat írtunk elő a modalitás felismeréséhez, hogy az a hétköznapi, folyamatos beszédben gyakran, kevés kivétellel előforduló eseteket maradéktalanul lefedje. A szabályok azt adják meg, hogy milyen (tag)mondat milyen (tag)mondatot követhet, és milyen (tag)mondatot nem, illetve milyen (tag)mondatok ismétlődhetnek, stb. Lényegében analóg szerepe van a beszédfelismerésnél használt nyelvi modellével, de statisztikai adatok hiányában, illetve a jóval kevesebb lehetőség miatt szabályokat adtunk meg.

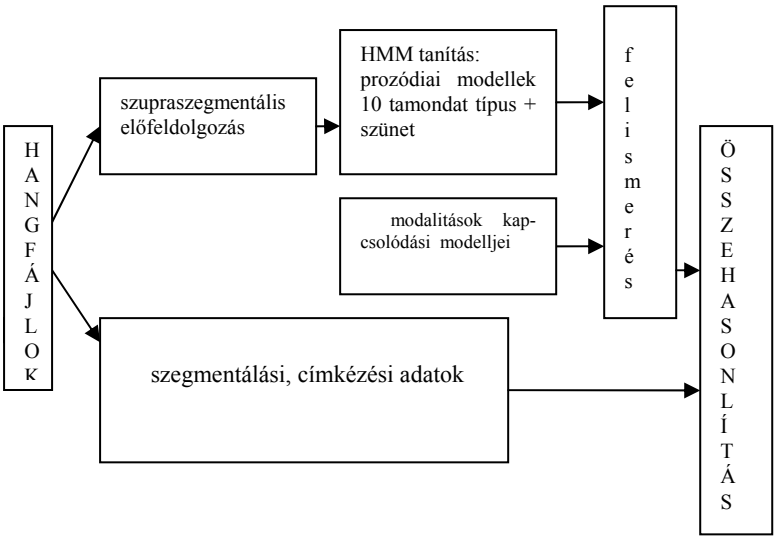


Fig. 3: A tesztelés folyamatábrája

A felismerés jóságát a helyes felismerés arányával *(Corr)* és a pontossággal *(Acc*) adjuk meg. A modalitás szerint helyesen felismert tagmondatok aránya:

$$Corr = \frac{H}{N} \cdot 100\% \tag{2}$$

A pontosság számítása:

$$Acc = \frac{H - I}{N} \cdot 100\% \tag{3}$$

ahol *H* a modalitás szerint helyesen felismert tagmondatok, *I* a beszúrások[3] és *N* az összes tagmondat száma.

A teszteléseket először az 1. táblázat szerinti 10 különböző címke betanításával és felismerésével kezdtük. Az eredmények feldolgozása során hamar kiderült, hogy egyes modalitás típusokból nincs elegendő számú minta a betanításhoz, valamint az energia- és alapfrekvencia-menet „hasonlósága" miatt egyébként is célszerű csoportosítást végezni az alábbiak szerint:

- A felsorolások (’F’), illetve a felkiáltó és felszólító mondatok tagmondataiból (’FFT’) is arányaiban kevés minta van, továbbá intenzitás- és alapfrekvencia-szerkezetükben is hasonlítanak. Ezeket gyakran felsorolásnak (’F’), vagy tagmondatnak (’T’) detektálja a felismerő. Továbbá mindegyikhez a vessző írásjel tartozik, ezért mindkét csoport összevonható a kijelentő mondat tagmondatával: ’F’, ’FFT’ → ’T’.
- A kiegészítendő kérdést tartalmazó mondatok tagmondatainak (’KT’) szerkezete nagy hasonlóságot mutat az egyetlen tagmondatból álló kiegészítendő kérdésével (’K’). A mondatok értelmezése szempontjából továbbá nem jelent különbséget, ha a *„Hová menne, és mit csinálna akkor?"* mondatot két kérdésként: *„Hová menne? És mit csinálna akkor?"* ismeri fel a modell. Ezért ezek összevonhatóak: ’KT’ → ’K’.

Így végül a csoportosítással (összevonással) 6 tagmondatmodellt, és egy szünet-modellt tanítottunk be, és használtunk a felismeréshez.

A további teszteléskor a szupraszegmentális jellemző vektorok átlagolási intervalluma (lásd a „Betanító anyag előkészítése" c. pontban), valamint a HMM (tag)mondatmodellek állapotainak száma függvényében vizsgáltuk a felismerés jóságát. Arra kerestük a választ, hogy az energia és alapfrekvencia jellegű jellemzők milyen időfelbontása szükséges ahhoz, hogy a különböző tagmondattípusokat optimálisan tudjuk felismerni.

Kérdés továbbá a HMM tagmondatmodellek állapotainak optimális száma. Nyilvánvalóan több állapotra van szükség, mint a fonémamodelleknél használt 3 állapot, de hogy ezen modelleknél hány állapotot kell felvennünk az optimális felismeréshez, azt a teszteléssel döntöttük el.

A két tényezőt együttesen vizsgáltuk, rögzített ($\log P_{ins}=0$) szóbeszúrási valószínűséggel[4] végeztük el a modalitás felismerést (ld. 4. ábra). Az eredményeket a 2. táblázat szemlélteti. A táblázat oszlopaiban találhatóak az átlagolási intervallum keretszámai. A sorokban a tagmondattípus modellekben beállított állapotok száma a változó értéke. A táblázat celláiban – százalékban kifejezve – találhatóak a helyes felismerés (*Corr*) eredményei.

[3] Beszúrásnak nevezik a beszédfelismerésben a valós tesztanyagban nem megjelenő, azonban a felismerő által feltételezett és így tévesen felismert elemet, mely esetünkben annak felel meg, hogy a modalitás felismerő egy további tagmondatokra nem bontható (tag)mondatot tévesen felbontott.

[4] A szóbeszúrási valószínűség a beszédfelismerők egy állítható paramétere, melynek kisebbre állításával – durva megfogalmazással – a felismerő mérsékli az adott hangmintára illesztett szimbólumok számát. Esetünkben a szóbeszúrási valószínűség a „(tag)mondat beszúrás" valószínűségének felel meg.

2. Táblázat A 7 különböző címke helyes felismerésére (*Corr*) %-osan

HMM Álla-potok száma		Időablak 10 ms-os keretenként							
		5	**10**	**20**	**26**	**30**	**36**	**40**	**50**
	5	-	-	60,24	59,76	60,00	58,31	58,31	58,80
	11	66,07	67,47	67,95	68,92	67,47	67,23	69,40	65,06
	15	-	66,99	66,51	66,27	67,47	66,99	64.58	66,47
	19	-	-	66,99	64,10	65,06	63,37	63,37	60,02

A legjobb eredményt 11 HMM állapot mellett kaptuk. Az időablak nem változtatja tendenciózusan az eredményeket 100 és 400 ms között. A legjobb átlagos modalitásfelismerés 69,4 % volt.

A tagmondattípusokra lebontott tévesztési mátrix a 11-es állapotszám és a 40 keretnyi átlagos intervallum mellett a 3. táblázatban látható.

A mátrix sorai jelentik azt, hogy mi volt az eredeti modalitás, az oszlopok jelentése pedig, hogy mit ismert fel a felismerő. Az utolsó, '*Ins'* feliratú sorban lévő tagmondat típusokat hamisan beszúrta a felismerő, és a '*Del'* feliratú oszlopban lévőket pedig törölte.

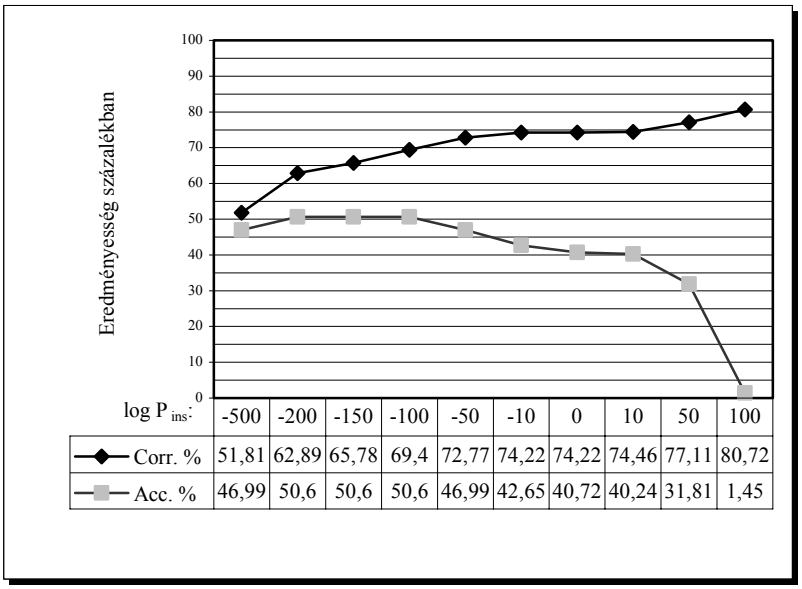
A tévesztési mátrix jobb oldalán látható, hogy bizonyos mondatfajtákra egészen elfogadható eredmények születtek: '**FF**' – 50%, '**T**' – 83,3%, '**S**' – 74,8%, és '**U**' – 96,0%. Ezek közül az első három érdemleges a modalitás-típusok felismerése, az utolsó pedig a mondathatárok detektálása szempontjából.

3. Táblázat. A tagmondattípusokra lebontott tévesztési mátrix a 11-es állapotszám és a 40 keretnyi átlagolási intervallum mellett, Corr=69.40 %, Acc=50.60 %

	S	**T**	**K**	**E**	**F**	**N**	**U**	**Del**	**[corr[%]]**
S	**83**	11	7	4	2	3	1	7	[74.8]
T	4	**70**	0	1	2	0	7	13	[83.3]
K	3	3	**4**	0	0	0	2	3	[33.3]
E	1	1	0	**2**	1	0	1	0	[33.3]
FF	0	2	1	0	**5**	0	2	4	[50.0]
N	3	4	0	0	2	**4**	2	14	[26.7]
U	0	5	0	0	0	0	**120**	11	[96.0]
Ins	6	32	3	2	4	4	27		

A (tag)mondat beszúrás valószínűségének optimalizálása is fontos szerepet játszik a felismerés hatékonyságában. Érdemes azonban odafigyelni arra is, hogy a helyesen felismert tagmondatok mellett a pontosság is fontos. A beszúrás valószínűségének növelésével a helyesen felismert modalitások mellett sok plusz címkét is elhelyezünk,

így például a mondathatárok meghatározása nagyon nehézkessé válik. A két eredményességi mutató változásait ezért együttesen figyeltük a tagmondatbeszúrás logaritmusának (a 4. ábrán $\log P_{ins}$) függvényében, amint azt a 4. ábra mutatja.



log P_{ins}:	-500	-200	-150	-100	-50	-10	0	10	50	100
Corr. %	51,81	62,89	65,78	69,4	72,77	74,22	74,22	74,46	77,11	80,72
Acc. %	46,99	50,6	50,6	50,6	46,99	42,65	40,72	40,24	31,81	1,45

Fig.4: Az eredményesség alakulása a (tag)mondatbeszúrás valószínűségéneklogarimusa ($\log P_{ins}$) függvényében

A helyesen felismert tagmondat modalitás típusok aránya a 100-as értéknél adódik a legmagasabbra. Az is megfigyelhető, hogy ekkor a pontosság százaléka nagyon alacsony (mindössze 1,45%), ami azt jelenti, hogy a felismerés tele van „felesleges" beszúrásokkal, és a mondathatárokat nem lehet helyesen felismerni. A pontosság három vizsgált értéknél haladja meg az 50%-ot (mindháromnál 50,6%), és ezek közül a helyesen felismert szavak aránya a -100-as értéknél a legmagasabb: 69,4%.

4 Értékelés

A fentiekben bemutatott szemantikai szintű modalitás felismerő nem túl nagy, és mondattípus eloszlásban is egyenetlen adatbázissal lett betanítva és tesztelve. Ennek ellenére a paraméterek optimális beállítása mellett a vártnál jobb eredmények adódtak. A legjobb felismerési eredményt akkor kaptuk, amikor az energia és az alapfrekvencia időfelbontása 100-400 ms közötti volt (átlagos olvasott beszédtempó mellett), a HMM tagmondattípus modellek állapotainak száma 11, és a mondatelem beszúrás valószínűségének a logaritmusa: -100 volt.

Ezen beállításokkal közel 70% a helyesen felismert címkék aránya, és a pontosság is több mint 50%-os értéket mutat, annak ellenére, hogy egy-egy tagmondatból több

száz darab csak az **S** és **T** mondatfajtáknál fordult elő. Az ’**S’** és ’**T’** típusok (kijelentő mondat és tagmondata) kb. 75%-os, illetve 83%-os eredménnyel detektálhatóak, továbbá az ’**FF’** mondatok helyes felismerése is eléri az 50%-ot annak ellenére, hogy betanításra, tesztelésre összesen csak 52 mondatunk volt. Továbbá a mondathatárok 96% pontosságú bejelölése szintén jó eredmény (ld. a.3. táblázat ’U’ feliratú sorában). A mondathatárok automatikus (’rec’) valamint kézi (’lab’) bejelölése összehasonlítására mutatunk példát az 5. ábrán.

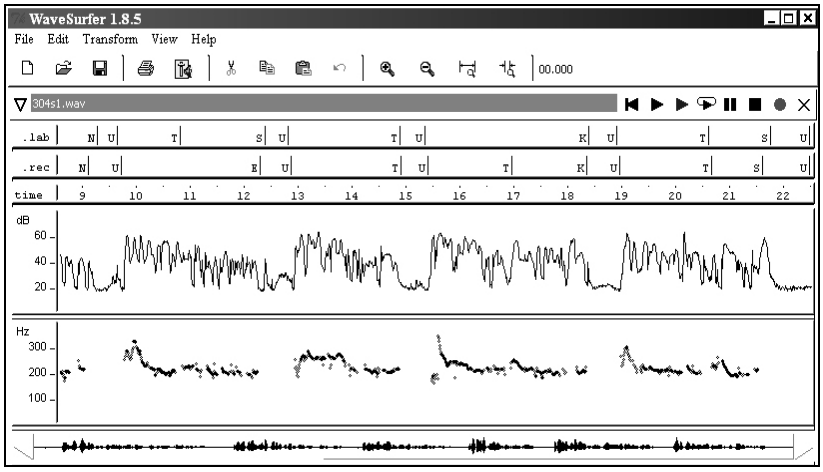


Fig.5. Tagmondathatárok kézi (’lab’) és automatikus (’rec’) bejelölése

Az eredményekből látható, hogy a nagyméretű adathalmazzal betanított tagmondatfajták (kijelentő mondatok és a kijelentő mondatok tagmondata) jó eredménnyel felismerhetőek, ezért célszerű a további mondatfajtákhoz is hasonló mennyiségű betanító és tesztelő anyag feldolgozása a jövőben.

A tagmondat alapú nyelvi modell nagyméretű adathalmazzal történő, statisztikai alapú kialakítása szintén jelentősen javíthatná a felismerés biztonságát.

Specifikusabbá lehet tenni a felismerőt, ha előre szerkesztett párbeszédekből felépült adatbázis betanításával építhetnénk fel a tagmondat modelleket, mert akkor az érzelmek (például a felkiáltó és felszólító mondatok prozódiai tulajdonságai) jobban kimutathatóak lennének, mint a most használt olvasott szöveg adatbázisokban végzett válogatás alapján.

A mondat és tagmondat típusok, valamint a mondathatárok felismerésének javítására a munkát tovább kell folytatni. Jelen cikkünkkel az volt a célunk, hogy bemutassuk, hogy érdemes ezen jellemzők vizsgálata, és bevonáshasznos az automatikus gépi beszédfelismerésbe.

Bibliográfia

1. Ainsworth, W.: Mechanisms of speech recognition. Pergamon Press. Oxford. (1976) 110-124.
2. Becchetti, C.-Ricotti, L. P: Speech Recognition,Theory and C++ implementation. Fondazione Ugo Bordoni, John Wiley. Rome .(1999)
3. Gordos, G.-Takács, Gy.: Digitális beszédfeldolgozás. Műszaki Könyvkiadó. Budapest. (1983) 239-240.
4. Gósy, M.: Fonetika, a beszéd tudománya. Osiris. Budapest. (2004) 182-243.
5. Langlais, P.-Méloni, H: Integration of a prosodic component in an automatic speech recognition system. 3rd European Conference on Speech Communication and Technology. Berlin. .(1993) 2007-2010.
6. Tóth, L.: Benchmarking Human Performance on the Acoustic and Linguistic Subtasks of ASR Systems. INTERSPEECH2007, Antverp. (2007) 382-385.
7. Olaszy, G.: A magyar kérdés dallamformáinak és intenzitásszerkezetének fonetikai vizsgálata. Beszédkutatás. Gósy Mária. MTA Nyelvtudományi Intézet. Budapest, (2002) 83-99.
8. Young, S., et al.: The HTK Book (for HTK Version 3.3). Cambridge University. Engineering Department, (2005)
9. Vicsi, K., Víg, A.: Az elsõ magyar nyelvû beszédadatbázis. Beszédkutatás. Gósy Mária. MTA Nyelvtudományi Intézete. Budapest. (1998) 163-177.
10. Vicsi Klára-Kocsor András-Teleki Csaba-Tóth László: *Beszéd adatbázis irodai számítógép-felhasználói környezetben*. II. Magyar Számítógépes Nyelvészet Konferencia. Szeged. Alexin Zoltán-Csendes Dóra. Szegedi Tudományegyetem. Informatikai Tanszékcsoport. (2004) 315-318.
11. Vicsi Klára-Szaszák György-Borostyán Gábor: *Folyamatos beszéd szó- és frázisszintű automatikus szegmentálása szupraszegmentális jegyek alapján*. II. Magyar Számítógépes Nyelvészet Konferencia. Szeged. Alexin Zoltán-Csendes Dóra. Szegedi Tudományegyetem Informatikai Tanszékcsoport. (2004) 319-326.
12. Vicsi Klára-Szaszák György: *Folyamatos beszéd szó- és frázisszintû automatikus szegmentálása szupraszegmentális jegyek alapján.* II. rész *Statisztikai eljárás, finn - magyar nyelvû összehasonlító vizsgálat.* Magyar Számítógépes Nyelvészeti Konferencia. Szeged. Alexin Zoltán-Csendes Dóra. Szegedi Tudományegyetem. Informatikai Tanszékcsoport. (2005) 360-370.

A beszéd érzelemi töltetének számítógépes felismerése

Tüske Zoltán[1,3], Simon Márta[2], Mihajlik Péter[1], Gordos Géza[1]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék
{tuske, mihajlik, gordos}@tmit.bme.hu

[2]Semmelweis Egyetem
Pszichiátriai és Pszichoterápiás Klinika
{simonmarta}@t-online.hu

[3]AITIA International

Kivonat: Új megközelítést mutatunk be a beszéd érzelmi tartalmának gépi felismerésére. Megmutatjuk, hogy statisztikai módszerekkel, csak a beszéd akusztikus jellemzői alapján, a szöveges tartalom figyelembe vétele nélkül megfelelő érzelemfelismerési eredményeket lehet elérni. Lineáris diszkrimi-náns alapján válogatott beszédjellemzők mennyiségét – azaz a jellemzővektor dimenzióját – adatvezérelt módszerekkel (PCA és LDA) radikálisan csökkent-jük, majd GMM osztályozókat tanítunk be. Sokbeszélős, hat érzelmi állapotra jellemző, magyar adatbázison átlagosan 42,9%-os felismerési pontosságot ér-tünk el. Felismerőnk 60,2%-kal ismerte fel az érzelmeket beszélőfüggő eset-ben. A megközelítés nyelvek közötti hordozhatóságát mutatja, hogy német adatbázison színészek által produkált felvételeken, kötött szöveges tartalom mellett, hét érzelemi osztállyal 71,8%-os beszélőfüggetlen felismerési ered-ményt értünk el, ami nemzetközi élvonalbelinek mondható.

1 Bevezetés

A beszédfeldolgozás területén az érzelemfelismerés mindinkább a figyelem középpontjába kerül. Az automatikus beszédfelismerővel ellátott rendszerekkel kapcsolatban felmerül az az igény, hogy a beszéd szöveges tartalmán kívül egyéb, non-verbális információt is – például a beszélő érzelmi állapotát – képes legyen figyelembe venni és felhasználni, ezáltal téve természetesebbé a felhasználó és a gép közötti kommunikációt.

Az érzelemfelismerési kutatások különböző forrásokból származó jeleken vizsgálódnak, úgymint fiziológiai, mimikai és beszédjelek. Ez a tanulmány a továbbiakban csak a beszédből géppel kinyerhető érzelmi információkkal foglalkozik.

Az ember képes még a telefonon keresztül érkező sávkorlátozott (400-3700 Hz) akusztikus jelből is a vonal túloldalán levő személy érzelmi állapotának meghatározására. Természetesen a vizuális információ, a gesztikuláció, az arcizmok igen kifinomult játékának hiánya gyakran vezet téves emocionális értékeléshez.

Habár a vokális csatorna közvetítette érzelmeket egyre többen vizsgálják, a számtalan kutatási eredmény ellenére nincs egyetértés abban, hogy az érzelmeket mely akusztikus jellemzők alapján lehet azonosítani, illetve egymástól elkülöníteni [8]. Az mindenesetre igazolt, hogy passzív érzelmek (pl. bánat) esetén az alapfrekvencia (F0) átlaga, tartománya és szórása csökken, míg aktív érzelmek esetén (pl. harag, öröm) növekszik.

Mind az emóciók kifejezése, mind azok észlelése jelentős kulturális, nyelvi, nemi és nem utolsó sorban egyéni különbségeket mutatnak, ebből következően minőségi és mennyiségi megjelenésük is jelentős eltéréseket tükröznek [1].

Az érzelem kifejeződése a verbális tartalomban is jelentkezhet. Általában más szavakat használ egy mérges, mint egy nyugodt ember. Schuller és társai [14] által készített érzelemfelismerő a kombinált vokális és verbális információval pontosabb felismerési eredményt ért el. Természetesen léteznek olyan szituációk, ahol érzelemtől függetlenül azonos mondatok hangozhatnak el, ilyenkor csak a vokális üzenet alapján történhet az érzelem felismerése. Az általunk használt felismerési módszer nem használja föl a beszéd szöveges tartalmát. Ezt azzal indokolhatjuk, hogy ugyan valamivel gyengébb hatásfokkal, de képes az ember egy számára teljesen idegen nyelven beszélő ember érzelmi állapotát is megállapítani [12].

Fontos megemlíteni, hogy azokban a kísérletekben, ahol az alanynak előre megadott öt-hat érzelem alapján kellett számára ismeretlen beszélővel készült felvételeket osztályoznia, az emberi felismerési képesség körülbelül a 60%-ot érte el [8, 10]. Hasonló tesztekben a bemondók a saját érzelmeiket kb. 80%-ban ismerték fel helyesen [10].

Mivel a szakirodalomban nincs egységes álláspont, hogy konkrét emóciókat milyen információk alapján lehet hatékonyan, szabályok alapján megkülönböztetni, ezért mi is statisztikai módon közelítettük meg az érzelemfelismerést.

A géppel történő felismerés pontosságát is, mint akármelyik statisztikus mintaillesztési feladatban, döntően befolyásolják a választott jellemzők, illetve az ezekből összeállított tulajdonságvektor. Tudomásunk szerint az automatikus érzelemfelismerés területén magyar nyelvre vonatkozóan még nem publikáltak hatékonyan használható paramétereket, a külföldi publikációk alapján azonban igyekeztünk minél több akusztikus tulajdonságot összegyűjteni. A szakirodalom által javasolt általános jellemzőket (pl. alapfrekvenciából és intenzitásból származtatott statisztikák – [4, 9, 15]), nem találtuk elég hatékonynak, ezért szükségesnek éreztük, hogy a rengeteg fellelt akusztikus paraméter közül – az általunk kifejlesztett módon – válogassunk, és csak az optimálisnak talált jellemzőkből képzett tulajdonságvektorral dolgozzunk. Ez utóbbi a mintafelismerés szempontjából is kívánatos, hiszen így elkerüljük a túl komplex modellezést, és az ebből eredő problémákat.

A hasznosnak vélt és publikált jellemzők nagy mennyisége, főként arra vezethető vissza, hogy az eredmények nagymértékben függnek a felhasznált adatbázisoktól. Tovább bonyolítja a helyzetet, hogy a statisztikai alapon működő beszélőfüggő és beszélőfüggetlen érzelemfelismerés más-más paramétereket tart hasznosabbnak. Előbbi esetben sokkal jobb eredményt értek el [9, 14], hiszen a tanuló rendszernek nem kell az egyéni különbségekből adódó változatosságot elsajátítania. Másik nagyon fontos tényező az érzelmes felvételek forrása, spontán avagy mesterségesen, színészek által keltett érzelmekről van-e szó. Utóbbi esetben biztosabb a felismerés. A

pusztán a beszéd prozódiájából történő felismerés esetén hasznos, ha érzelmenként azonos szöveges tartalommal rendelkező felvételeket használhatnánk, így a megfelelőnek ítélt akusztikus jellemzők biztosan az érzelmek közötti prozódiai eltéréseket ragadnák meg. Ebben az esetben le kell mondanunk arról az igényről, hogy a tanításra használható adatbázis felvételei spontán érzelemkifejezést tartalmazzanak.

Érdemes kiemelni, hogy a felismerendő érzelmek számának növelésével a publikált eredmények drámaian romlanak. Például, egy telefonos beszédinterfészen keresztül irányított ügyfélszolgálatnak érdeke, hogy az ideges ügyfeleket valódi operátorokhoz kapcsolja. Ebben az esetben két érzelmi állapot elegendő a felismerési feladat szempontjából, a publikált eredmények 90% fölöttiek [7, 14]. Egy diagnosztikai rendszer esetében komplex érzelmeket kell kezelni, ezért tíznél is több állapot lenne szükséges, spontán, 5 osztályos klasszifikáció esetén az 50%-os hatékonyság is már jónak számít [7]. A csak prozódiai jellemzők alapján történő érzelemfelismerésről szóló kutatási eredmények túlnyomó többségében az alapérzelmek szintjén megállnak, nem elég hatékonyak, így az összetettebb érzelmek felismerése még várat magára.

Kutatásunk hat érzelmi állapot - *harag, szomorúság, undor, neutrális, öröm, meglepődés* - pusztán vokális információ alapján történő automatikus felismerését magyar felvételeken tűzte ki célul. Ezt kétféle, beszélőfüggő és beszélő független saját adatbázison végeztük el. Célunk volt érzelmekhez köthető jellemzők keresése, és az ezekkel elérhető hatásfok összehasonlítása a nemzetközi tapasztalatokkal. Bemutatjuk az általunk hasznosnak talált jellemzőket és a válogatásukra használt, nyilvános német adatbázison is tesztelt módszerünket. Az egyes adatbázisokon nyert akusztikus paraméterek közlése után ismertetjük a felismerőrendszerünkkel elért eredményeinket.

2 Adatbázisok

A kísérletekhez kétféle adatbázis készült, mindkettő 44.1 kHz-es mintavételezéssel és 16 bites kvantálással. Az első korpusz (HU_SI) 34 beszélőtől tartalmaz érzelmenként 2-3 példamondatot. Összesen 243 spontán bemondásból áll.

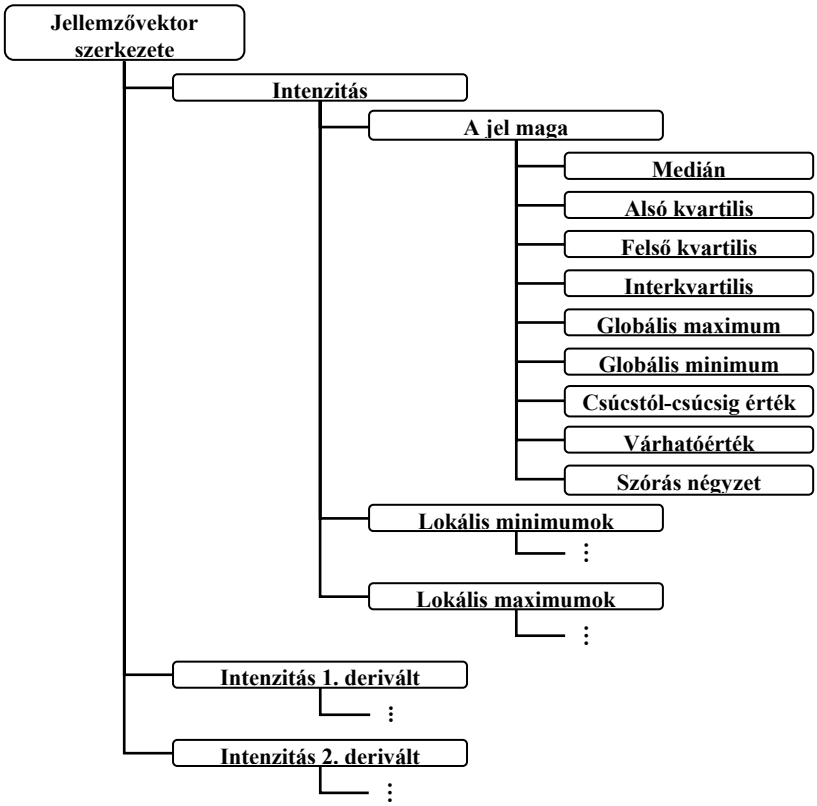
A második adatbázis csupán két beszélőtől felvett érzelmes mondatokból áll. Tartalmaz olyan nem-spontán bemondásokat, amelyek mindkét beszélőtől minden érzelemmel elhangoznak; érzelmenként különböző, de minkét beszélő esetében azonos tartalmú mondatokat; valamint spontán, érzelmenként és beszélőnként is különböző, egyedi mintákat, összesen 198 felvételt (HU_SD).

A szöveges tartalom nélküli megközelítés előnye, hogy lehetőség volt – apróbb módosítások után – német nyelvű, a Berlini Műszaki Egyetemen készült, nyilvános, érzelmes beszédadatbázison [3] is tanítani és tesztelni. A korpusz színészek által mesterségesen keltett, hétféle érzelemmel készült: *semleges, harag, félelem, öröm, bánat, undor, unalom*. Összesen 537 felvételből áll, és 10 beszélővel készült. A szöveges tartalom beszélőnként, érzelmenként azonos volt. (DE_SI)

3 Akusztikus jellemzők

A fellelhető szakirodalomban nem találni egyértelmű javaslatot a sikeres érzelemfelismeréshez szükséges jellemzőkre vonatkozóan. A legtöbb kutatási eredmény [2, 6, 14] a hosszúidejű jelszakaszokból (mondat, több szó; kb. néhány másodperc) nyert paraméterekből indul ki, így rendel minden egyes bemondáshoz egy jellemzővektort.
Általánosan alkalmazottak az alapfrekvencia (F0) és az energia (E) időjeleiből származtatott statisztikák (szórás, átlag, minimum, maximum stb.). A beszédjel energiáját általában további alsó és felső energiára osztják [16], a határ 4-600 Hz körüli. Fontos a beszéd sebessége és annak ingadozása is. A beszédfelismerési tapasztalatokból ismert, hogy a beszéd rövid idejű szakaszait (kb. 32 ezredmásodperc) igen tömören jellemzik a kepsztrális együtthatók (MFCC = Mel Frequency Cepstral Coefficient). Általános módszer, hogy ezekből az együtthatókból származtatott hosszúidejű statisztikákat is bevonják az érzelemfelismerésbe [7].

A fentiek alapján tehát adott bemondásra mértük a következő alábbi időjeleket: intenzitás, alsó energia, felső energia, alapfrekvencia, MFCC vektor hossza, 10 darab MFCC. Beszédfelismerőt alkalmazva lehetőség adódott az elhangzott szavak rejtett Markov-modellből történő kijelölésére, így a beszélő által egységnyi idő alatt kiejtett hangok és szavak mennyiségének (artikulációs sebesség és „szóráta") mérésére is. Számoltuk az első és a második deriváltakat (sebesség, gyorsulás) is. Ezekből a jelekből további „hosszú idejű" jeleket származtattunk. Ezzel az időjelek szélsőértékeinek változásit igyekeztünk figyelembe venni: lokális maximumok, lokális minimumok. Majd minden hosszúidejű jelen számoltuk a következő statisztikákat: medián, alsó kvartilis, (a legkisebb és a medián között középen elhelyezkedõ adat számértéke a rendezett mintában), felső kvartilis (hasonlóan a medián és a legnagyobb érték között van középen), interkvartilis (felső és alsó kvartilis különbsége), maximum, minimum, maximum és minimum különbsége (csúcstól-csúcsig érték), tapasztalati várható érték, tapasztalati szórás. Összesen 1377 (=17*3*3*9) darab jellemzőt vizsgáltunk (*1. ábra*).



1. Ábra: Az előállított jellemzővektor szerkezetének illusztrálása a beszédintenzitás jeléből származtatott statisztikákkal

4 A jellemzők válogatása

A beszédjelből nyert paraméterek vizsgálata egyenként történt. A Fisher-féle lineáris diszkrimináns analízisből ismert osztályok közötti és osztályon belüli variancia számítás alapján képzett hányadosok mutatják az egyes jellemzők szeparáló képességét. Esetünkben ennek alkalmazása úgy történt, hogy vettünk egy érzelmes osztályt (pl. harag), míg a többi érzelemhez tartozó adatokat összevontuk egy közös osztályba (pl. nem-harag). Ezután minden egyes jellemzőre kiszámoltuk az erre a két osztályra vonatkozó szeparáló képességet. Ezt minden érzelmi osztályra elvégeztük, majd a legdiszkriminálóbb jellemzőket gyűjtöttük össze az egyes szeparációvizsgálatból. Összesen 40 darab különböző jellemzőt.

Azért, hogy valóban a legjobb jellemzőket találjuk meg, a fent leírt módszert a keresztkiértékelésből ismert leave-one-out módszerrel használtuk. Beszélőfüggetlen esetben minden egyes tesztből kihagytunk egy beszélőt, beszélőfüggő esetben érzel-

menként az adatok 1/10-ét. Így például a HU_SI adatbázison 34 darab negyven elemű vektort kaptunk Végül csak az összes tesztben szereplő jellemzőket tartottuk meg. Az *1. táblázat*ban az egyes adatbázisokon ilyen módon nyert jellemzők számát láthatjuk.

1. Táblázat: Az egyes adatbázisokból kinyert leghasznosabb jellemzők száma

ADATBÁZIS	JELLEMZŐK SZÁMA
HU_SI	24
HU_SD	16
DE_SI	18

A *2.* és *3. táblázat*ban az egyes magyar adatbázisokon nyert néhány hasznos jellemző látható. Kaptunk olyan paramétereket, melyek a többszörös teszt alapján egyértelműen egy érzelem többitől való megkülönbözetésére szolgál, valamint olyan általános akusztikus tulajdonságokat, amik minden teszben jó szereparálási képességet mutat, de szorosan egyik osztályhoz sem köthető. Ami meglepő, hogy beszélőfüggetlen esetben csak az MFCC együtthatókból származtatott statisztikákat találunk, beszélőfüggő esetben a paraméterek 1/5-e az intenzitásból származik.

2. Táblázat: Néhány a beszélőfüggetlen (HU_SI) adatbázison nyert érzelemhez köthető és általánosan jól teljesítő jelparaméter

Harag	3. MFCC szórása MFCC vektor hossz szórása
Undor	10. MFCC 2. deriváltjának csúcstól-csúcsig értéke 10. MFCC 2. deriváltjának szórása
Öröm	10. MFCC maximumainak mediánja 10. MFCC alsó kvartilise
Neutrális	1. MFCC felső kvartilise 1. MFCC maximumainak mediánja
Szomorúság	MFCC vektor hosszának maximumainak szórása 10. MFCC szórása
Meglepődés	1. MFCC maximumainak csúcstól-csúcsig értéke 1. MFCC maximumainak minimuma
Általános	10. MFCC együttható maximumainak szórása 9. MFCC 2. deriváltjának felsőkvartilise 1. MFCC együttható maximumainak mediánja 1. MFCC együttható felső kvartilise

3. Táblázat: Néhány a beszélőfüggő (HU_SD) adatbázison nyert érzelemhez köthető és általánosan jól teljesítő jelparaméter

Harag	intenzitás medián
Undor	MFCC vektor hosszának alsó kvartilise intenzitás maximumainak mediánja
Öröm	intenzitás maximumainak alsó kvartilise
Neutrális	felső energia mediánja
Szomorúság	*nem találtunk egyértelmű jellemzőt*
Meglepődés	MFCC vektor hosszának alsó kvartilise
Általános	MFCC vektor hosszának csúcstól-csúcsig értéke MFCC vektor hosszának minimumainak alsókvartilise Felső energia 1. deriváltjának alsó kvartilise

5 Tanítás és felismerés

Adott felvételből képzett vektort a Bayes-döntéssel soroltunk egyik vagy másik érzelmes osztályba, azaz a legnagyobb valószínűségű érzelemre döntöttünk, az egyes érzelmek valószínűségét azonosnak tételeztük fel.

$$\hat{C} = \arg\max_i \{P(C_i|z)\} = \arg\max_i \{P(z|C_i)P(C_i)\}$$

Ahol z jelenti a döntés előtt álló, beérkezett vektort, C_i pedig az egyes érzelmi osztályokat. A döntéshez szükséges feltételes eloszlásfüggvényeket Gauss függvények keverékével (Gaussian Mixture Modell = GMM) becsültük. A válogatott mennyiségű jellemzőkön az alábbi transzformációk elvégzése után kapott vektorokkal tanítottuk az egyes érzelmek modelljeit. A tanítás és felismerés során alkalmazott lépéseket a *2. ábra* foglalja össze.

Standardizálás: A tanításhoz használt adatok alapján egységnyi szórásúvá és nulla várhatóértékűvé tettük az egyes dimenziókat, standardizáltuk az adatokat.

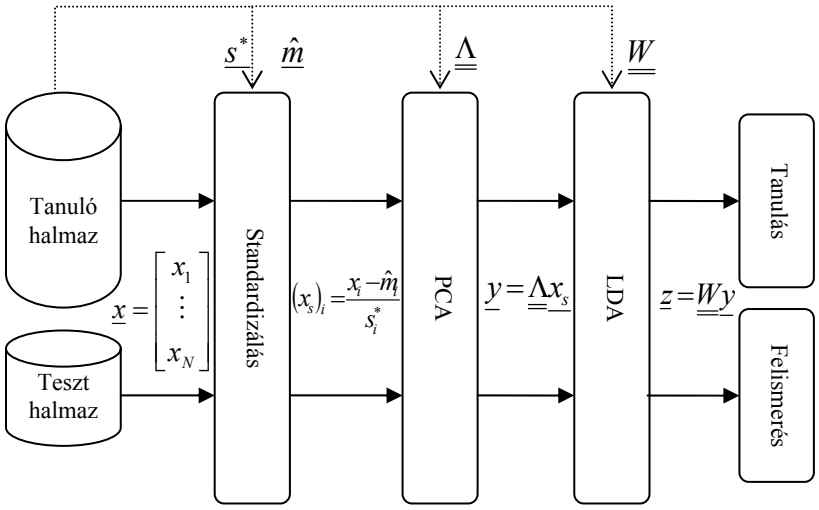
Főkomponens analízis (PCA): A kiválasztott jellemzők között előfordulhatnak olyanok, melyek között szoros összefüggés, korreláció lehet. A standardizált adatok korrelációs mátrixa az alábbi módon becsülhető.

$$\underline{\underline{R}} = \frac{1}{n-1}\sum_{\underline{x}} \underline{x}_s \underline{x}_s^T$$

Érdemes a paraméterek számát oly módon csökkenteni, hogy a túlzottan korreláló paraméterek helyett csak azok valamilyen lineáris kombinációját tartjuk meg. Az ilyen kapcsolatok feltárására, ezáltal dimenziócsökkentésre használhatjuk a korreláci-

ós mátrix legnagyobb sajátértékeihez tartozó sajátvektorok (főkomponensek) alapján képzett transzformációs mátrixot ($\underline{\Delta}$). Erre többek között azért van szükség, mert a következő lépés numerikus problémákat vet fel, ha az adatok túlságosan korrelálnak [13].

LDA: A főkomponens analízis után az adatvektorokat kisebb dimenziójú térbe vetítettük a Fisher-féle diszkrimináns analízisnek (LDA) megfelelően [5] kapott mátrix segítségével ($\underline{W}$). Az így nyert ötdimenziós vektorokkal végeztük a tanítást – ahol 1 illetve 2 Gauss függvény keverékével próbáltuk a sűrűségfüggvényeket közelíteni – és a felismerést.



2. Ábra: A válogatott jellemzővektoron képzett transzformációk tanítás és felismerés előtt

A gépi felismerő rendszerek teljesítőképessége keresztkiértékeléssel jellemezhető. Esetünkben ez azt jelenti, hogy például a 34 beszélővel készített adatbázison, 34 tanítási és felismerési tesztet futtatunk, az egyik beszélőt mindig kihagyva a tanításból, a felismerési teszteket pedig a kihagyott beszélő adatain mértük. A 34 teszt eredményét átlagolva kaptuk meg a rendszerünk felismerési eredményét.

6 Eredmények

A sokbeszélős magyar adatbázison (HU_SI) elért beszélőfüggetlen eredmények a *4. táblázat*ban láthatók, az átlagos felismerési pontosság 42,9%. Figyelembe véve, hogy nem színészek által produkált érzelmeket hordozó, tartalmilag kötetlen felvételekről van szó, az eredmény a nemzetközi publikációkkal összemérhető, és az emberi közelítőleg 60%-os hatásfokhoz képest is bíztató.

4. Táblázat: Magyar, beszélőfüggetlen (HU_SI) érzelemfelismerés eredménye

Érzelem	Felismerési arány [%]
Harag	42,7
Undor	43,5
Öröm	33,3
Neutrális	62,0
Szomorúság	36,7
Meglepődés	39,0
Átlag	**42,9**

A beszélőfüggő esetben – ahol beszélőnként külön-külön tanítottunk és teszteltünk, majd a független eredmények átlagát vettük – felismerőnk az alábbi eredményeket mutatta. (*5. táblázat*).

5. Táblázat: Kétbeszélős magyar adatbázison (HU_SD) elért átlagos felismerési hatásfokok

Érzelem	Felismerési arány [%]
Harag	50,0
Undor	80,0
Öröm	80,0
Neutrális	60,0
Szomorúság	53,3
Meglepődés	38,0
Átlag	**60,2**

Ebben az esetben a felismerő sokkal jobban teljesített, érzelmi kategóriánként átlagolva 60% körül. A kevesebb tanítóminta dacára a beszélőfüggő felismerési eredmények lényegesen jobbak lettek.

Azért, hogy rendszerünket másokéval is összehasonlíthassuk, a kísérleteket lefuttattuk a német adatbázison is (*6. táblázat*).

6. Táblázat: Tízbeszélős, német adatbázison (DE_SI) elért felismerési eredmények

Érzelem	Felismerési arány [%]
Harag	65,6
Unalom	76,5
Undor	80,3
Félelem	73,0
Öröm	51,3
Semleges	73,7
Bánat	82,0
Átlag	**71,8**

Meglepően magas felismerési eredményt sikerült elérni, mely a nemzetközi irodalomban használt komplexebb tanuló rendszerek (például SVM) eredményeivel is összevethető [14]. Véleményünk szerint ez a magas felismerési eredmény annak köszönhető, hogy az adatbázisban kötött a szöveges tartalom, és ez korlátozza az érzelemkifejezés lehetőségeit. Nem szabad elfelejteni azt sem, hogy itt színészek által produkált felvételekről van szó, melyek nem adhatják vissza az egyes érzelmek teljes skáláját. Általában is elmondható, hogy a színészekkel készült felvételek „hevesebb" érzelmeket tartalmaznak.

7 Összefoglalás

Megmutattuk, hogy statisztikai módszerekkel, pusztán a beszéd akusztikus jellemzői alapján, a szöveges tartalom figyelembe vétele nélkül megfelelő érzelemfelismerési eredményeket lehet elérni. Ez különösen beszélőfüggő esetben lehet igen hatékony. Annak érdekében, hogy ilyenkor ne kelljen egy teljesen új felismerőt betanítani, amihez sok adat kell, érdemes lenne a beszédfelismerésnél is gyakran használt beszélőadaptációt alkalmazni – ebben az irányban tervezzük a további vizsgálatokat.

Az eredmények alapján arra is következtethetünk, hogy az „amatőrök" és a színészek által keltett beszéd érzelmi töltete különböző jellegű, melyek közül az utóbbinak a felismerése jóval eredményesebb lehet.

8 Köszönetnyilvánítás

A kutatást az NKFP-2/034/2004-es projekt keretében az NKTH támogatta.

Bibliográfia

1. Bernáth, László – Révész, György 1994. A pszichológia alapjai. Tertia, Budapest, 1994.
2. Blouin, Christophe - Maffiolo, Valerie 2005. A study on the automatic detection and characterization of emotion in a voice service context. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal 469-472.
3. Burkhardt, Felix - Paeschke, Astrid – Rolfes, Miriam – Sendlmeier, Walter - Weiss, Benjamin 2005. A Database of German Emotional Speech, In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 1517-1520.
4. Cichosz, Jaroslaw – Slot, Krzysztof 2005. Low-Dimensional Feature Space Derivation for Emotion Recognition. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 477-480.
5. Duda, Richard O. - Hart, Peter E. - Stork, David G. Pattern Classification (Second Edition) 2000. John Wiley & Sons Inc, ISBN: 0-471-05669-3, New York
6. Fernandez, Raul – Picard, Rosalind W. 2005. Classical and Novel Discriminant Features for Affect Recognition from Speech. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 473-476.
7. Kwon, Oh-Wook – Chan, Kwokleung – Hao, Jiucang – Lee, Te-Won 2003. Emotion Recognition by Speech Signals. In Proceedings of EUROSPEECH-2003. Geneva, Switzerland, 125-128.
8. Laukka, Petri 2004. Vocal Expression of Emotion, PhD Thesis. Uppsala University, Uppsala.
9. Luengo, Iker - Navas, Eva - Hernáez, Inmaculada – Sánchez, Jon 2005. Automatic Emotion Recognition using Prosodic Parameters. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 493-496.
10. Petrushin, Valery A. 2000. Emotion recognition in speech signal: experimental study, development, and application. In Proceedings of ICSLP-2000. vol.2. Beijing, China, 222-225.
11. Scherer, Klaus R. 2000. A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology. In Proceedings of ICSLP-2000, vol.2. Beijing, China, 379-382.
12. Scherer, Klaus R – Banse, Rainer - Wallbott, Harald G. 2001. Emotion Inferences from Vocal Expression Correlate Across Language and Cultures. Journal of Cross-Cultural Psychology. vol. 32. No. 1. 76-92.
13. Schlüter, Ralf - Zolnay, András - Ney, Hermann 2006. Feture Combination using Linear Discriminant Analysis and its Pitfalls. In Proceedings of INTERSPEECH-2006. Pittsburgh, Pennsylvania, 345-348.
14. Schuller, Björn – Müller, Ronald - Land, Manfred – Rigoll, Gerhard 2005. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features Within Ensenmbles. In Proceedings of INTERSPEECH-2005. Lisbon, Portugal, 805-808.
15. Ververidis, Dimitrios – Kotropoulos, Constantine – Pitas, Ioannis 2004. Automatic emotional speech classification. In Proceedings of ICASSP'04. vol. 1. Philadelphia, Pennsylvania, 593-596.
16. Ververidis, Dimitros - Kotropoulos, Constantine 2004. Automatic Speech Classification to five emotional states based on gender information. In Proceedings of EUSIPCO-2004. Vienna, Austria, 41-344.

III. Morfo-fonológia a beszédfeldolgozásban

Statisztikai és szabály alapú morfológiai elemzők kombinációja beszédfelismerő alkalmazáshoz

Németh Bottyán[1], Mihajlik Péter[1], Tikk Domonkos[1], Trón Viktor[2]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem, TMIT, Budapest Magyar Tudósok Körútja 2. 1117,
{bottyan, mihajlik, tikk}@tmit.bme.hu
[2] International Graduate College Language Technology and Cognitive Systems
University of Edinburgh and Saarland University,
v.tron@ed.ac.uk

Kivonat: A magyar nyelvű számítógépes beszédfelismerésnél célszerűnek tűnik, hogy ne a szavakat, hanem a morfémákat vegyük alapegységnek a nyelvi modell felépítéséhez. Ehhez viszont szükséges, hogy a szavakat a morfémáknak megfelelő szegmentumokra bontsuk. A cikk egy új szegmentálási technikát ismertet, ami két különböző morfológiai szegmentáló módszer egyesítéséből született, és mindkét ősénél jobban alkalmazható számítógépes beszédfelismeréshez. Ennek a rendszernek az egyik pillére egy szabály alapú morfológiai elemző, a hunmorph, a másik pedig egy statisztikai alapokra épülő morfológiai szegmentáló, a morfessor. A kompozíció során igyekeztünk mindkét rendszer előnyeit megtartani, hátrányos tulajdonságait orvosolni. Ez nagyrészt sikerült is, leszámítva, hogy a morfessor által biztosított nyelvfüggetlenség a hunmorph bevonásával elveszett.

1 Bevezetés

A számítógépes beszédfelismerésben általában szó alapú nyelvi modellt használnak a felismeréshez. Ez a magyar nyelv esetén, ahol egy szónak rengeteg különböző alakja lehet, nem tűnik a legésszerűbb megoldásnak. Azt várnánk, hogy a morfémákra épülő nyelvi modellel pontosabb felismerési eredményeket érhetünk el. Elsőként [7] alkalmazott magyar nyelvű beszédfelismerésnél (diktáló rendszerben) morféma alapegységeket, azonban a felismerés pontosságát nem vetette össze a szó alapú megközelítésével. Később [8] szintén próbálkozott a morféma alapú nyelvi modellezéssel orvosi diktáló rendszerben, de még a morféma felismerési pontossága is lényegesen gyengébbnek adódott a szó alapú felismerésnél mért szófelismerési pontosságnál. Azonos metrikával előttünk nem hasonlították össze a morféma és szó alapú magyar nyelvű beszédfelismerési eredményeket. Más nyelveken számos sikeres kísérletet végeztek morféma alapú nyelvi modellek beszédfelismerési alkalmazásával [4], [5].

Mi először a morfessort [1], egy statisztikai tanuló algoritmus alapján működő szegmentálót próbáltuk ki. A kapott felismerési eredmények jobbak az egyszerű szó alapú felismerésnél, de a kapott szegmentumok nem feltétlenül felelnek meg valós

nyelvtani elemeknek, és ha mégis, akkor sem tudjuk a kapott morfémák nyelvtani jelentését [9], [3].

A további fejlesztések szempontjából viszont az aktuális felismerési pontosság mellett fontos, hogy a szegmentumok jelentéssel bíró egységek legyenek, mert ez teszi lehetővé, hogy további nyelvfeldolgozási szinteket építsünk a beszédfelismerő fölé. Így kipróbáltuk a hunmorph-ot [6] is a szegmentumok előállítására. A hunmorph eredetileg morfológiai elemzésére lett kialakítva, nem pedig szegmentálásra, így ehhez először némi átalakításra volt szükség, hogy a szegmentumokat is előállítsa a program. A hunmorph használatánál további problémát jelentett, hogy a program néha összekötött két külön morfémának megfelelő szegmentumot, illetve egy szóra nagyon sok alternatív elemzési megoldást kínált. Az alternatívák közül valamilyen egyszerű heurisztikával igyekeztünk kiválasztani egyet (pl.: a leghosszabb elemzés). Feltehetően ez utóbbi okok miatt a hunmorph-os szegmentálásra épülő beszédfelismerő eredményei rosszabbak voltak a morfessorra épülőénél, viszont még mindig jobbak, mint a pusztán szó alapú elemző eredményei.

Szerettük volna a szabály alapú szegmentáló gyengeségeit kiküszöbölni, mivel az volt az alapfeltevésünk, hogy a természetes morfémahatárokat használatával jobb eredményeket érhetünk el. Hogy javítsunk a hunmorph-os szegmentálás eredményein, megpróbáltuk a felajánlott alternatívák közül a statisztikailag legértelmesebbet kiválasztani. Ehhez a választáshoz a morfessorban a szegmentálás tanulására használt módszert alkalmaztuk.

2 Szegmentáló használata a beszédfelismerőben

A beszédfelismerőkben használt nyelvi modellek általában szó alapúak, vagyis a felismerés alapegysége a szó. Az agglutináló nyelvek esetén ez a megközelítés jelentős hátrányokkal jár. Mivel egy szónak rendkívül sok különböző alakja lehet, amit ez esetben mind különböző szónak kell tekintenünk, a beszédfelismerőnek egy igen nagyméretű szótárban kell keresnie. Nagy szótár esetén a felismerés során sok lehetséges megoldás közül kell választanunk, és ráadásuk az egyes szavakra jutó tanítópéldák száma is kicsi lesz.

A problémák leküzdéséhez érdemes a morfémákat választani a felismerés alapegységeként. Azonban ennek a megközelítésnek is megvannak a maga buktatói. Először is a szavakat morfémákra kell bontani, ami korántsem magától értetődő feladat. Vannak olyan szóelemző alkalmazások, amelyek a szótő mellett a szóhoz kapcsolódó morfémákat is meghatározzák, de ezek nem a beszédfelismeréshez lettek fejlesztve, és nem adják meg, hogy a szóban mely betűk (még érdekesebb, mely hangok) felelnek meg az egyes nyelvtani elemeknek. Csak ízelítőnek nézzünk néhány problémásabb esetet.

1. Táblázat: A szótő nem állítható elő szegmentumként

Szó	**Elemzés**
lenniük	van <INF><PERS><PLUR>
hass	hat <SUBJUNC-IMP><PERS<2>>
arannyal	arany <CAS<INS>>
hússzor	húsz [MULTIPL-ITER]/ADV
borókásban	boróka [ATTRIB]/ADJ<CAS<INE>>

Amikor szegmentálni szeretnénk egy szót, sokszor az eredeti szótőt nem kapjuk vissza, néha az egész szótő, gyakran csak a szótő vége módosul. Kérdés az is, hogy ha kötőhangot használunk a toldalékoláskor, azt melyik részhez kapcsoljuk.

Egy szónak gyakran több lehetséges jelentése és ezzel együtt több lehetséges szegmentálása van. Ezt a problémát valamilyen egyértelműsítés alkalmazásával lehet megoldani. Az emberek a szövegkörnyezet alapján könnyen meg tudják határozni, hogy mi az éppen megfelelő elemzés, de a számítógépnek ez bonyolultabb feladat. De nem csak a feladat nehézsége jelent gondot, hanem az is, hogy a beszédfelismeréshez használt (trigram) nyelvi modell nem alkalmas az ilyen jellegű egyértelműsítésre, ezért valamilyen egyszerűbb szabály alapján kell választanunk a különböző lehetőségek közül. A legegyszerűbb ilyen lehetőségek a leghosszabb, legrövidebb vagy a legelső elemzés választása. Ésszerű lenne a leggyakoribb megoldás választása, de nem állt rendelkezésünkre egyértelműsítő, így nem tudtunk gyakoriságot számolni a tanítókorpuszon.

Tegyük fel, hogy rendelkezésre áll a morféma alapú nyelvi modell és a hozzá megalkotott a beszédfelismerő, és feldolgozunk vele egy felvételt. Ekkor a kimeneten egy morfémasort fogunk kapni, amit még nem elég, mert mi kimenetként szavakat várunk a beszédfelismerőtől. Hogy ezt megtehessük, egy egyszerű trükköt alkalmazunk. Bevezettünk egy új szimbólumot (#), ami a szóhatárokat jelöli. A nyelvi modell építésekor ezt egyszerűen egy új morfémának tekinthetjük. Így a felismerési eredmény tartalmazza a szóhatárokat, és a szavak könnyen összerakhatók. Ennek a módszernek annyi hátránya van, hogy a szóhatároknál csökkenti a modell kontextusérzékenységét. A hatás különösen akkor jelentős, ha a szavak átlagosan kevés morfémából épülnek fel. (A rendelkezésünkre álló korpuszban a szavankénti morféma szám 1,6 körül van.) Ha ellensúlyozni szeretnénk a hatást, akkor hosszabb kontextust kellene figyelembe vennünk, ami jelentősen növeli a szükséges számításokat, ezért ezt nem alkalmaztuk.

3 Az eredeti módszerek

3.1 Hunmorph

A hunmorph egy nyílt forráskódú morfológiai elemző. A morphdb.hu morfológiai szótár segítségével elemzi a szavakat, és előállítja azok morfológiai elemzését, vagyis

meghatározza és annotálja a szótövet és a toldalékokat. Ahhoz, hogy az elemzést használni lehessen a beszédfelismerésben, olyan módon kellett átalakítani a szoftvert, hogy a leválasztott toldalékokat az aktuális szóban szerepelő formában adja vissza. A következő példa szemlélteti a kimeneten véghezvitt változtatást.

Az „odatették” eredeti elemzése:

```
oda/PREV+tesz/VERB<PAST><PLUR><DEF>
```

És az ennek megfelelő szegmentálást is tartalmazó elemzés (szegmentumok a „{}” jelek között):

```
{odate} oda/PREV+tesz/VERB {tték} <PAST><PLUR><DEF>
```

Természetesen az elemző teljes átalakítására nem volt lehetőség, a szegmentumokat olyan formában állnak elő, ahogy az a program belső működésének leginkább megfelel. Így az eredmény néha eltér az intuitív megoldástól, és elég sajátosnak tűnhet.

2. Táblázat: Problémák a szegmentálással

Szó	**Szegmentálás**	**Hunmorph kimenete**
kossal	ko-ssal	{ko} kos /NOUN {ssal} <CAS<INS>>
ontással	ont-á-ssal	{ont} ont {á} /VERB[GERUND]/NOUN {ssal} <CAS<INS>>
állásomban	áll-ás-omban	{áll} áll {ás} /VERB[GERUND]/NOUN {omban} <POSS<1>><CAS<INE>>
elfordította	el-fordít-otta	{el} el/PREV+ {fordít} + fordít {otta} /VERB<PAST><DEF>

A 2. táblázat első két példája a szavak és toldalékok néhol furcsa szétválasztását mutatja be. Ez a probléma elő-előfordul ugyan, de a szegmentálások jelentős része jóval közelebb áll az emberi intuícióhoz (legalábbis valamelyik a felajánlott szegmentálások közül). A második két eset azt mutatja be, hogy a program gyakran nem vág szét olyan toldalékokat, amiket még tovább lehetne darabolni. Ezekre jellemző, hogy ugyan egy szegmentumként szerepelnek a kimeneten, de a szegmentum után megtalálható mindkét részhez tartozó nyelvtani címke. Ezt a megfigyelést kihasználva egy utólagos feldolgozással e címkék jelentős részét tovább tudjuk darabolni. Ennek lényege, hogy eltároljuk az összes olyan szegmentumot, amihez csak egy címke tartozik. Nevezzük ezeket atomi szegmentumoknak. Ezek után az összetett szegmentumokhoz keresünk olyan atomi szegmentumokat, amelyek címkéje megtalálható az összetett szegmentumhoz tartozó címkék között, és a szegmentum egy részét lefedi. Ha sikerül egy szegmentumot teljesen és átlapolódásmentesen lefedni atomi szegmentumokkal, akkor a fedésnek megfelelően feldaraboljuk azt.

3.2 Morfessor

A morfessor egy statisztikai módszerek alapján működő szegmentáló program. Eredetileg finn nyelvre fejlesztették ki, mert a finnben egy szónak annyira sok alakja lehetséges, hogy egy szabály alapú morfológiai elemző elkészítése túl bonyolult lenne. A

program pusztán egy címkézetlen korpuszból képes megtanulni, hogy hogyan kell szegmentálni egy adott szöveget, és semmilyen nyelvspecifikus szabályt nem tartalmaz, ezért alkalmazása igen egyszerű volt számunkra. További előnye a szabály alapú vetélytársával szemben, hogy minden egyes szóra csak egyetlen szegmentálást ad, vagyis az egyértelműsítés problémáját kiküszöböli.

Érdemes kicsit közelebbről is megnézni, hogy milyen elvek alapján működik a morfessor, már csak azért is, hogy később jobban megérthessük a két módszer kombinálásának lényegét. A következőkben a [2]-re fogunk támaszkodni. Az „alap" morfessornak több javított változata is készül, de mi csak a legegyszerűbb verzióval fogunk részletesebben foglalkozni. Természetesen a méréseknél az összes verziót kipróbáltuk, és a morfessor eredményeként ezek közül a legjobbat közöljük.

A program feladata, hogy előállítsunk egy nyelvi modellt egy címkézetlen korpuszból. A modell pedig nem más, mint a morfémák egy halmaza és egy rajtuk értelmezett nyelvtan. Tehát egy olyan modellt szeretnénk találni felügyelet nélküli tanulással, aminek segítségével tömören le tudjuk írni a tanulókorpuszt és ráadásul a morfémakészletünk is tömör marad. A probléma megfogalmazható egy maximális a poszteriori paraméterbecslési feladatként (MAP):

$$\begin{gathered}\arg\max_M P(M \mid korp) = \arg\max_M P(korp \mid M)P(M), ahol \\ P(M) = P(szótár, nyelvt)\end{gathered} \tag{1.1}$$

A MAP becslés két részből áll: a nyelvi modell valószínűségéből és a korpusz modellre vetített feltételes valószínűségének maximum likelihood becsléséből. Látható az is, hogy a modell valószínűsége a szótár és a nyelvtan együttes valószínűségével egyezik meg. Érdemes megjegyezni, hogy a képletben szereplő valószínűségek bayesi értelemben vett valószínűségek, tehát nem előfordulások valószínűségei, hanem priori hiedelmek bizonyosságát fejezik ki.

Annak a valószínűsége, hogy egy adott szótárat állítsunk elő, a szótárban szereplő morfémák együttes valószínűségével arányos, pontosabban, ha *M* a különböző morfémák száma, akkor *M!*-szorosa, mert a morfémák ennyiféle különböző sorrendben kerülhetnek a szótárba. A morfémáknak két tulajdonsága van, amelyek befolyásolhatják a szótár valószínűségét: a morfémák előfordulási gyakorisága és az őket felépítő betűsor, ami tartalmazza a morféma hosszát is. Így a szótár valószínűségére a következő képletet kapjuk, ahol $s_{\mu i}$ a μ_i morfémát reprezentáló karaktersor és $f_{\mu i}$ a morféma előfordulási gyakorisága.

$$P(szótár) = M!P(f_{\mu_1}, f_{\mu_2}, f_{\mu_3}, ...)P(s_{\mu_1}, s_{\mu_2}, s_{\mu_3}, ...) \tag{1.2}$$

Az előfordulási gyakoriságokból adódó tagot az egész szótárra globálisan tudjuk számolni. Legyen *N* az összes morféma összes előfordulásának száma.

$$P(f_{\mu_1}, f_{\mu_2}, f_{\mu_3}, ...) = 1/\binom{N-1}{M-1} = \frac{(M-1)!(N-M)!}{(N-1)!} \tag{1.3}$$

A második tag számolásához azzal az egyszerűsítő feltevéssel élünk, hogy a morfémákat alkotó karaktersorozat független a többi morfémát alkotó karaktersorozattól. Így az együttes valószínűség megegyezik a morfémák valószínűségének szorzatával. Feltesszük továbbá azt is, hogy a morfémákat alkotó karakterek valószínűségei is függetlenek egymástól, és így a morféma valószínűsége az őt alkotó karakterek valószínűségének szorzatával azonos.

$$P\left(s_{\mu_1}, s_{\mu_2}, s_{\mu_3}, \ldots\right) = \prod_{i=1}^{M} P\left(s_{\mu_i}\right), \acute{e}s$$
$$P\left(s_{\mu_i}\right) = \prod_{j=1}^{l_{\mu_i}} P\left(c_{ij}\right) \tag{1.4}$$

A morféma hosszát implicit modellezzük a már említett morfémavég szimbólum (#) bevezetésével, amit minden morféma végére odaírunk a szótárban. A „#” valószínűségéből könnyen számolható egy *l* hosszú morféma valószínűsége, mivel a morféma először tartalmaz *l* „#”-től különböző szimbólumot végül pedig egy „#”-t. Ennek valószínűsége egy szimpla exponenciális eloszlásból adódik.

$$P(l) = \left[1 - P(\#)\right]^{l} P(\#) \tag{1.5}$$

A nyelvtan azt határozza meg, hogyan lehet az egyes nyelvi elemeket kombinálni. A legegyszerűbb morfessor egyáltalán nem veszi figyelembe a kontextus az elemek kombinációjánál, ezért nem is igazán lehet nyelvtanról beszélni, vagyis a szótár és a nyelvtan együttes valószínűsége a szótár valószínűségére redukálódik. Ez azt is jelenti, hogy egy morféma ugyanolyan valószínűséggel fordulhat elő bármilyen morféma után, vagy a szó elején és végén. A morféma gyakorisága tehát egy egyszerű maximum likelihood becslés. Ha f_μ a μ morféma előfordulási gyakorisága, akkor ez így írható le:

$$P(\mu_i) = \frac{f_{\mu_i}}{N} = \frac{f_{\mu_i}}{\sum_{j=1}^{M} f_{\mu_j}}. \tag{1.6}$$

A korpusz összes szava felbontható a szótárban található morfémákra, gyakran több felbontás is lehetséges. A MAP modellt használva mindig a legvalószínűbb felbontást fogjuk választani. A korpusz valószínűsége egy adott nyelvi modell esetén a következő, ahol *W* a szavak száma és n_j a *j.* szóban található morfémák száma, a morfémák előfordulási valószínűsége pedig az (1.6) képlet alapján számolandó:

$$P(korp \mid M) = \prod_{j=1}^{W} \prod_{k=1}^{n_j} P(\mu_{jk}) \,. \tag{1.7}$$

Az eddigi képletek meghatározzák a szegmentálás tanulása során maximalizálandó függvényt. A maximum megtalálásához egy mohó algoritmust javasoltak. Induláskor a szótár a korpuszban található szavak halmaza. A tanulás során az algoritmus sorban veszi a szavakat, és megpróbálja őket különféle módon szegmentálni, majd az alternatívák közül a legnagyobb valószínűségűt tartja meg. A módszert mindaddig folytatjuk, amíg szignifikáns javulást tapasztalunk.

A keresés során nem közvetlenül a valószínűségeket számoljuk, hanem azok (kódhosszként is értelmezhető) negatív logaritmusát, mert így a szorzás helyett összeadást tudunk alkalmazni.

Az algoritmus egy speciális adatstruktúrát használ, ahol a korpusz minden egyes szavának megvan a saját bináris vágási fája. A fában a levelek, vagyis azok az elemek, amelyek nincsenek tovább darabolva, felelnek meg a szótárban szereplő morfémáknak, és csak ők számítanak bele a kódhosszba. Minden egyes csomópontnál tároljuk az adott elem előfordulási gyakoriságát, ami megegyezik a szülők előfordulási gyakoriságának összegével. Továbbá minden csomópont csak egyszer szerepelhet ebben az adatstruktúrában, tehát ha két szó vágási fájában egy részfa átlapolódik, akkor azt a részfát csak egyszer tároljuk el.

Az algoritmus fő művelete a csomópont újradarabolása. Ez egy rekurzív művelet, amely először megkeresi, hogy az adott csomópontot hol kell kettévágni ahhoz, hogy a legjobb értékeket kapjuk, majd a kapott két csomópontot tovább darabolja. Ezt mindaddig folytatja, amíg lehet olyan vágást találni, amely csökkenti a szótár teljes kódhosszát. Tehát ahogy említettük, induláskor maguk a szavak a morfémák, majd véletlen sorrendben végigmegyünk az összes szón és újradaraboljuk őket. Ha végeztünk az összes szóval, akkor kezdjük elölről, mindaddig, amíg egy megadott korlátnál jobban tudjuk csökkenteni a globális kódhosszt.

```
újravágás(csp)
// EGY CSOMÓPONT EGY SZÓNAK VAGY EGY DARABJÁNAK FELEL MEG
// 1. ELTÁVOLÍTJUK AZ AKTUÁLIS REPREZENTÁCIÓJÁT A CSP.-NEK
if csp megtalálható a struktúrában then
   for all a csp-ben gyökerez• részfa elemeire (m) do
      csökkentsd számláló(m) számláló(csp)-vel
      if m levél vagyis egy morféma then
         csökkentsd az L(korp | M) -et és L(f_μ1, f_μ2, ...) -et
      if számláló(m) = 0 then
        m eltávolítása
        if m levél then
           csökkentsd L(s_μ1, s_μ2, ...) -et
```

```
// EL•SZÖR MEGPRÓBÁLJUK AZ EGÉSZ SZÓT EGY MORFÉMÁNAK TEKINTENI
csp-t levélként visszarakjuk a struktúrába számló(csp)-vel
növeld L(korp|M)-et és L(f_μ1, f_μ2, ...)-et
növeld L(s_μ1, s_μ2, ...)-et
legjobb megoldás ← [L(korp|M); csp]

// PRÓBÁLJUK KI A CSP. ÖSSZES LEHETSÉGES KÉT RÉSZRE VÁGÁSÁT
Távolítsd el csp-t L(M|korp)-ból, de hagyjuk az adatstruktúrában
Mentsük el az állapotot X-be
for all pre + suf = csp do
   for m in [pre; suf] do
      if az adatstruktúra tartalmazza m-et then
         for all n, m-ben gyökerez• részfa csomópontra do
            növeld számláló(n) számláló(m)-mel
            if n levél then
               növeld L(korp|M)-et és L(f_μ1, f_μ2, ...)-et
      else
         add hozzá m-et az adatstruktúrához számláló(m)-mel
         növeld L(korp|M)-et és L(f_μ1, f_μ2, ...)-et
         növeld L(s_μ1, s_μ2, ...)-et
   if L(M|korp) < legjobb megoldás then
      legjobb megoldás ← [L(korp|M); pre; suf]
   állítsuk vissza az X-be elmentett állapotot

// VÁLASSZUK KI A LEGJOBB VÁGÁST VAGY „NEM VÁGÁST"
Állítsuk az adatstruktúrát és az L(M|korp)-t a „legjobb megoldás"-nak megfelel•en
if pre + suf = m vágás történt then
   pre és suf szül•jének állítsuk be m-et

// FOLYTASSUK A VÁGÁST REKURZÍVAN
újravágás(pre)
újravágás(suf)
```

3 A hibrid megoldás

Amint az első kísérletek után kiderült, hogy a beszédfelismerésben a morfessorral készített szegmentálás jobb eredményt ad, elkezdtünk gondolkodni, hogyan lehet egyesíteni a két módszert. Arra gyanakodtunk, hogy a hunmorph gyengesége abból adódik, hogy nincs megfelelő módszerünk az általa generált lehetőségek közül a legjobbat kiválasztani. De valójában már azzal is megelégedtünk volna, ha a hunmorph segítségével hasonló eredményeket tudunk elérni, mint a morfessorral, mert a szabály alapú rendszer esetén, mintegy melléktermékként megkapjuk a szó nyelvtani elemzését is, amiből reményeink szerint későbbiekben felhasználhatunk. Ezért egy hibrid megoldás elkészítése mellett döntöttünk. Az alapötlet ehhez igen egyszerű: használjuk a morfessor jól bevált statisztikai modelljét a hunmorph által felkínált szegmentálások közül a legmegfelelőbb kiválasztására.

A cél érdekében némileg változtattunk a morfessor kódján. Módosításunk lényege, hogy a szavak újravágásakor a költségfüggvényt kiszámoljuk az összes hunmorph által javasolt alternatívára, de nem az összes lehetséges vágásra, és ez alapján választjuk ki a „legjobbat”. Ehhez természetesen először le kell futtatnunk a hunmorphot a korpuszon, és az eredményt meg kell adnunk a morfessornak. Ezek után egy szó újravágásánál már csak a meglévő variációkat próbáljuk ki, és ezek közül választjuk a legjobbat. Ennek érdekében a programot tulajdonképpen csak a pszeudokódban szürkével megjelölt helyeken kellett módosítani, valamint a rekurziónál nyilván kell tartani, hogy éppen melyik hunmorph variációt vizsgáljuk. Tehát az első szürke résznél annyit változtatunk, hogy nem az összes lehetséges vágást vizsgáljuk, hanem csak azokat a vágásokat, amelyek megfelelnek az adott hunmorph elemzésnek. A második szürke soron a módosítás lényege, hogy nem engedjük meg azt, hogy ne vágjunk szét egy morfémát, ha a hunmorph még tovább bontaná azt. Sőt a vágást akkor is elvégezzük, ha ez növeli a kódhosszt..

Fontosnak tartjuk még jegyezni a hibrid megoldással kapcsolatban, hogy ez a morfessor használatát is módosítja. A morfessor működése ugyanis eredetileg két ciklusra bontható. Az első ciklusban egy nagyméretű korpusz alapján a program megtanulja, hogyan kell elemezni a szavakat, és elmenti az elemzéshez használt modellt. Ezek után a program működéséhez nincs szükség a korpuszra, hanem az elmentett modell alapján akár egyes szavakat külön-külön is képes elemezni. Ezzel szemben a hibrid modell csak tanuló üzemmódban használható, mivel a tanuló algoritmust használjuk fel a hunmorphos variációk szűrésére. Persze a morfessor továbbra is elmenti a modellt, és ez alapján működőképes lesz, de nincs garancia rá, hogy az ismeretlen szavakhoz olyan szegmentálást rendel, amit a hunmorph is felkínálna, illetve ilyen esetekben nem adható a szegmentálás mellé morfológiai elemzés. Ez a megkötés számunkra szerencsére nem akadály, mivel csak a beszédfelismerésben használt nyelvi modell építéséhez szeretnénk használni, ami egy offline folyamat, így taníthatjuk az egész korpuszon a morfessort.

4 Kísérleti eredmények

A felismerési teszteket a magyar MALACH korpuszon végeztük [3]. A MALACH (Multilingual Access to Large Spoken Archives) projekt célja, hogy hatékony hozzáférést biztosítson a Holokauszt túlélőinek beszámolóihoz. Az interjúkat 32 nyelven vették fel, és egy jelentős részük, mintegy 2000 órányi hanganyag magyar nyelvű. Ebből a 2000 órából eddig mindössze 31 órányit rögzítettek írásban is. A kísérleteket a lejegyzett részen végeztük. A tanításhoz 26 órányi, a teszteléshez a maradék 5 órányi anyagot használtuk fel (további részletek az akusztikus modell-tanításról a [3]-ban találhatók). A trigram nyelvi modellt a tanítókészlet 200 ezer szava alapján építettük a SRILM eszköz segítségével [10].
Röviden a következő eredményeket kaptuk. Ezek alapján látható, hogy az új megközelítés nem csak a nyelvtani információkat őrzi meg, hanem a felismerési pontosságban is a legjobb (3. táblázat).

3. Táblázat: Szó- és betűhiba arány (WER és LER) a MALACH korpuszon különböző szegmentálási módszereket használva. A feltüntetett hibák két különböző tesztalanyon mért hibák átlagai.

Nyelvi modell	**WER**	**LER**
Szó alapú	50,48	22,72
Szó alapú + beszélő adaptáció	45,1	18,42
Morfessor	47,58	21,26
Morfessor + beszélő adaptáció	39,77	16,11
Hunmorph + Morfessor	47,04	21
Hunmorph + Morfessor + beszélő adaptáció	**39,22**	**16,09**

Összefoglalás

A leírt kísérletek körüljárták a problémát, hogy hogyan lehet előállítani olyan szegmentálást, ami eredményesen használható a morféma alapú beszédfelismerés során. Kipróbáltunk két különböző elven működő, már meglévő módszert. A morféma alapú beszédfelismerési eredmények jobbnak bizonyultak a szó alapúnál, igazolva feltevésünket, hogy magyar nyelven a szó alapú beszédfelismerés nem optimális. Az első pozitív eredmények után megpróbáltuk finomítani a szegmentáló módszerünket. Ehhez a két különböző módszer kombinációjával próbálkoztunk. A megalkotott kombinációnk sikeresnek bizonyult, mert a hibrid módszer segítségével kaptuk a legjobb felismerési eredményeket, és a szabály alapú szegmentáló által előállított nyelvtani elemzés is a rendelkezésünkre áll a szegmentálás mellett.

A jövőben több érdekes folytatása is lehetséges a munkának. Egyrészt ha rendelkezésünkre állna egy egyértelműsítő, ami kiválasztja a korpuszban egy szónak az adott helyen legmegfelelőbb morfológiai elemzését, akkor megpróbálhatjuk a szabály alapú szegmentálásból mindig a legvalószínűbbet kiválasztani és ebből állítani elő a

nyelvi modellt. Egy másik kutatási irány annak vizsgálata, hogy a meglévő morfológiai elemzés segítségével hogyan javíthatóak a beszédfelismerési eredmények.

Köszönetnyilvánítás

Köszönet szeretnénk mondani a MOKK munkatársainak, különösen Halácsy Péternek a hunmorph eszközzel kapcsolatos technikai segítségért, valamint a beszédfelismerős – szövegfeldolgozós – számítógépes nyelvész kutatók eszmecseréjének aktív előmozdításáért, amelynek révén - reményeink szerint – nem csak e cikk társszerzői gazdagodtak.
A kutatást – részben – az NKFP-2/034/2004-es projekt keretében az NKTH támogatta.

Bibliográfia

1. Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Pylkkönen, J., and Virpioja, S., "Unlimited vocabulary speech recognition with morph language models applied to Finnish.", Computer, Speech and Language, Vol. 20, Issue 4 (2006) 515–541
2. Creutz, M. and Lagus, K., "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.", Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March, (2005)
3. Mihajlik, P., Fegyó, T., Nemeth B., Tüske Z., and Trón V., " Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages – Hungarian ASR for the MALACH Project TSD 2007, Pilsen
4. Kwon, O.–W. and Park, J., "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," Speech Communication, Vol. 39, Nos. 3–4 (2003) 287–300
5. Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Pylkkönen, J., and Virpioja, S., "Unlimited vocabulary speech recognition with morph language models applied to Finnish.", Computer, Speech and Language, Vol. 20, Issue 4 (2006) 515–541
6. Trón Viktor, Halácsy Péter, Rebrus Péter, Rung András, Simon Eszter, és Vajda Péter: morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-05), pp. 169–179, Szeged, 2005.
7. Szarvas, Máté: "Efficient large vocabulary continuous speech recognition using weighted finite-state transducers - The development of a Hungarian dictation system" PhD. Thesis, TITECH, Tokyo, 2003
8. Vicsi Klára at al.„Középszótáras, folyamatos beszédfelismerő rendszer fejelsztési tapasztalatai" III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, (p. 348-359), 2005
9. Péter Mihajlik, Tibor Fegyó, Zoltán Tüske, and Pavel Ircing, „A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian" In Proc. of INTERSPEECH-2007, pp.: 1497 – 1500, Antwerpen, Belgium, August 27-31, 2007
10. Stolcke, A., "SRILM – an extensible language modeling toolkit", In Proc. Intl. Conf. on Spoken Language Processing, Denver (2002) 901–904
11. Viktor Trón, László Németh, Péter Halácsy, András Kornai, György Gyepesi, and Dániel Varga, „Hunmorph: open source word analysis", In: Proceeding of ACL., 2005

Fonetikus morfológiai elemző beszédfelismeréshez

Gyepesi György[1], Kertész Zsuzsa[1, 2], Serény András[1],

[1] Alkalmazott Logikai Laboratórium,
1022 Budapest, Hankóczy J. u. 7.
{ggyepesi, kzsuzsa, sandris}@all.hu
[2] ELTE Angol Nyelvészeti Doktori Program,
1088 Budapest, Rákóczi út 5.

Kivonat: Ebben a tanulmányban azt mutatjuk be, hogy hogyan alakítottunk ki a magyar nyelvre **fonetikus morfológiai elemzőt**. Noha szegény morfológiájú nyelvek (mint amilyen az angol) esetében jó eredményt adnak a szóalak n-gramm nyelvmodellek, ragozó nyelvekhez olyan nyelvmodelleket érdemes kialakítani, amelyek a puszta szósorrend helyett a szavak szerkezetét, alaktanát is figyelembe veszik. Ezek hatékony működéséhez azonban az szükséges, hogy bemenetük ne csupán a szóalak legyen, hanem az ahhoz tartozó morfológiai elemzés is. Ennek egyik módja az, ha fonematizáljuk a betűalapú morfológiai elemzőnket. Ennek alkalmazásával elkerülhető számos, a kiejtésszótár haszálatakor felmerülő buktató is.

1 Bevezetés

A beszédtechnológiai kutatásokban nem új keletű az a felismerés, hogy hatékony beszédfelismeréshez nem lehet elegendő csupán az akusztikus jelsorozatok „megfejtése". A beszédfelismerő rendszer nem állhat pusztán az akusztikus komponensből, amely az akusztikus inputhoz egy fonémaszimbólum-sort (fonémajel-sorozatot) rendel, mert így az adott nyelvben lehetetlen fonémakapcsolatok, szavak, szósorozatok ugyanolyan a priori valószínűséggel jelennének meg, mint a lehetségesek. Ahogyan az emberi beszédfelismerés során is támaszkodunk szókincsbeli és grammatikai ismereteinkre, a gépi felismerésben is olyan rendszert érdemes kialakítani, amely a lehető legpontosabban modellálja ezt a folyamatot. Emiatt az akusztikus szintet mindig valamilyen **kiejtésszótárral** és a mondattani mintákért felelős úgynevezett **nyelvmodellel** kell támogatni.

A klasszikus felépítése egy ilyen beszédfelismerő rendszernek a következő: az akusztikus jelsorozatot az **akusztikus komponens** dolgozza fel, az általa kimenetként adott fonémajel-sorozatra az úgynevezett **kiejtésszótár** elemeit illesztjük, végül az így kapott szósorozatok közül a **nyelvmodell** segítségével választunk. A kiejtésszótár olyan szólista, amely a szavak írásképe mellett felsorolja azok lehetséges kiejtéseit.

Az angolban például a *read* alaknál egyaránt szerepel a R IY D[1] (a jelen idejű alak kiejtése) és a R EH D (múlt idejű alak kiejtése) kiejtés.

A kiejtésszótár bemenetét tehát az akusztikus felismerő által adott **fonémajel-sorozatok** adják (például R IY D), a kimenete pedig az adott hangsor „szótári alakja" (*read*). Azokat a fonémajel-sorozatokat, amelyekre nem illeszthető értelmes szótári szavak sorozata, a felismerő eldobja.

A teljesnek talált illesztéseket (azaz a kiejtésszótár által adott szósorozatokat) kapja meg a nyelvmodell, amely valószínűségük szerint értékeli őket: amelyik az adott nyelvben valószínűtlen, ritkán előforduló sorrend, az kis értéket kap, a gyakori szóegyüttállások értelemszerűen nagyot.

1.1 Mik a problémák a kiejtésszótárral?

A fent említett *read* ige példája máris felvet egy problémát a kiejtésszótár működési elvével kapcsolatban. Míg a kiejtett alaknál (az akusztikus komponens kimeneténél) még tudtuk, hogy R IY D vagy R EH D hangzott el, a kiejtésszótár a „neutralizált" *read* alakot fogja kiadni, mindenféle egyéb információ nélkül és ezt a szóalakot kapja meg a nyelvmodell. A nyelvmodell tehát nem használhat a szóalakokra vonatkozó egyéb információt.

Angolra jól működnek a szóalak-alapú (tehát hozzáadott alaktani információt nélkülöző) n-gramm modellek (lásd többek között [4]), ezért kisebb a jelentősége annak, hogy az adott helyen a *read* ige jelen vagy múlt idejű alakja szerepelt-e.

Ugyanakkor egy erősen ragozó nyelvre, mint amilyen pl. a magyar, a finn vagy a török a szóalak n-gramm modelleknél jobb eredményt adnak azok a modellek, amelyek morfológiai információt is tudnak kezelni (erről részletesebben lásd a 2.1. pontot). Egy ilyen nyelv esetében nagyon nem mindegy, hogy a kiejtésben tükröződő morfológiai vagy jelentésbeli különbséget sikerül-e a nyelvmodell számára elérhetően reprezentálni[2].

A kiejtésszótárak nem képesek kódolni a szóhatárokon fellépő hasonulásokat[3]. Vegyük példának a brit angolban az *r* sajátos disztribúcióját és az emiatt fellépő bonyodalmakat. Izoláltan az angolban szó végén nem fordulhat elő *r*; a szóvégi r-ek csak akkor jelennek meg, amikor egy magánhangzóval kezdődő szó vagy toldalék következik: *far away*.

A kiejtésszótár komoly akadályba ütközik, amikor a F AH R AX W EY fonémajel-sorozatot látja: ugyanis a két szó „között" megjelenő *r*-rel sehogy nem tud elszámol-

1 Az angol szavak fonetikus átírását a DARPA fonetikus ábécé szerint végeztük, de itt az egyszerűség kedvéért a hangsúlyjelöléseket elhagytuk; a magyar szavak átírására a helyesírást nagyjából tükröző saját rendszert alakítottunk ki.

2 A magyarban igen kevés olyan alak van, amelynél egy írott alakhoz több lehetséges kiejtés társul különböző jelentéssel (ezeket homográfoknak hívják a lexikológiában). Egy lehetséges példa az *egészség* szó két különböző ejtése: a gyakoribb, „betegség hiánya" jelentésben a *-ség* képző *s*-éhez hasonul az előző *-sz*; a ritkább (sokak szerint nem is létező), „valaminek az egész volta" jelentésű alakban ez nem történik meg.

3 Az egyszerűség kedvéért itt hasonulásnak nevezünk minden, morféma- vagy szóhatáron fellépő hangtani folyamatot (így a brit angolban megfigyelhető $r \sim \varnothing$ váltakozást is).

ni. Ha gondos kiejtésben hangzott el a szókapcsolat, akkor a F AH R AX W EY megoldásnak kell a legnagyobb akusztikus valószínűséggel bírnia; emellett persze megjelennek kisebb akusztikus valószínűséggel olyan (téves) megoldások is, mint például F AH AX W EY vagy F AH W EY stb. Ebből az a hiba származhat, hogy a kiejtésszótárban megjelenő (illetve nem megjelenő) alakok miatt épp a legnagyobb akusztikus valószínűségű – és valóban elhangzott – megoldást kell eldobni, és egy vagy több kisebb valószínűségűt megtartani és közvetíteni a nyelvmodell felé. Innen a nyelvmodellnek már lehetetlen korrigálnia a téves hipotézist.

Az ilyen problémákat a kiejtésszótáron belül csak úgy lehet orvosolni, ha minden szónak felvesszük az összes olyan alakját, amely egy szóhatáron fellépő folyamat miatt létrejöhet. Ez a megoldás viszont legalizálja azokat a fonémasorokat is, ahol egy szónak olyan kiejtése jelenik meg, amit a következő nem indokol (például a zöngésségi hasonulás miatt felvett alakokat elfogadja olyan esetben is amikor nincs hasonulás, olyanokat eredményezve, mint A B L A G A L A T[4], *ablak alatt*). Emellett ez az út a kevésbé ragozó nyelvek esetében is nehézkes, és nagy bővítéssel járna együtt.

2 Fonematizált morfológia

2.1 Miért kell morfológiai elemzés?

Morfológiailag gazdag nyelvek esetében is meg lehet valósítani – bár feltételezhetően kevésbé hatékonyan – a felismerést anélkül, hogy morfológiai elemzést vennénk igénybe. Nézzük, hogyan nézne ki a fenti rendszer a magyarra.

Először is ki kell alakítani a kiejtésszótárat, amely ez esetben egy szóalaktár. Ez azt jelenti, hogy elvileg az összes lehetséges szóalakot és azok kiejtéseit tartalmazza. Ez első ránézésre reménytelennek tűnik: egy magyar névszó lehetséges alakjainak száma száznál is több lehet, az igék esetében ez a szám pedig még több.

A probléma azonban csak látszólagos: a korpuszok vizsgálatával ugyanis azt lehet kideríteni, hogy az előforduló szóalakok száma mindössze kb. 2-3-szorosa a szótári alakoknak [6]. Ez azt jelenti, hogy minden szónak 2-3 (de mondjuk maximum 5-6) alakját kellene felsorolnunk, és ez már nem tűnik olyan lehetetlennek. Ehhez persze a korpuszhoz kellene igazítani a szótárt, ez pedig nem kis munka, de kivitelezhető.

Mivel a magyarra viszonylag egyértelműek és világosak a kiejtési szabályok, a szóalaktár fonematizálása – azaz a „kiejtésszóalaktár" – létrehozása sem okoz nagyobb gondot. Tehát a kiejtésszótár létrehozása megoldható: okozhat ugyan némi bizonytalanságot a felismerésnél – ha például olyan alakkal találkozik, amely véletlenül épp nincs benne a korpuszhoz igazított szótárban – de ez nem okoz jelentős romlást az eredményekben.

Egy másik alternatíva a Mathias Creutz által [2] bemutatott *Morfessor* nevű alkalmazás elvén működő elemzés: a szótárban nem csak szavak, hanem toldalékok is szerepelnek. Ha ezeket fonematizáljuk, kész a kiejtésszótárunk a ragozó nyelvre is.

[4] Technikai okok miatt a szegmentumok hosszúságát nem kezeljük.

Ezzel ugyan elfogadhatóvá válnak értelmetlen szavak is, de mivel az akusztikus bemenetben ritkán fordulnak elő nem létező alakok, elvileg itt is bízhatunk abban, hogy a felismerést ez nem rontja számottevően.[5]

Azonban hiába működnének ezek a megoldások, mindegyiknél megmarad a szóhatáron előforduló hasonulások problémája. Ennek megoldására csak az a lehetőség marad, hogy a különböző alakokat felsoroljuk a szótárban (lásd például F AH mellett F AH R). Ennek hiányában torzulhat az akusztikus felismerés, ahogyan már az angol példán bemutattuk.

A legfőbb motivációnk azonban arra, hogy a morfológiai elemzést ne kerüljük meg, az a **nyelvmodell** jellege. Mint azt már fentebb említettük, a magyarra és a ragozó nyelvekre az n-gramm modellek rosszabb eredményt nyújtanak, mint egy morfológiailag egyszerűbb nyelvre. Ennek több, a nyelvi tipológiából fakadó oka van. Egyrészt a gazdag morfológia általában – bár nem feltétlenül – együtt jár a szórendi kötöttségek elmaradásával. Egy angol mondat szórendje szigorúbb, mint egy magyar mondaté, és még ha a magyar szórendi variációk többnyire fontos grammatikai információt hordoznak is (lásd például topik, fókusz, stb.), a nyelvmodell szempontjából ez kevésbé lényeges. Csak az látható számára, hogy ugyanarra a szóhalmazra a szavak számától függően akár 4-6 legális szórend is előfordulhat (lásd például *Peti tavaly abbahagyta a hegedülést*; *Peti hagyta abba a tavaly a hegedülést*; *Peti tavaly hagyta abba a hegedülést; Peti tavaly a hegedülést hagyta abba*).

A szórendi kötöttségekhez az angolban az is hozzátartozik, hogy a jelentéstanilag és/vagy szintaktikailag összetartozó szavak – vagy szószerkezetek, frázisok – nem, vagy csak bizonyos mértékben szakíthatók el egymás mellől: az igét többnyire közvetlenül követik a legfontosabb bővítményei (tárgy vagy egyéb kötelező vonzatok), a különböző határozók (idő-, mód-, hely-, stb.) nem férkőzhetnek be ezek közé. A magyarban ez utóbbi „beférkőzésnek” gyakorlatilag nincs akadálya: emiatt az összetartozó szószerkezetek olykor igen messze, több szó távolságra is kerülhetnek egymástól, amit egy szóalak n-gramm modell nem tud kezelni. Noha a magyarban is akadnak ilyen „szétszakíthatatlan” szerkezetek, ezek jóval ritkábbak, mint az angolban.

Másrészt a ragozás során előálló viszonylag sokféle szóalak miatt adatritkasággal találkozunk, ami ismét ront a nyelvmodell tanulási hatékonyságán. Azt gondoljuk, hogy ezen okok miatt olyan nyelvmodellel kell dolgoznunk, amely nem hagyja figyelmen kívül a szavak belső szerkezetét. Egy ilyen nyelvmodell (legtipikusabb morfológiai nyelvmodellek az úgynevezett faktorizált nyelvmodellek lásd pl. [1]) például képes morfológiai szabályszerűségeket tanulni: például a *fut+nak* szóalak statisztikáiból következtetni tud az esetleg ritkábban előforduló *szalad+nak* alakra is. Az viszont világos, hogy ez egy szóalak alapú nyelvmodellnél nem történhet meg, ugyanis amíg a szavak belső szerkezete láthatatlan a nyelvmodellnek, addig az ilyen analógiákat nem lehet vele felfedeztetni[6].

---

5 Természetesen ha nem izolált szavakról, hanem szósorozatokról van szó, akkor igenis okozhat problémát az, hogy nem létező szavakat is elfogad az elemző, ugyanis előfordulhat, hogy tévesen azonosítja a morfémákat és ezáltal a szavakat is.

6 Meg kell azonban jegyezni, hogy az említett „távoli” grammatikai összefüggéseket a faktorizált nyelvmodellek sem képesek jól kezelni. Ez a probléma még mindig nem megoldott a nyelvmodellek kialakításában.

2.2 Miért legyen fonematizált a morfológiai elemző?

A fentiek alapján láthatjuk, hogy olyan eszköz megvalósítása lenne gyümölcsöző, amely egyrészt morfológiai annotációval tud bemenetet nyújtani a nyelvmodellnek, másrészt pedig a szó- és morfémahatáron történő hasonulásokat is jól tudja kezelni. Egy olyan morfológiai elemzőt képzelünk tehát el, amelynek a bemenete betűsor helyett fonémajelsor, a kimenete pedig a szokásos módon az adott sztring morfológiai elemzése.

Tehát, adva van egy fonémajelekből álló input (pl. B AA NY A); a morfológiai elemzőnk (ispell-alapú, lásd később) erre kiadja a következő elemzéseket

1. bánya/NOUN
2. bán/VERB<DEF>
3. bán/VERB<SUBJUNC-IMP><DEF>
4. bán/NOUN<POSS>

Ezután már a nyelvmodellen múlik, hogy melyiknek tud nagyobb valószínűséget adni.

Mik az előnyei egy ilyen fonetikus (fonematizált) morfológiai elemzőnek? Egyrészt mivel a magyar fonematizálási szabályok és a morféma- és szóhatáron lejátszódó hangtani folyamatok egyaránt kódolva vannak benne, nincs szükség arra, hogy felvegyük azokat a szóalakokat, amelyeket más esetben csak a szóhatáron történő hasonulások miatt kellene felvenni a szótárba (lásd az angol *far away* példát). Másrészt mivel a morfológiai elemző (ha jól működik), elő tudja állítani az összes lehetséges szóalakot – ráadásul ideális esetben csakis a létezőket állítja elő –, fel sem merül a szóalakok felsorolásának igénye. Harmadszor pedig fontos megemlíteni, hogy ezáltal meg lehet spórolni az egész kiejtésszótárat, eggyel kevesebb modulunk lesz az építményben, ami miatt az egész rendszerünk egyszerűbb, elegánsabb, és vélhetően gyorsabb is lehet.

3 Megvalósítás

3.1 A transzducer felépítése

Ahhoz, hogy a fonetikus morfológiai elemzőt meg tudjuk valósítani, szükség van (i) egy fonematizáló algoritmusra, és (ii) egy speciális szöveges morfológiai elemző transzducerre. A fonematizáló feladata megadni egy tetszőleges szöveg fonetikus átiratát. A morfológiai elemző transzducer abban az értelemben speciális, hogy az élein nem karakterek, hanem szótövek és toldalékok (tehát morfémák) szerepelnek.

Első lépésben azt az algoritmust ismertetjük, amely a fonematizáló és a speciális transzducer használatával kialakítja a fonetikus morfológiai transzducert. Végül

pedig azt mutatjuk be, hogyan építettünk fonematizálót és speciális morfológiai transzducert a magyar nyelvre.

Inicializáljuk a morfológiai elemző transzducert úgy, hogy vesszük a speciális morfológiai transzducert, és minden élen kicseréljük az inputot (a szótövet vagy toldalékot) annak fonetikus átiratával. Ezzel együtt megőrizzük az eredeti, karakteres (tehát „helyesírás szerinti") inputokat is. A fonetikus átiratokat a fonematizáló algoritmus adja.

Ezek után a transzducer *minden* egymást követő élpárját sorra vesszük. Legyen E és F két egymásutáni él, továbbá *ei* az E él inputja, *eo* az E él outputja; *fi* az F él inputja, *fo* pedig az F él outputja. Az output jelen esetben morfológiai annotációt, elemzést jelent. Ezenkívül legyen *epi* az E él inputjának, *fpi* pedig az F él inputjának fonematizált alakja.

Az *ei* és *fi* konkatenációjaként létrehozott *ei^fi* karaktersorozatot fonematizáljuk, az eredményt nevezzük *p*-nek.

Amennyiben a *p* fonémajelsorozat megegyezik *epi^fpi*-vel (tehát az *ei* és *fi* fonetikus alakjainak egymás után fűzésével), akkor nem történik semmi, mert ez azt jelenti, hogy az E és F éleken megjelenő inputok olyanok, hogy konkatenációjukkor a fonematizálásban nem történik hasonulás (pl. *ház+ban* = H AA Z B A N).

Ha azonban *p* eltér *epi^fpi*-től, akkor a fonematizált transzducerhez hozzáveszünk **egy új élet** E kezdőpontjából F végpontjába. Az új él inputja *p*, outputja pedig *eo^fo*: *ablak+ban* = A B L A G B A N, lásd az 1. ábrát.

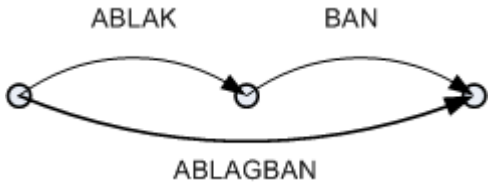


Fig. 1. A morféma- illetve szóhatáron történő hasonulásokat úgy reprezentáljuk, hogy az első él kezdőpontjából a második él végpontjába behúzunk egy harmadik élet, amelyen a két morféma konkatenációjának fonematizált alakja szerepel.

Mit érünk el ezzel a módszerrel? Azt, hogy az új transzducer a hasonulást tükröző *p* fonémajelsorozatot elfogadja, és azt az outputot generálja hozzá, mint az eredeti, hasonulást nem reprezentáló változathoz.

Az élpárok mentén való hasonítás időnként kihagy olyan hasonulásokat, amelyek három vagy több él mentén jelennek meg. A magyarban igen ritkák az olyan hasonulások, amelyekhez élhármasokat kell figyelembe venni: ilyen például a *hajt+s+d*, ahol egyrészt a *t+s* összeolvadását (*cs*), valamint a *d* által kiváltott visszafelé történő zöngésségi hasonulást is kezelni kell valahogy (*hajdzsd*). Az algoritmusnak paramété-

re az, hogy élhányasokkal dolgozzon, tehát nagyobb számra is beállítható; a magyarra azonban az esetek legnagyobb részére elégségesnek találtuk az élpárokat.

Az algoritmus eredménye tehát olyan fonetikus morfológiai transzducer, amelynek élein fonémajelsorozatok vannak.

2.2 A transzducer a felismerésben

Van tehát egy transzducerünk, amelynek élein egy véges ábécéből (a fonémajelekből) alkotott sorozatok szerepelnek inputként. Egy transzducer esetében azonban definíció szerint egy véges ábécé jelei az inputok, minden élen egy jellel: a bemutatott transzducer látszólag nem felel meg ennek a definíciónak. Ugyanakkor véges sok éle van, a véges sok élen pedig véges sok input, tekintsük tehát ábécének azt a halmazt, amelynek elemei a transzducer élein szereplő inputok.

Hogyan működik ez a transzducer mint felismerő? Legyen adott egy véges fonémajelsorozat ***P***. A transzducerünk input ábécéjében fonémajelsorozatok vannak, ***P*** viszont *fonémajelek* és nem *fonémajelsorozatok* sorozata. Nem tudunk mást tenni, mint hogy ***P***–t feldaraboljuk minden lehetséges módon, és minden darabolást felismertetünk a transzducerünkkel, összegyűjtjük a kapott kimeneteket, és ezek halmazát adjuk vissza.

Így is megvalósítható a felismerés, de ez a módszer exponenciálisan lassabb, mintha olyan transzducerünk lenne, aminek minden élén egy jel (azaz egy fonémajel) az input.

A transzducerünket tehát **betűsíteni** kell: ha adott egy transzducer, amelynek élein egy *véges ábécé jeleiből álló sorozatok* az inputjelek, akkor a betűsítő algoritmussal készítjük el azt az eredetivel ekvivalens transzducert, amelynek élein az *ábécé jelei* az inputok. A betűsítő algoritmus szétvágja az egynél hosszabb éleket annyi élre, amennyi jelből az eredeti él inputja áll, az utolsó élre írja az eredeti él kimenetét, és természetesen elkészíti a szétvágáshoz szükséges új csúcsokat. Az első él indulóállapota az eredeti él indulóállapota, az utolsó él végállapota az eredeti él végállapota, a közbenső csúcsok az újak, és az újak egyike sem elfogadó.

A speciális morfológiai elemző transzduszert a [3]-ban ismertetett módon készetjük el ispell, illetve magyarra a **morphdb** [8] erőforrásból. Megjegyezzük, hogy a speciális transzducerből fent ismertetett betűsítő algoritmussal készítünk karakter-inputú morfológiai transzducert.

Fonematizáló transzducert a magyar nyelv fonetikus szabályaiból az [5] szerint építettük.

Bibliográfia

1. Bilmes, J. A., Kirchhoff, K.: Factored language models and generalized paralel backoff. In Proceedings of HLT/NAACL, (2003) 4–6.
2. Creutz, M.: Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition**.** Doctoral Dissertation (2006)

3. Halácsy P. et al. Végesállapotú transzducerek mindenkinek. V. Magyar Számítógépes Nyelvészeti Konferencia, Szegeed, 2007. december 6–7.
4. Jurafsky, M. Martin, J. H.: Speech and Language Processing: an Introduction to Natural Language Processing, Speech Recognition, and Computational Phonology. (2000) Prantice-Hall.
5. Kaplan, R., Kay, M.: Regular models of phonological rule systems. Computational Linguistics, 20:3 (1994) 331-379.
6. Kornai, A.: Frequency in morphology. In Kenesei I. (ed): Approaches to Hungarian IV. (1992) 246-268.
7. Koskenniemi, K.: Two-level morphology: A general computational model of word-form recognition and production. Tech. rep. Publication No. 11, Department of General Linguistics, University of Helsinki. (1983).
8. Trón V. et al.: morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In Alexin Zoltán, Csendes Dóra (szerk.). *III. Magyar Számítógépes Nyelvészeti Konferencia.*, SZTE, Szeged, 2005, p. 169-179.

Fonémaosztályok felügyelet nélküli tanulása
Absztrakt

Gyarmati Ágnes, Vásárhelyi Dániel

{MTA Nyelvtudományi Intézet, ELTE BTK Elméleti nyelvészet doktori program}
1068 Budapest, Benczúr u. 33.
{aagnes, vad}@nytud.hu

Kivonat Írásunkban különféle természetes fonémaosztályok különféle felügyelet nélküli tanulóalgoritmusok általi tanulását mutatjuk be különböző korpuszokon. Ezek az algoritmusok kizárólag az egyes fonémák korpuszon belüli eloszlása alapján, mindenféle fonológiai vagy bármilyen más előzetes ismeret nélkül alkalmasak bizonyos természetes osztályok elkülönítésére.

1. Bevezetés

A hagyományos fonológiai elemzés alapja a megkülönböztető erővel bíró egységekre, a *fonémákra* való szegmentálás. Az egyes fonémákat többnyire négy alapkritérium: a szembenállás, a kiegészítő eloszlás, a fonetikai hasonlóság és a szabad váltakozás elve alapján szokás azonosítani. Erre a műveletre már a strukturalista irányzatokban, a számítógép megjelenése előtt létezett algoritmikus módszer [1]. A generatív fonológia elterjedésével a szamítógépes módszerek háttérbe szorultak, és csak az utóbbi időben kerültek előtérbe.

A fonológiai általánosítások, szabályok vagy megszorítások megfogalmazásában, reprezentációjában alapvető szerepet játszanak a természetes osztályok. Azokban az esetekben, amikor egy természetes osztály egy másik természetes osztályban vett komplementere (azaz egy őt tartalmazó másik természetes osztályból való kivonás eredménye) szintén természetes osztály, értelmezhetjük a két komplementer osztályt, mint a nagy természetes osztály *partícióját* vagy másképpen (kétértékű) *fonológiai jegyet*. Ilyen fonémaosztályok – a teljes igénye nélkül – a magán- és mássalhangzók, az előbbin belül a magyarban az elöl- és hátulképzettek, az utóbbiban a zöngések és zöngétlenek, és így tovább.

Írásunkban három különböző tanulási módszert ismertetünk, melyek közül a legrégebbi Szuhotyin algoritmusa [7], amely az alapján sorolja egy osztályba az elemeket, hogy bizonyos tulajdonságaik mennyire térnek el egymástól. A 2-means klaszterezés egy meghatározott metrika szerinti közelség alapján osztja fel két osztályra a fonémák halmazát [5]. Harmadik módszerünk az összes partícionálás közül választja ki azt, amelyre egy célfüggvény értéke a lehető legnagyobb lesz.

Az algoritmusokat különféle nyelvekből vett korpuszokra teszteltük: magyar, angol, francia, japán, cseh, maori és hawaii. Az algoritmusok bemeneti adatként fonémasorozatokat várnak, ezért a nyelvtől és az adott nyelv helyesírási konvencióitól függően a szokványos írott szövegek általában nem használhatók. Megoldás lehet speciális korpuszok, fonémikus reprezentációban írt korpusz használata, illetve a szöveges korpusz a mi céljainknak megfelelő előfeldolgozása. Az általunk használt angol, japán és francia korpusz fonémikusan átírt szólisták [3], a cseh, maori és hawaii helyesírás jól közelíti a fonémikus reprezentációt. Az írott magyar nyelvű korpusz azonban gondos előfeldolgozást igényel, ezt automatikus eszközökkel oldottuk meg (pl. az azonos értékű fonémák/fonémasorozatok egy közös szimbólummal/szimbólumsorozattal való helyettesítése ($ly \rightarrow j$, $qu \rightarrow kv$). Az eredmény ugyan nem adott tökéletes fonémikus reprezentációt, de a mi céljainknak már megfelelt.

Szuhotyin kifejezetten a mgh/msh-elkülönítés automatizálására tervezte algoritmusát, két alapfeltevésre építve: i, egy szöveg leggyakoribb szegmentuma mindig magánhangzó, ii, a magánhangzók és a mássalhangzók gyakrabban váltakoznak, mint nem. Az algoritmus kezdeti lépésként minden szegmentumot mássalhangzónak címkéz fel, majd a korpuszban található bigramok gyakoriságából kiindulva iteratív lépésekben keresi meg azokat a szegmentumokat, melyek a legnagyobb valószínűséggel magánhangzók, minden lépésben egyet, amíg talál megfelelő jelöltet. Goldsmith és Xanthos szerint az eredmény függ a nyelvtől, továbbá az algoritmus más feladatra nem alkalmazható[4]. Ezzel szemben mi azt találtuk, hogy az eredmény elsősorban az átírás minőségétől függ, nem függ a fonémakészlet felépítésétől, ami nem is meglepő, hiszen a fonémák disztribúcióját veszi alapul. Továbbá az algoritmus más problémák esetében is adhat értelmes eredményt, amennyiben a vizsgálat tárgyában váltakozó tendencia rejlik (pl. egyes nyelvekben a hangsúly). Ekkor természetesen az alapfeltevéseket a megfelelő módon kell módosítani. Az i, tulajdonképpen, csak az egyik osztály szerepének kitüntetésére szolgál, nevezhetnénk az első elem címkéjét *1*-esnek is.

A klaszterezés az egymáshoz való hasonlóság alapján osztályozza az elemeket, így nincs szükség semmilyen előzetes tudásra vagy feltevésre a kategóriákat vagy az osztályozandó elemeket illetően.

Mi a *k-means* algoritmust használtuk, amely egy vektortér pontjait k partícióba sorolja be oly módon, hogy az egyes partíció középpontjaitól való távolságok négyzetösszegét minimalizálja. A vizsgált nyelv fonémakészletének számosságából adódó n dimenziós vektortérben a vektorokat a relatív bigramgyakoriságokból képezzük, esetünkben $k = 2$, mivle két kategóriába kívánjuk sorolni a fonémákat. A klaszterezést a Cluster 3.0 szoftver [2] segítségével végeztük.

Mivel a két klaszter közül egyiknek sem volt kitüntetett szerepe, így fordulhatott elő, hogy a tesztelés közben a 0 jelű clusterbe hol a magánhangzók, hol a mássalhangzók kerültek. A magyar, cseh, francia, maori és hawaii nyelvre tökéletes osztályzást kaptunk ezzel a módszerrel, a japán és az angol fonémák osztályzása kisebb hibát tartalmazott. Az angol esetében viszont elmondhatjuk,

hogy gyakorlatilag ugyanazt az osztályozást kaptuk a hangsúlyt jelölő, illetve a hangsúlyt nem jelölő reprezentációra is. Az egyetlen különbség a klaszterek címkéjében volt.

A k-means clusterezés előnye, hogy szélesebb körben alkalmazható kategorizálásra, az egymással alternáló komplementer osztályok mellett képes olyan természetes osztályok megtanulására is, mint a magyar mássalhangzók elölségi osztálya, amelyekre váltakozás helyett az osztály elemeinek harmóniája a jellemző.

Ugyanezeknek az osztályoknak a tanulására még transzparensebben alkalmazható az a módszer, amelyben minden partícióhoz hozzárendeltünk egy számot, amely értéke nőtt abban az esetben is, ha egy *alternáló*bb[1] és abban az esetben is, ha egy *harmonizáló*bb[2] partíciót választottunk. A legalternálóbb vagy legharmonizálóbb partíció által meghatározott osztályokat tanulta meg az algoritmus.

2. Korpuszok

A cikkben ismertetett módszereket több, különböző fonémarendszerel rendelkező nyelvekre teszteltük. Mivel az algoritmusok bemenete *fonémasorozatok*, a korpuszokat ennek a követelménynek megfelelően kellett kiválasztani, illetve alakítani. Ebből a szempontból a korpuszok háromfélék lehetnek.

- fonémikusan adott
- a fonémikus reprezentációt jól közelítő ortografikus
- a fonémikus reprezentációtól jelentősen eltérő ortografikus

Egy átlagos korpusz az utóbbi két kategória egyikébe tartozik, bár léteznek speciális korpuszok is, melyeket korábban már átírtak fonémikussá, és úgy tették hozzáférhetővé[3].

A tesztelés során használt angol, francia és japán korpuszok a John Goldsmith honlapján szabadon hozzáférhető szólisták. Alapjában véve megőriztük az ott alkalmazott átírási rendszert, azonban az angol korpuszt két változatát is figyelembe vettük, az eredeti, a hangsúlyt is jelölő ArpaBet ábécében, illetve a hangsúlyjelölést elhagyva is.

[1] Azaz a partíció által meghatározott komplementer természetes osztályok elemei egymással többet váltakoztak.

[2] Azaz a partíció által meghatározott komplementer természetes osztályok elemei gyakrabban fordultak elő együtt.

[3] Természetesen fonetikusan átírt szövegeket is használhatunk bemeneti adatként, amennyiben a fonetikai átírás nem sokban különbözik a fonémikustól. Ha ugyanis az átírásban alkalmazott fonetikai rendszer túl finom, akkor elveszhetnek a hangrendszerrel kapcsolatos fontos információk, melyek megragadására a fonológia absztraktabb szintje kínál lehetőséget (pl. a fonémaazonosság kérdése). Erre a későbbiekben még visszatérünk

A maori és a cseh helyesírás gyakorlatilag fonémikus, néhány kisebb átalakítási művelettől eltekintve (pl. a szöveg egységes kisbetűsítése) nem változtattunk rajtuk. A cseh és a maori korpuszt a világhálón szabadon hozzáférhető szövegekből gyűjtve állítottuk össze.

1. táblázat. Extracts from the corpora used

Hungarian
2003-10-11 14:00 Golgota Közösségi Ház.
Bejárat a Hold utca felől.
English (w stress)
BECAME 247 B IH0 K EY1 M
BECAUSE 884 B IH0 K Aó Z
BECKETT 11 B EH1 K IH0 T
BECKONED 8 B EH1 K AH0 N D
BECOME 360 B IH0 K AH1 M
English (w/o stress)
BECAME 247 B IH K EY M
BECAUSE 884 B IH K AO Z
BECKETT 11 B EH K IH T
BECKONED 8 B EH K AX N D
BECOME 360 B IH K AH M
French
abrasif 21 A b r A z i f
abreuver 1 A b r ö v é
abri 39 A b r i
abri-sous-roche 1 A b r i s u r O S
Japanese
bonresuhamu 1 b o n r e s u h a m u
bonryo 1 b o n r y o
bonryuu 1 b o n r y u :
bonsai 1 b o n s a i
Czech
Za další hodinu se vrátil.
Pod paží třímal objemný balík.
Maori
I taku haerenga mai ki te Tari Toko i te
Ora i te marama o Hōngongoi 1993

A legalaposabb előfeldolgozásra a magyar szövegek esetén volt szükség:

– Először eltávolítottuk a szövegből a nem a magyar ábécé betűit jelölő karaktereket, majd kisbetűsítettünk, hogy elkerüljük a kis- és nagybetűk megkülönböztetéséből fakadó fonématöbbszöröződéseket.

– A magyarban, mint tudjuk, néhány fonémát kétjegyű, egy esetben (*dzs*) pedig háromjegyű betűvel jelölünk, ezek a korpusz előfeldolgozása során különös figyelmet igényelnek. A többjegyű betűkkel jelölt hangból képzett geminátákat az első karakter megkettőzésével írjuk le:
  - short: *cs, (dz), dzs, gy, ly, ny, sz, ty, zs*
  - long: *ccs,(ddz),ddzs,ggy,lly,nny,ssz,tty,zzs*

  Ezeket a geminátákat a feldolgozás során felbontottuk, pl. $ssz \rightarrow sz + sz$.
– További átalakítások:
  - $ly = j$
  - $q + u = k + v$
  - $w = v$
  - $x = k + sz$
  - $y = i$
  - $ch = h$

Két különböző magyar korpuszt használtunk az egyik a magyar webkorpusz [6], a másik Jókai Mór *Az arany ember* című regénye volt.

3. Algoritmusok és alkalmazásaik

3.1. Szuhotyin algoritmusa

Szuhotyin algoritmusa [7] egy régi egyszerű algoritmus, mely a magánhangzók és mássalhangzók elkülönítésére szolgál. Egy szöveget beolvasva az abban szereplő szegmentumokat automatikusan két diszjunkt halmazba sorolja be, az algoritmus kimenete ez a két halmaz, a mássalhangzókat, illetve a magánhangzókat tartalmazó halmaz. Ez a felügyelet nélküli gépi tanulási módszer két alapvető feltevésre épül:

– egy átírt szöveg leggyakoribb szegmentuma magánhangzó
– a magánhangzók és mássalhangzók gyakrabban váltakoznak, mint nem

Ebben a fejezetben röviden ismertetjük az algoritmust, majd bemutatjuk tesztelési eredményeinket.

Leírás. Tekintsünk egy nyelvet, mely hangrendszere n fonémából áll: $P = \{p_1, \ldots, p_n\}$, és egy korpuszt az adott nyelvből. A program ezek ismeretében legelőször egy $(n \times n)$ dimenziós négyzetes R mátrixot hoz létre és tölt felt, melynek elemeire: $r_{ij} = r_{ji} = f_{ij} + f_{ji}$, ha $i \neq j$, ahol f_{kl} a $p_k p_l$ bigram előfordulási gyakorisága a korpuszban. Az elemekre vonatkozó szabályt általánosnak véve a főátlóban az egyforma fonémákból álló bigramok gyakoriságának kétszeres értékeinek kellene megjelennie, de az algoritmus működése szempontjából a főátlóbeli értékek definíció szerint 0-k. R=

$$\begin{pmatrix} 0 & f_{12}+f_{21} & \dots & f_{1n}+f_{n1} \\ f_{21}+f_{12} & 0 & \dots & f_{2n}+f_{n2} \\ \vdots & & \ddots & \vdots \\ f_{n1}+f_{1n} & \dots & & 0 \end{pmatrix} \tag{1}$$

A fonémák fel vannak címkézve, kezdetben minden fonémá címkéje *mássalhangzó*. Az algoritmus ezután egy iteratív fázisba lép, a ciklusmag minden egyes lefutásakor az R mátrix adataiból kiindulva megkeresi és átcímkézi azt a fonémát, mely a legnagyobb valószínűséggel valójában nem is mássalhangzó, hanem magánhangzó. A további lépésekben az „új magánhangzónak"az R mátrixbeli összes adatát figyelmen kívül kell majd hagyni.

A második alapfeltevés alapján az a fonéma lehet jelölt a magánhangzóságra, mely jóval gyakrabban előz meg vagy követ mássalhangzókat, mint magánhangzókat, így mindenképp pozitívnak kell lennie annak azon gyakoriságok különbségének, melyek megmutatják, hogy a korpuszban hányszor állt az adott fonéma mássalhangzó, illetve magánhangzó közvetlen környezetében. Ez a különbség iteratívan hozzá van rendelve minden egyes p_i fonémához a megfelelő $v(p_i)$ értékben:

$$v(p_i) = \sum_{1 \le j \le n, j \neq k_v} r_{ij} - \sum_{1 \le j \le n, j = k_v} r_{ij}$$

ahol k_v azon fonémák indexei, melyek már korábban új, magánhangzós címkét kaptak. Minél magasabb ez a $v(p_i)$ érték, annál valószínűbb, hogy p_i magánhangzó. Az algoritmus mohó, az átcímkézésre mindig a legmagasabb $v(p_i)$ értékkel rendelkező fonémát választja ki (illetve közülük az egyiket, ha több ilyen is van). Amikor a $v(p_i)$ értékek között nincs pozitív, az algoritmus leáll, és visszaadja a felcímkézett fonémák listáját.

Alkalmazások.

3.2. Eredmények

Goldsmith és Xanthos [4], miután Szuhotyin algoritmusát három különböző nyelvre alkalmazták (angol, francia, finn), megállapitják, hogy habár az algoritmus maga nyelvfüggetlen, pontossága erősen függ a bemenő adatoktól. Az algoritmust mi magyar, angol, cseh, maori nyelveken teszteltük.

Az eredmény a magyarra tökéletes volt, azaz minden általunk jól ismert magánhangzó magánhangzó lett, és csak a valódi mássalhangzók maradtak a mássalhangzók között. A jó eredmények arra ösztönöztek minket, hogy megvizsgáljuk Goldsmith és Xanthos azon kijelentését, mely szerint az adatok erősen befolyásolják az algoritmus pontosságát. Angol nyelvű korpuszra is alkalmaztuk az algoritmust, Goldsmithékéhez hasonló eredménnyel.

2. táblázat. Szuhotyin algoritmusa a Magyar webcorpusra

Cluster	Fonémák
Magánhangzók	a,á,e,é,i,í,o,ó,ö,ő,u,ú,ü,ű
Mássalhangzók	b,c,cs,d,dz,dzs,f,g,gy,h,j,k,l,m,n,ny,p,r,s,sz,t,ty,v,z,zs

3. táblázat. Szuhotyin algoritmusa angolra (hangsúlyjelöléssel)

Cluster	Fonémák
Mgh	AH0,R,S,IH0,L,ER0,N,M,EY1,WHH,Y,ER1,AÉ,AY0,EY0,OW0,DH
Msh	AA0,AÁ,AE0,AH1,AO0,AÓ,AW0,AW1,AY1,B,CH,D EH0,EH1,F,G,IH1,IY0,IY1,JH,K,NG,OW1,OY0 OY1,P,SH,T,TH,UH0,UH1,UW0,UW1,V,Z,ZH

Amint azt Goldsmith és Xanthos is megjegyzik, a fonetikai átírás félrevezeti az algoritmust azáltal, hogy olyan szimbólumok is megkülönböztetendők, melyek ugyanazt a fonémikus tartalmat jelölik, de eltérő a hangsúlycímkéjük (0 vagy 1 (hangsúlytalan, illetve hangsúlyos)). A hangsúlyjelölések eltávolítása az „igazi" fonémák kisebb halmazához vezet, egyúttal jelentősen javítja az algoritmus által elérhető eredményeket is. Az egyetlen hibája, hogy az R-t tévesen magánhangzónak osztályozza.

Az algoritmus ugyanazokat a lépéseket végzi, és a korpusz is ugyanaz. Az eredményt tehát nem a korpusz változtatta meg, hanem a szimbólumok halmaza, azaz az ábécé. Mivel az algoritmus célja eredetileg is az volt, hogy tanulja meg a mássalhangzók és magánhangzók kategóriáját, ha adottak a nyelv *fonémái*, az ArpaBet hangsúlyjelölőit egyszerűen figyelmen kívül kell hagyni, hiszen a hangsúly az angolban nem fonémikus.

4. táblázat. Szuhotyin algoritmusa angolra (hangsúlyjelölés nélkül)

Cluster	Fonémák
Mgh	AA,AE,AH,AO,AW,AX,AY,EH,ER,EY,IH,IY,OW,OY,R,UH,UW
Msh	B,CH,D,DH,F,G,HH,JH,K,L,M,N,NG P, S, SH, T, TH, V, W, Y, Z, ZH

Az *r* szokatlan viselkedésére (angolban és franciában is az egyetlen mássalhangzó, mely átkerült a magánhangzók közé) többféle magyarázat lehetséges. Az egyik szerint az *r* a különböző nyelvekben különböző mértékben konszonantális (ld. a fonetikai megvalósítas változatosságát). Ez további vizsgálatokat igényel. Meg kell jegyezni azonban, hogy a cikkben használt korpusz a sztenderd ameri-

kai normát követi, rotikus dialektus átirata. Az *r* gyakran szótagképző, valamit mássalhangzó előtt is megmarad (nemrotikus dialektusokban a prekonszonantális *r*-ek nincsenek[4].

Az algoritmust cseh nyelvű korpuszon is teszteltük. A cseh nyelv egyik jellegzetessége, hogy az *l* és az *r* lehetnek szótagképzők. Várhatnánk, hogy az *l*, de

5. táblázat. Szuhotyin algoritmusa korpuszra

Cluster	Fonémák
Magánhangzók	ý,ó,a,é,y,o,ě,á,e,í,u,ú,i,u
Mássalhangzók	ň,k,v,w,č,ž,l,x,m,b,c,n,z d,ch,p,f,ř,g,r,š,ď,s,h,t,j

főleg az *r* átkerüljön a magánhangzók közé, de ez mégsem történik meg. A szillabikus és a nem-szillabikus *r* fonémikusan azonosak, tehát az algoritmus nem tesz különbséget közöttük. Az algoritmus az osztályozást a fonémák disztribúciójára alapozva végzi, így az a tény a döntő, hogy mely helyzetben, milyen környezetben a legnagyobb a gyakorisága. Az *r* gyakrabban áll tiszta mássalhangzós pozícióban, szótagmagban, így nem kerül a magánhangzók közé.

A maori fonémarendszer az eddig tárgyalt nyelvekétől eltér, mindössze kilenc magánhangzóból és tíz mássalhangzóból áll, így méretében és arányaiban is más. Nyelvfüggetlen tanuló algoritmusunk tökéletes eredményt ad. Szuhotyin

6. táblázat. Szuhotyin algoritmusa maori korpuszra

Cluster	Fonémák
Magánhangzók	a,ē,ō,o,e,ī,ā,i,ū,u
Mássalhangzók	k,w,m,n,ng,p,wh,r,h,t

ezt az algoritmusát kifejezetten a magánhangzók és mássalhangzók automatikus megkülönböztetésére tervezte. Goldsmith és Xanthos ezt egyik hátrányának is tartják, az algoritmus eredeti célja annyira speciális, hogy más feladatra nem alkalmazható. Vizsgáljuk meg azonban a két kezdeti alapfeltevést.

– egy átírt szöveg leggyakoribb szegmentuma magánhangzó
– a magánhangzók és mássalhangzók gyakrabban váltakoznak, mint nem

[4] Ez elméletfüggő, hogy a prekonszonantális, illetve szünet előtti helyzetben egyáltalán nincs-e *r*, vagy csak a felszínen nem jelenik meg

Az első tulajdonképpen azt a célt szolgálja, hogy az algoritmus ki tudja választani a megfelelő címkét a két halmazból. Ez azonban csak egy kitüntetett szerep megjelölése, de az algoritmus lényegi működését nem érinti, fel lehet címkézni a *magánhangzók* helyett 1-essel is.

A második feltétel pedig csak annyit vár el, hogy a szegmentumok valamilyen szabály(osság) szerint váltakozzanak (jelen esetben a tendencia a magánhangzók és a mássalhangzók váltakozása). Ha eltekintünk a konkrét szerepektől, és csak a szabályossagra fordítjuk figyelmünket, felfedezhetjük, hogy bármely két osztály elemeinek besorolására használható, amennyiben a két osztály elemei váltakozást mutatnak.

Ilyen váltakozást mutat pl. az angolban a hangsúlykiosztás: nem állhat két hangsúlyos szótag egymás mellett. Ebben az esetben az ArpaBet hangsúlyjelölése már nem irreleváns információ. Valóban, ha korpusznak az eredeti korpusz magánhangzóból álló sorozatot tekintjük, akkor az algoritmus eredményeként az ArpaBet 0-s, illetve 1-es hangsúlyjelű elemeit szetválogatva. Mivel a schwa a legtöbbször előforduló magánhangzó a szövegben, így az 1-es kategóriába a hangsúlytalanok kerülnek. Érdekesebb kérdés, hogy mit ad az algoritmus, ha a

7. táblázat. Szuhotyin algoritmusa az angol magánhangzókra

Cluster	Fonémák
Mgh 1-es kategória (hangsúlytalan)	AA0,AE0,AH0,AO0,AW0 AY0,EH0,ER0,EY0,IH0 IY0,OW0,OY0,UH0,UW0
Mgh 2-es kategória (hangsúlyos)	AÁ,AÉ,AH1,AÓ,AW1 AY1,EH1,ER1,EY1,IH1, IY1,OW1,OY1,UH1,UW1

hangsúlyt nem jelölő ábécét használjuk. Azt már tudjuk, hogy a schwa a leggyakoribb szegmentum, így most is ez fog legelőször sorra kerülni, és az 1-es kategóriát megnyitni. Hipotézisünk az volt, hogy a schwához azok a magánhangzók fognak csatlakozni az 1-es kategóriába, melyek *tipikusan hangsúlytalan* szótagban fordulnak elő, míg a másik osztályba a *tipikusan hangsúlyos* magánhangzók kerülnek. A hipotézishez hasonló eredményt kaptunk.

8. táblázat. Szuhotyin algoritmusa az angol magánhangzókra (2)

Cluster	Fonémák
Mgh 1-es kategória	AX,ER,IH
Mgh 2-es kategória	AA,AE,AH,AO,AW,AY,EH,EY,IY,OW,OY,UH,UW

3.3. 2-means klaszterezés

A klaszterezési algoritmusok valamilyen hasonlóság alapján osztanak részekre egy objektumhalmazt. A számítógépes nyelvészetben elsősorban valamiféle jegyek szerinti csoportosítás felügyelet nélküli megtanulására használják őket, mivel nem szükséges hozzájuk semmiféle előzetes feltevés a csoportokról. Számos klaszterezési módszer ismeretes, ezek közül mi a k-means algoritmust használtuk.

Leírás. A k-means algoritmus a klaszereket a tömegközéppontjuk segítségével definiálja. Egy valós V vektortér elemeit klaszterezzük. Kezdetben véletlenszerűen választunk ki k darab pontot középpontnak ($\mu_j \in V$), majd i, minden pontot besorolunk abba a klaszterbe, amelynek középpontjához legközelebb esik, ii, újraszámítjuk a tömegközéppontok helyét

$$\mu_j = \frac{1}{|C_j|)} \sum_{x \in C_j} x$$

i és ii iterálását addig folytatjuk, míg a középpontok helye már nem változik. Mivel az algoritmus nem feltétlenül a megoldáshoz konvergál, ezért többszöri futtatásra van szükség.

Távolságfüggvénynek használható az euklideszi távolság vagy valamilyen másik távolságfügvény (korreláció, city block). Mi az euklideszi távolságot

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

használtuk, de nem okozott jelentős különbséget más metrika használata sem.

Alkalmazások. Az algoritmus Szuhotyinéhez hasonlóan jól teljesített a magán- és mássalhangzók szétválasztásakor és kiváló eredményt adott a magyar korpuszokon a magánhangzók elölségi osztályozásakor annak ellenére, hogy a két jegy viselkedése gyökeresen különbözik, az előbbi értékei alternálni, az utóbbiak harmonizálni szeretnek.

Magán és mássalhangzók. A különféle korpuszokon többé-kevésbé hibátlanul a magán- és mássalhangzókat kaptuk meg a teljes fonémakészleten való futtatáskor.

Az angol NG és a japán n hibás osztályozására magyarázatot adhat azok speciális eloszlása, az, hogy sohasem fordulnak elő szótagkezdetben és általában magánhangzó előtt.

Elölségi harmónia. A 2-means algoritmust a magyar korpuszok magánhangzóin futtatva a két elölségi osztályt (a,á,o,ó,u,ú ill. e,é,i,í,ö,ő,ü,ű) kaptuk meg, némi bizonytalansággal az i-k besorolásában, ami jól illeszkedik a magyar fonológiai ismereteinkhez.

Corpus	Cluster	Phonemes
Czech	0	ý,ó,a,é,y,o,ě,á,e,í,ů,ú,i,u
	1	ň,k,v,w,č,ž,l,x,m,b,c,n,z,d,c,p,f,ř,g,r,š,ď,s,h,t,j
Hungarian	0	ó,a,é,ő,ö,o,á,e,í,ú,ű,i,ü,u
	1	ny,v,k,l,b,m,zs,c,n,d,z,ty,p,cs,q,f,dzs,sz,r,g,gy,h,s,t,j
English w accents	0	lY0,AE0,lY1,AY0,AO0,EH0,AE1,UW0,AO1,NG,AY1,ER0,EH1,AW0,ER1,UW1,Z,AA0,AW1,OY0,EY0,AA1,IH0,UH0,IH1,OY1,EY1,UH1,OW0,OW1,AH0,AH1
	1	HH,V,K,W,L,B,M,Y,ZH,N,DH,D,CH,P,F,R,G,JH,S,TH,T,SH
English w/o accents	0	HH,V,K,W,L,B,M,Y,ZH,N,DH,D,CH,P,F,R,G,JH,S,TH,T,SH
	1	AW,AA,UW,AY,ER,AO,NG,EH,AE,Z,lY,AH,UH,OW,IH,EY,O
French	0	@,è,A,é,y,õ,Ö,O,o,ö,E,ã,^,i,Ü,u
	1	k,v,!,w,l,b,m,n,z,d,p,f,g,r,s,S,t,J,*,j
Maori	0	k,w,m,n,ng,p,wh,r,h,t
	1	a,ē,ō,o,e,ī,ā,i,ū,u
Japanese	0	a,n,o,:,e,i,u
	1	v,k,w,x,m,b,C,y,z,d,p,f,',G,g,r,h,S,sT,t,J,j

1. ábra. 2-means clustering results for each corpora

Egyéb osztályok. Az algoritmus az angol korpuszon a mássalhangzókra futtatva többé-kevésbé a zöngés-zöngétlen distinkciót adta (HH, K , Y , Z , CH, P , F , JH, S , TH, T , SH és V , L , B , M , ZH, NG, N , DH, D , G , R), ami egy tipikusan harmonizáló jegy, míg a zöngéseken belül az egymással alternáló szonoráns-obstruens osztályok jöttek ki.

3.4. Függvénymaximalizálás

A számítástudományban megszokott módszer, hogy egy feladat optimális megoldását egy célfüggvény maximalizálásával keressük. A következőkben egy olyan egy adott korpusz fonémáinak partícióin értelmezett függvényt mutatunk, amely maximumá egy olyan partíción veszi fel, amelynek elemei maximálisan harmonizálnak vagy alternálnak a korpuszban.

Leírás. Tegyük fel, hogy adott egy korpusz bigram gyakorisági $n \times n$ mátrixa M. Ekkor a fonémák egy p $0,1$-értékű n hosszú vektorral adott partíciójára összegezhetjük azokat a bigram gyakoriságokat, amelyek egy osztály elemein belüli bigramokhoz tartoznak (f_h) illetve azokat, amelyek két különböző osztály elemei közötti bigramokéi ($f_a = C - f_h$, ahol C egy konstans, az összes bigram

gyakoriság öszege, ezzel normálva $f'_a = 1 - f'_h$). A két szám különbsége így annak négyzete $f = (f'_a - f'_h)^2 = (2f'_a - 1)^2$ akkor maximális, ha *valamelyik* gyakorisági összeg maximális.

Az $f(p)$ például az alábbi módon számolható ki:

- Legyen e egy n hosszú 1-vektor $((1, 1, ..., 1))$
- Legyen $M_p = \begin{pmatrix} tr((e-p)^T M(e-p)) & tr((e-p)^T Mp) \\ tr(p^T M(e-p)) & tr(p^T Mp) \end{pmatrix}$
- Végül legyen $f(p) = \left(1 - 2tr\left(\frac{M_p}{\|M_p\|_{tc}}\right)\right)^2$ az optimalizálandó függvény, ahol a $\|\|_{tc}$ a mátrix taxi normája, azaz $\sum_{i,j} m_{ij}$

Keressük tehát a adott M-re $ArgMax\, f(p)$-t. Ezt egy mohó rekurzióval végezzük, egy véletlen partícióból indulunk ki és egy-egy elemet osztályozunk át, valahányszor az $f(p)$ értékét ez növeli.

A mohó rekurzió

- Egy véletlen p választása, $f(p)$ kiszámítása
- Amennyiben létezik p', amely egyetlen elem osztályának megváltoztatásával áll elő és $f(p') > f(p)$, akkor a rekurzió meghívása p'-re

Alkalmazások. Ez az algoritmus a klaszterezéssel szinte teljesen megegyező eredményeket adott a teljes fonémakészleten (magán- és mássalhangzók), a magyar magánhangzókon (elölségi osztályok) és az angol mássalhangzókon (zöngésség és szonoritás).

4. Összefoglalás

A fentiekben bemutattunk három gépi tanulási módszert, amelyek segítségével különböző korpuszokon, különféle fonémaosztályok megtanítását végeztük el. Az eredmények azt mutatják, hogy – legalábbis bizonyos– természetes osztályok megtanulhatók mindenféle előzetes fonológiai, akusztikai vagy fonetikai feltételezések nélkül is.

4.1. További feladatok

Az ismertetett algoritmusok lehetőségeit még korántsem merítettük ki. Számos nyelvet és korpuszt kell még megvizsgálni ahhoz, hogy kiderüljön, a fenti módszerek valamelyike, vagy esetleg többük is alkalmas-e egy korpusszal adott nyelv (legalább részleges) fonológiai reprezentációjára.

Szükséges lenne továbbá az algoritmusok teljesítményének optimalizálására is.

Az ismertetett algoritmusokon kívül további feladat más módszerek keresése és alkalmazása is fonémaosztályozási feladatokban, mint például a rejtett Markov-modell (HMM) használata.

Hivatkozások

1. J. Durand, Siptár P.: Bevezetés a fonológiába. Osiris, Budapest (1997)
2. M. Eisen, M. de Hoon: Cluster 3.0 Manual for Windows, Mac OS X, Linux, Unix. Standford University (1999), University of Tokyo (2002)
3. J. Goldsmith: Phonological Complexity. Software and corpora at http://hum.uchicago.edu/%7Ejagoldsm/PhonologicalComplexity/
4. J. Goldsmith, A. Xanthos: Learning phonological categories Draft at http://hum.uchicago.edu/%7Ejagoldsm/Papers/phonolcat.pdf
5. A. K. Jain, R. C. Dubes: Algorithms for clustering data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1988)
6. Kornai, A, Halácsy, P, Nagy, V, Oravecz, Cs, Trón, V, and Varga, D (2006). Web-based frequency dictionaries for medium density languages. *In: Proceedings of the 2nd International Workshop on Web as Corpus, edited by Adam Kilgarriff, Marco Baroni ACL-06, pages 1–9*
7. Sukhotin, B. V.: Eksperimental'noje vydelenie klassov bukv s pomoščju evm. Problemy strukturnoj lingvistiki, 234:189–206 (1962)

IV. Ontológia

Igék szemantikai klaszterezése bővítménykereteik alapján

Gábor Kata[1], Héja Enikő[1]
[1] MTA Nyelvtudományi Intézet, Korpusznyelvészeti osztály, Postafiók 701/518, H-1399 Budapest, Magyarország

{gkata,eheja}@nytud.hu

Kivonat: A bemutatott kísérlet célja a magyar igék bővítménykereteik alapján történő szemantikai csoportosítása nem felügyelt tanulási módszerrel. A bővítménykereteket a Szeged Treebankből nyertük ki. Az igéket hierarchikus klaszterezési eljárással csoportosítottuk. A cikkben bemutatjuk az eljárást, az eredményeket, valamint kitérünk a szemantikai osztályok kiértékelésének nehézségeire, és javaslatot teszünk egy új értékelési módszerre.

1 Bevezetés

Az utóbbi évtizedben a nyelvtechnológiai kutatás fontos célkitűzésévé vált a nagy lefedettségű, robusztus, több nyelvi szintet lefedő lexikonok létrehozása. Ennek egyik oka, hogy a többszintű struktúra megkönnyíti a karbantartást és lehetővé teszi az automatikus bővítést, mivel a strukturált szerkezetnek köszönhetően a műveletek egyedi tételek helyett szóosztályokra alkalmazhatók, vagyis lehetővé teszik a nyelvészeti általánosításokat. Másfelől pedig a lexikális információ automatikus kinyerése kevésbé idő- és munkaigényes, mint az adatbázisok kézi felépítése. A korai automatikus lexikonépítési kísérletekben nem számítógépes célokra készült, kétnyelvű vagy értelmező szótárak elektronikus változatát használták nyersanyagként. Ez a megközelítés rendelkezik azzal az előnnyel, hogy a nyersanyagából a zajt már emberi munkaerővel kiszűrték, ám a módszerben rejlő lehetőség épp a nyersanyag korlátolt mennyisége miatt viszonylag hamar kimerült, a későbbi automatikus frissítésre pedig nem ad lehetőséget. A szótár használatánál robusztusabb megközelítést jelent az igei vonzatkeret-információ automatikus kinyerése nagyméretű korpuszokból. Napjainkban már a legtöbb európai nyelvre, így a magyarra is rendelkezésre állnak olyan erőforrások (morfológiailag elemzett és egyértelműsített korpusz: [15], szintaktikailag annotált treebank: [4]), melyek lehetővé teszik gépi tanulási módszerek alkalmazását a vonzatkeret-információ és a szemantikai tulajdonságok kinyerése céljából. A lexikai tuljdonságok gépi tanulásával foglalkozó kutatások többsége az igei szubkategorizációs keret korpuszból való kinyerését tűzte ki célul ([12], [3], magyarra: [Sass, 2007]). Ugyanakkor a lexikai szemantika-szintaxis interfész elméleti kutatásával párhuzamosan történtek

kisérletek az igék automatikus szemantikai csoportosítására is kvantifikálható szintaktikai jellemzőik alapján. ([5], [13], [14], [11]). Ezen kutatások közös elméleti kiindulópontját a szintaxis-szemantika interfésszel foglalkozó elméletek közös feltevése képezi, mely szerint az igei argumentumok vonzatokként való megvalósulásáról az ige jelentéséből kiindulva adhatunk számot. Ez a széles körben felhasznált elméleti előfeltevés (*Semantic Base Hypothesis,* [8], [10]) azon a megfigyelésen alapul, hogy a szemantikailag hasonló igék hasonló szintaktikai környezetben fordulnak elő, azaz hasonló vonzatkeret-mintázatokat mutatnak. Az igék automatikus szemantikai csoportosítását célzó munkák vagy a [10]-ben meghatározott igeosztályokat kívánják igazolni korpusz-adatok segítségével, vagy olyan algoritmusok fejlesztésén dolgoznak, amelyek lehetővé teszik új igék szemantikai kategorizációját. A fentiekkel szemben jelen kutatás célja azoknak a lexikai-szemantikai tulajdonságokkal meghatározható igeosztályoknak az azonosítása, amelyek relevánsak a magyar igei argumentumrealizáció leírása szempontjából. Mivel nem tudjuk előre, milyen szemantikai osztályokba szeretnénk besorolni az igéket, nem felügyelt tanulási módszerhez kell folyamodnunk. Így [13] nem felügyelt módszerét követtük. A Szeged Treebank ([4]) 150 leggyakoribb igéjét soroltuk csoportokba hierarchikus agglomeratív klaszterezési eljárással, szintaktikai bővítményekereteik alapján. A kísérlet kettős célt szolgált: egyrészt magát a tanulási módszert akartuk tesztelni, vagyis arra kerestük a választ, hogy bővítménykeret-információ alapján kinyerhetők-e szemantikailag koherens osztályok a korpuszból. Amennyiben ige, úgy a kísérlet megerősíti a *Semantic Base Hypothesis*-t, hiszen alátámasztja, hogy a szemantikailag hasonló igék szintaktikailag is hasonló viselkedést mutatnak. Másrészt azt akartuk megtudni, hogy melyek azok a szemantikai jelentéskomponensek, melyek köré az alapvető igeosztályok szerveződnek. Előfeltevésünk szerint a leggyakoribb igékből előűlló csoportok tükrözni fogják a legalapvetőbb igei jelentéskomponenseket.

A következő fejezetekben bemutatjuk a felhasznált jegykészletet [2] és a klaszterezési eljárást [3], majd ismertetjük az eredményeket [4]. A kiértékelés nehézségeire az [5] részben térünk ki. Végül felvázoljuk a kutatás további lehetséges irányait [6], majd összegezük az elmondottakat [7].

2 A jegykészlet

Mivel a jelenleg rendelkezésre álló magyar szintaktikai elemzők (Babarczy et al. 2005, Gábor és Héja 2005) nem végeznek teljes elemzést, nem terjednek ki az igei vonzatkeret automatikus felismerésére. Ezért úgy döntöttünk, hogy az első kísérlethez eleve szintaktikailag elemzett korpuszt, a Szeged Treebank-et használjuk. A korpuszt alkotó szövegek különböző forrásból származnak: üzleti hírek, napi sajtó, szépirodalom, jogi szövegek és iskolások fogalmazásai alkotják. A kísérlethez valamennyi részkorpuszt felhasználtuk. A treebank mérete 1.2 millió szó. Az igei bővítménykeret annotálációjában nem szerepel a vonzat és a szabad határoz kategóriája, az ige és bővítménye közti relációt a bővítmény esetragja definiálja.

A klasszifikációs (felügyelt tanulási) és klaszterezési (nem felügyelt) eljárások alkalmazásakor a siker szempontjából alapvető fontosságú kérdés, hogy milyen, a korpuszból kinyer-

hető, kvantifikálható tulajdonságok tükrözik leginkább azokat a lexikális tulajdonságokat, melyek köré a nyelvészetileg releváns osztályok szerveződnek. [2] reguláris mintákat használ. [13] és [14], [3] szintaktikailag elemzett korpuszból kinyert szubkategorizációs kereteket használnak. [13] kísérletképpen a szemantikai szelekciós információval is kiegészítette a kereteket, ám az összehasonlításból kiderült, hogy a pusztán szintaktikailag jellemzett keretek használata pontosabb eredményre vezet[1]. [7] az igeosztályokra jellemző szintaktikai alternációkat nyelvész szakértők által definiált jegykészletekkel közelítik. Módszerük előnye, hogy csak POS-taggelést igényel, ám hátránya, hogy az igeosztályok halmazát előre definiálni kell, az osztályokra jellemző jegykészlet kiválasztásához szakértői munkára van szükség, és időigényes. Ennek a hátránynak a kiküszöbölésére [11] létrehozott egy általános jegykészletet, mely tetszőleges osztályba tartozó angol igék felügyelt csoportosítására használható.

A magyar igék automatikus csoportosításakor nem állt rendelkezésünkre nyelvészek által meghatározott besorolás, így nem tudtuk előre, milyen szemantikai osztályokat szeretnénk eredményül kapni. Ezért mindenképpen nem felügyelt módszerhez kellett folyamodnunk. Igeosztályok híján nem ismerhettük az osztályokra jellemző alternációk halmazát sem, ami kizárta a magyar alternációkra jellemző közelítő jegyek definiálását. Így [13]-hoz hasonlóan az igéket szubkategorizációs kereteikkel jellemztük.

A Treebank annotációjának megfelelően a vonzatok mellett a szabad határozókat is a szubkategorizációs keret részének tekintettük. Ennek a döntésnek nem csupán gyakorlati oka van. A magyar nyelvre eddig nem született megbízhatóan alkalmazható vonzatteszt, aminek az az oka, hogy a vonzatok és a szabad határozók szintaktikai viselkedése hasonló. A felszíni sorrend alapján nem különböztethetjük meg őket, hiszen az ige és a vonzata közé beékelődhet egy vagy több szabad határozó is (szemben például az angollal). A szintaktikai funkciót jelölő esetragok különböző igék mellett különböző szerepeket kódolnak. Emmellett azt gondoljuk, hogy az ige melltt megjelenő adjunktumok legalább annyira jellemzőek az ige jelentésére, mint a vonzatok. Az adjunktumok sem produktívak abban az értelemben, hogy tetszőleges ige mellett használhatnánk őket: olyan igék mellett tehetők ki, melyek jelentése kompatibilis az adjunktum esetragja által kódolt szemantikai szereppel. Ezen megfontolásokból úgy döntöttünk, hogy az adjunktumokat is figyelmbe vesszük az igék szintaktikai környezetének jellemzésénel.

A treebankből kinyert szubkategorizációs kereteket azok az esetragos NP-k és főnévi igenevek alkotják, melyek az ige alá tartozó csomópontokban helyezkednek el. Az annotáció jellegéből fakadóan e gyermek-csomópontok kinyerése egy nem rendezett listát eredményez, melynek elemei az ige szintaktikai bővítményei. A szubkategorizációs keretek hosszának, vagyis a keretben szereplő bővítmények számának nem szabtunk felső határt. A kerettípusokat úgy állítottuk elő az egyedi keretekből, hogy elhagytuk a bővítmények lemmáját, morfoszintaktikai elemzésükből pedig csak a szófaj taget és az esetragot tartottuk meg. A bővítmények sorrendjét nem vettük figyelembe. Az általánosítás eredényeként 839 féle keretet kaptunk, melyeket mind megtartottuk.

[1] Az angol igék csoportosításának teszteléséhez Schulte im Walde Levin igeosztályait használta.

3 A klaszterezési eljárás

Mivel a kísérlet eredményeként a magyar igékre jellemző szemantikai igeosztályokra voltunk kíváncsiak, a klaszterezési eljárást a 150 leggyakoribb magyar igére alkalmaztuk. Az igék reprezentálásánál és a klaszterezési módszer kiválasztásánál [13]-ban leírt módszerre támaszkodtunk.Az igék korpuszbeli előfordulásait az ige és a különböző szubkategorizációs keretek együttes előfordulásainak maximum likelihood becslése adja:

$$p(t|v) = f(v,t) / f(v)$$

ahol *f(v)* az ige gyakorisága, *f(v,t)* pedig az ige és a keret együttes gyakorisága. Az értékeket a 150 igére és mind a 839 szubkategorizációs keretre kiszámoltuk.

A valószínűségi eloszlások összehasonlításához különbözőségi mértékként a relatív entrópiát használtuk:

$$D(x||y) = \sum_{i=1}^{n} x_i \cdot log\ (x_i / y_i)$$

A szubkategorizációs keretek nagy száma miatt az igék valószínűségi eloszlásában sokszor szerepel nulla érték. A relatív entrópia választása azzal jár, hogy ezeket az értékeket simítással korrigálni kell. Másfelől viszont nem akartuk elveszíteni azt az információt, amit a nulla számú előfordulás hordoz – vagyis az ige és a keretben megjelenő esetrag(ok) szemantikai inkompatibilitását. Mivel a leggyakoribb igékkel dolgoztunk, feltételeztük, hogy ezek a hiányok nem véletlenszerűek, és olyan simítási módszert akartunk választani, mely megfelel annak az elképzelésnek, hogy az együttes előfordulás hiánya egy gyakori ige esetében nagyobb eséllyel jelez inkompatibilitást, mint egy kevésbé gyakori lemma esetében (ahol inkább beszélhetünk véletlenről). Ezért az alábbi módszerrel simítottuk az adatokat:

$$f_e = 0{,}001 / f(v) \quad ha \quad f_c(t,v)=0$$

ahol f_e a becsült, f_c pedig a korpuszban mért gyakoriság.

Ezután két hierarchikus agglomeratív klaszterezési eljárást alkalmaztunk az adatokra:

1) Először kiindulásként a 150 ige mindegyikét egyelemű klaszternek tekintettük. Az interáció minden lépésénél kiszámoltuk az összes klaszter közti távolságot, azaz a relatív entrópiákjukat, és minden lépésben összevontuk a két leghasonlóbb klasztert. A klaszterek közötti távolságot minden lépésben újra kiszámoltuk. Ahogy Schulte im Walde is megjegyzi, ennek a módszernek az

a hátulütője, hogy néhány iteráció után az igék a legnépesebb klaszterek köré csoportosulnak, így kevés, nagy elemszámú csoportot eredményez. Hogy elkerüljük a problémát, meghatároztuk klaszterek maximális elemszámát: a negyedik elem után több igét nem olvasztottunk a klaszterbe. Az összevonást addig folytattuk, míg az intuitív értékelés alapján hasznosnak találtuk. Ez a módszer, valamint a korpusz mérete a 150 igéből 120 besorolását tette lehetővé, az összevonások folytatása ezután csökkentette volna a csoportok belső koherenciáját. Mindazonáltal így is szembesültünk a lánceffektussal (*chaining effect*), ami azt jelenti, hogy az eredményül kapott csoportok némelyikének legkevésbé hasonló tagjai közötti távolság túl nagy volt.

2) A második kísérletben a lánceffektus elkerülése céljából az elemszám helyett a klaszterek legkevésbé hasonló elemei közötti távolság maximális értékét határoztuk meg. Csak akkor soroltunk be egy új igét egy klaszterbe, ha a klaszter elemei közül a tőle legtávolabb eső igétől mért távolsági értéke kisebb volt, mint a – tesztfutások alapján meghatározott – maximális érték. Ezzel a módszerrel 71 igét sikerült 23 osztályba sorolnunk. Az első módszerrel szemben a második előnye, hogy képes nagy elemszámú, mégis koherens csoportok kialakítására, ami esetünkben különösen fontos, mivel a legalapvetőbb magyar szemantikai igeosztályokra vagyunk kíváncsiak. Ugyanakkor éppen ebből az okból későbbi terveink között szerepel nem agglomeratív, top-down módszerek kipróbálása is, melyek alkalmasabbak az adatok szerkezetének áttekintésére.

4 Eredmények

Mindkét fent ismertetett módszernél azt tapasztaltuk, hogy az igék egy része kis számú, népes csoportokba szerveződik, míg a maradékuk általában egy közeli szinonímájával (pl.: zár - végez) vagy ellentétpárjával (pl.: ül - áll) alkot egy klasztert. Természetesen az 1) módszer, vagyis a csoporton belüli igék számának korlátozása kevesebb klasztert ereményez és értékesebb eredményeket ad a kevésbé gyakori igék esetében. Ezzel szemben a második módszer, vagyis az egy csoportba sorolt igepárok közti maximális távolság korlátozása hatékonyabb az alapvető, nagy elemszámú igeosztályok meghatározására. Mivel az a célunk, hogy Levinéhez hasonló igeosztályokat találjunk a magyar nyelvben, a következő lépés annak megvizsgálása, hogy az igeosztályok koherensek-e szemantikailag, és ha igen, elemeik milyen jelentéskomponensekben osztoznak. Elsőként a három legnagyobb osztályt vizsgáltuk meg, mert a leggyakoribb igék közül sokat tartalmaznak, mégis – a módszer jellegzetességéből adódóan –belső koherencia jellemzi őket. A 71 kategorizált ige egyharmada a három

legnagyobb osztály valamelyikébe lett besorolva. Az alábbiakban ismertetjük az osztályokat[2].

C-1: létigék: *marad, van, lesz, nincs*

C-2: modálisok: *megpróbál, próbál, szokik, szeret, akar, elkezd, fog, kíván, kell*

C-3: mozgásigék: *indul, jön, elindul, megy, kimegy, elmegy*

Míg a C-1 és a C-3 osztályok erős szemantikai koherenciát mutatnak, a C-2 osztály leginkább a szintaktikai *modális* leírással jellemezhető. C-2 egy alosztályára laklamazható az a leírás, hogy valamilyen cselekvés elvégzésével kapcsolatos mentális attitűdöt fejeznek ki (*szeret, akar, kíván*), de ay osztály többi tagja esetében nehéz közös szemantikai tulajdonságot találni.
Általánosságban elmondható az eredményül kapott igeosztályokról, hogy lehorgonyozhatók valamilyen szemantikai jelentéskomponenshez, vagy jól jellemezhetők valamely argumentumuk közös szemantikai szerepével. Például: állapotávltozást jelentő igék: *erősödik, gyengül, emelkedik*; beneficiens argumentummal rendelkező igék: *biztosít, ad, nyújt, készít*; képességet jelentő igék: *tud, lehet, sikerül*. Néhány csoportot a fentieknél specifikusabb metapredikátummal jellemezhetünk – például külön csoportot alkotnak a megjelenést vagy az ítélkezést jelentő igék. Más esetekben viszont a szemantikai reláció sokkal kevésbé szoros, mint például az „ágenses, folytonos cselekvést jelentő igék" címkével leírható csoport esetében: *ül, áll, lakik, dolgozik*. A skála másik végén elhelyezkedő klaszterek olyan igéket tartalmaznak, melyek szemlátomást nem osztoznak közös jelentéskomponensben, pusztán „véletlenül" ugyanolyan esetraggal járó vonzatuk miatt kerültek egy csoportba: *foglalkozik, találkozik, rendelkezik.*

5 Értékelés

A szemantikai igeosztályok kiértékelésére még nincs bevett módszer. A létező eljárások két csoportba sorolhatók. Az első csoportba tartozó módszerek a csoporton belüli koherenciát vizsgálják. Másképpen szólva azt ellenőrzik, hogy egy független távolsági mérték használatával is azt az eredményt kapjuk-e, hogy a csoporton belüli elemek közelebb vannak egymáshoz, mint más csoportok elemeihez. Ezzel az eljárással azonban nem sokat tudunk meg a csoportok szemantikai koherenciájáról. A másik megoldást valamilyen kézileg előállított csoportosítással való összevetés jelenti. Az angol kísérletekben Levin osztályozását használják, ami a magyarra nem áll rendelkezésünkre. Kézenfekvő megoldás lehet a magyar WordNettel ([9]) való összehasonlítás is, azonban ezzel szemben is kifogások merülnek fel. Az osztályokat akkor tudjuk

[2] Mivel nem voltak előzetes osztályaink, az elnevzéseket utólag adtuk, az intuitív értékelés megkönnyítése céljából.

rávetíteni a hálóra, ha az osztály igéi közötti szemantikai kapcsolatot sikerül leképezni a WordNet-hierarchiára. Ha azonban az igei jelentések közti kapcsolatot a háló csomópontjaiban mérve fogalmazzuk meg, problémát jelent, hogy a WordNet csomópontok közti távolság nem egyenletes.
Mivel végsősoron Levin-típusú osztályzást akarunk kidolgozni a magyarra, az osztályokat kiértékelhetjük úgy is, hogy megpróbálunk az osztály igéire jellemző szintaktikai alternációkat keresni. Ehhez azonban figyelembe kell vennünk a magyar szintaxis jellegzetességeit, melyek megnehezítik az alternációk leírását. A lehetséges szubkategorizációs keretek nagy száma és a vonzatok nagy részének elhagyhatósága miatt túl sok lehetséges alternációval kell számolnunk. Ezért úgy döntöttünk, hogy a kezdetekben megpróbáljuk leszűkíteni vizsgálódásunk tárgyát. Kiindulásként megpróbáltuk meghatározni az egy csoportba sorolt igék közös jelentéskomponenseit. Ezután megvizsgáljuk, milyen szemantikai szerepeket kódoló bővítmények megjelenését engedélyezik ezek a jelentéskomponensek. Ha az egy osztályba sorolt igék ugyanazokkal a szemantikai szerepekkel kompatibilisek, és megegyeznek abban is, hogy milyen esetrag kódolja a szerepet, akkor az osztályt koherensnek tekintjük. Precízebb megfogalmazásban ez azt jelenti, hogy az igeosztályokat mátrixokkal ábrázoljuk, melyek oszlopait a főnévi esetragok, sorait az igei lemmák töltik ki, a cellákban pedig az a szemantikai szerep szeerepel, melyet az adott esetraggal megjelenő bővítmény az ige mellett betölt. Az osztály akkor koherens, ha a hozzá tartozó mátrix megfelel az alábbi két követelménynek:

1) A mátrix specifikus az adott igeosztályra.
2) Az egy oszlopba tartozó cellák ugyanazt a szerepet tartalmazzák.

A következő táblázatok koherens és osztályspecifikus mátrixokat tartalmaznak.

1. Táblázat: A C-3 klaszter igéi és a hozzájuk tartozó szemantikai szerepek

	ACC	**INS**	**ABL**	**ELA**
indul	-	eszköz - társ	forrás	Forrás
jön	-	eszköz - társ	forrás	Forrás
elindul	-	eszköz - társ	forrás	Forrás
megy	-	eszköz - társ	forrás	Forrás
elmegy	-	eszköz - társ	forrás	Forrás
kimegy	-	eszköz - társ	forrás	Forrás

2. Táblázat: A C-1 klaszter igéi és a hozzájuk tartozó szemantikai szerepek

	ACC	**INS**	**ABL**	**ELA**
marad	-	társ	ok	Összetevő
van	-	társ	ok	Összetevő
lesz	-	társ	ok	Összetevő
nincs	-	társ	ok	Összetevő

Amint az 1. táblázat mutatja, a C-3 csoport igéi mellett ablatívusz vagy elatívusz rag is jelölheti a forrás szerepű összetevőt. Az esetragok közti választás a főnévi csoporton múlik. A C-1 csoport igéi mellett ezek az esetragok más szemantikai szerepeket kódolnak.
Fontos megjegyezni, hogy nem rendelkezünk a szemantikai szerepek egy előre meghatározott halmazával. A megbízható kiértékeléshez szükséges, hogy az oszlopok kitöltését egymástól függetlenül több ember végezze.

6 Konklúzió és további teendők

A 150 leggyakoribb magyar ige szemantikai osztályokba sorolása egy kezdeti lépés az igei szintaxist meghatározó szemantikai tulajdonságok feltérképezése felé. Mivel nem volt előfeltételezésünk az eredményként vár igeosztályokról, és nem voltak az értékeléshez használható, kézzel kialakított csoportjaink sem, az eredmények kiértékeléséhez nyelvészeti elemzésre van szükség. Mindazonáltal a 4 részben bemutatott intuitív értékelés alapján azt mondhatjuk, hogy a kapott osztályok meglepően erős szemantikai koherenciát mutatnak. Figyelembe kell vennünk továbbá, hogy ezek a biztató eredmények rendkívül kis korpusz használatával születtek, ami megerősíti, hogy a *Semantic Base Hypothesis* jó eredménnyel használható szemantikai osztályok automatikus kinyeréséhez.
A feladat és az eredmények tükrében a további teendők két részterületre oszthatók. Egyfelől tervezzük további klaszterezési eljárások (például a [13] –ban leírt top-down eljárás) kipróbálását, valamint a jegykészlet finomítását, egyrészt a zajos jegyek kiszűrésével, másrészt újabb, releváns morfoszintaktikai jegyek bevonásával. Hosszabb távon szükséges a vizsgálódást kiterjeszteni a Magyar Nemzeti Szövegtár adataira, ehhez azonban szükség lesz a korpusz szövegének legalább részleges automatikus szintaktikai elemzésére. Másik fontos feladatunk a jövőben az igeosztályok nyelvészeti elemzése, mely az eredmények kiértékelésével szorosan összefügg. A kiértékeléshez használt mátrixok előállítását az igékhez társított példamondatok segítségével képzeljük megvalósítani.

Bibliográfia

1. Babarczy Anna, Gábor Bálint, Hamp Gábor, Kárpáti András, Rung András, Szakadát István.: *Hunpars. Mondattani elemző alkalmazás.* In: *Alexin Z., Csendes D. (szerk): A Harmadik Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete,* Szeged Egyetemi nyomda, 2005. Szeged, pp. 20-28 .
2. Brent, M.: From Grammar to Lexicon: unsupervised learning of lexical syntax.. *Computational Linguistics*, 19(2): 243-262. MIT Press, 1993, Cambridge, MA, USA
3. Briscoe, T., Carrol, J.: Automatic Extraction of Subcategorization from Corpora. In: *Proceedings of the 5th Conference on Applied Natural Language Processing,* 1997, pp. 356-373, Washington DC, USA
4. Csendes Dóra, Csirik János, Gyimóthy Tibor, Kocsor András: The Szeged Treebank. *LNSC series Vol. 3658,* pp. 123-131

5. Dorr, Bonnie J., Jones, Doug: Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In: *Proceedings of the 14th International Conference on Computational Linguistics (COLING-96)*, 1996, pp. 322–327., Kopenhagen, Denmark.
6. Gábor, K., Héja, E.:.Vonzatok és szabad határozók szabályalapú kezelése. In: *Alexin Z., Csendes D. (szerk): A Harmadik Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete,* Szeged Egyetemi nyomda, 2005. Szeged, pp. 245-257 .
7. Joanis, E., Stevenson, S.: A general feature space for automatic verb classification. In: *Proceedings of the 10th Conference of the EACL,* 2003, pp. 163-170, 200. Budapest, Hungary.
8. Koenig, J-P, Mauner, G., Bienvenue. B..: Arguments for Adjuncts. *Cognition,* 2003*,* 89, pp.67-103.
9. Kuti J., Vajda P., Varadsi K.: Javaslat a magyar igei WordNet kialakítására. In: *Alexin Z., Csendes D. (szerk): A Harmadik Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete,* Szeged Egyetemi nyomda, 2005. Szeged, pp. 79-87.
10. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. Int. J. Digit. Libr. 1 (1997) 108–121
11. Merlo, P., Stevenson, S.: Automatic Verb Classification Based on Statistical Distributions of Argument Structure In: *Computational Linguistics. 27:3*, 2001, pp. 373-408.
12. Pereira, C. N. Fernando, Tishby, N., Lee, L.: Distributional Clustering of English Words. *31st Annual Meeting of the ACL*, 1993, Columbus, Ohio, USA, pp. 183-190.
13. Schulte im Walde, Sabine: Clustering Verbs Semantically According to their Alternation Behaviour. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, 2000, Saarbrücken, Németország, pp. 747–753.
14. Schulte im Walde, Sabine and Brew, Chris: Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, Philadelphia, PA, USA, pp. 223–230.
15. Váradi, T.: The Hungarian National Corpus. Proceedings of the Third International Conference on. Language Resources and Evaluation, 2002. Las Palmas pp.385-389

NP-koreferenciák feloldása magyar szövegekben a Magyar WordNet ontológia segítségével

Miháltz Márton[1], Naszódi Mátyás[1], Vajda Péter[2], Varasdi Károly[2]

[1] MorphoLogic Kft., 1126 Budapest, Orbánhegyi út 5,
{mihaltz, naszodim}@morphologic.hu
[2] MTA Nyelvtudományi Intézet, 1399 Budapest, Benczúr u. 33.
{vajda, varasdi}@nytud.hu

Kivonat: A cikkben bemutatunk egy tudásalapú anafora-feloldó rendszert, mely személyes névmások, zéró névmások, határozott névelős köznevek, valamint tulajdonnevek közötti koreferencia-viszonyok azonosítását végzi. A rendszer mély szintaktikai elemzésre, szintakszis-elmélet tételekre, pszicholingvisztikai kutatásokra, valamint a Magyar WordNet ontológiában tárolt nyelvi tudásra támaszkodik. A bemutatott módszerek a különböző típusú jelenségeket átlagosan 70%-os pontossággal képesek kezelni.

1 Bevezetés

Természetes nyelvű szövegekben az NP-koreferenciák feloldása egy adott dokumentumban eltérő pontokon megjelenő, de azonos entitásra referáló főnévi csoportok (NP-k) közötti viszonyok azonosítását jelenti. A feladat megoldása fontos olyan nyelvtechnológiai alkalmazások számára, mint a gépi fordítás, az információkivonatolás és egyéb szövegfeldolgozó alkalmazások ([7]).

A cikkben bemutatott, jelenleg még folyamatban lévő munka az alábbi koreferencia-jelenségek kezelésére – a visszautaló elem (anafora) és a szövegben korábban előforduló, vele koreferens NP (antecedens) közötti kapcsolat azonosítására – tesz kísérletet magyar nyelvű szövegekben ([10] alapján):

1. Táblázat: a vizsgált koreferencia-típusok, példák (az egymással koreferens NP-k félkövérrel kiemelve)

Típus	Példa
Ismétlés	Tegnap találkoztam **egy ismerősömmel**. **Az ismerősöm** nagyon sietett, mindössze pár percet beszéltünk.
Tulajdonnév-variáns	**Kovács Jakab** tegnap sajtótájékoztatót tartott. Az eseményen **Kovács úr** bejelentette az új termékeket.
Szinonima	Tamás kapott **egy biciklit**. Én is láttam **a kerékpárt**.
Hiper-/hiponima	Bejött **egy puli**. **Az állat** fáradtnak tűnt.
Névmás	Beszéltem **Julival**. Megadtam **neki** a számodat.
Zérónévmás	**Viktor** ismeri **Ferit**, de (**ő**) nem kedveli (**őt**) túlságosan.

Jelenleg nem foglalkozunk a személyes névmáson és bizonyos mutató névmásokon kívüli egyéb névmás-típusok (visszaható és kölcsönös névmások, vonatkozó névmások stb.) feloldásával. Főnévi csoportok alatt a mondatban előforduló maximális NP-ket értjük, melyek jellemzően a mondat főigéjének vonzatai, illetve főnévi eredetű szabad határozói. Jelenleg nem foglalkozunk a komplex, hierarchikus szerkezetű főnévi csoportok (pl. birtokos szerkezetek, koordinált NP-k stb.) összetevőivel. Így jelenleg nem foglalkozunk a birtokos szerkezetekben a morfológiailag jelölt számú-személyű, de a mondatban a birtoktól akár távolabbra is kerülő, vagy akár hiányzó birtokosnak megfelelő szerkezetek azonosításával. Szintén nem foglalkozunk az anaforát a szövegben követő antecedensű koreferencia-típus (katafora), valamint az epithetonnak nevezett jelenség („***Balázs*** *nem találta a kulcsát.* ***A szerencsétlen*** *nem tudott bejutni a lakásba.*") kezelésével.

A következő részben bemutatjuk a jelenleg működő, szabályalapú megközelítést alkalmazó koreferencia-feloldó rendszerünket, majd a kiértékelésére kialakított, annotált korpuszokat használó környezetet. Az utolsó részben részletesen ismertetjük a további lehetséges fejlesztések irányát.

2 A koreferencia-feloldó algoritmus

A koreferencia-feloldás kutatásának irodalmában az utóbbi időben megfigyelhető hangsúlyeltolódás a tudásalapú megközelítésektől az adatalapú, gépi tanulásra támaszkodó, a szabályalapú rendszerek teljesítményével vetekedő megközelítések felé ([8]). Számunkra azonban a munka kezdetekor nem állt rendelkezésre magyar nyelvre az adatalapú megközelítésekhez elengedhetetlen, jellemzően több ezer, kézzel annotált példából álló tanítókorpusz, így kénytelenek voltunk a tudásalapú megközelítések felől indítani.

Rendszerünk többféle tudásra támaszkodik. A legfontosabb inputot a MetaMorpho fordítóprogram-projektben fejlesztett magyar mondatelemző elemzéseiben kapott morfológiai, szintaktikai és szemantikai jegyek, nyelvtani szerepek, mélyszerkezeti elemzési struktúrák stb. jelentik. Ezekre támaszkodnak a kötéselmélet ([4]) és a magyar mondatmegértés kutatásainak ([9], [10]) eredményeire támaszkodó szabályaink. További, világismereti tudásra alapuló szabályok forrásaként a magyar WordNet ontológiát ([3]) használjuk. Végül a tulajdonnevek közötti referencia-azonosság felismeréséhez karakteralapú megközelítéseket alkalmazunk.

A feldolgozandó dokumentumokban balról jobbra haladva vizsgáljuk az egyes NP-ket. Minden, anaforikusnak feltételezett NP-hez legfeljebb egyetlen, korábbi NP antecedenst – a szövegben hozzá legközelebb esőt – rendelünk, a megközelítésünkben így a visszautalások láncokba szerveződhetnek (szemben a mindig a szövegben legelső antecedensre visszautaló annotálási megközelítésekkel.) Így a névmások, zéró névmások antecedensei lehetnek korábbi névmások, zérónévmások is.

A koreferencia-feloldás a teljes input dokumentum nyelvi elemzésével kezdődik. A bekezdésekre tagolt szöveg mondatainak mindegyikéhez a MetaMorpho elemzővel előállított szintakszisfák egyszerűsített változatát rendeljük, melyek a gyökércsomópont alatt csak a tagmondatoknak (CP), maximális igei frázisoknak (VP) és a főnévi csoportoknak (NP) megfelelő csomópontokat tartalmazzák. A szintaktikai elemző

gyakran (főként a hosszabb, összetett mondatok esetében) nem képes teljes, a mondat minden szavát lefedő elemzési fát előállítani, ilyenkor a rendelkezésre álló részelemzéseket használjuk fel (VP-k, NP-k, illetve főnévi eredetű határozói csoportok (ADVP)). Az azonosított főnévi csoportokban 25 jegy reprezentálja a MetaMorpho segítségével meghatározott pozicionális, lexikai, morfológia, szintaktikai és szemantikai tulajdonságokat.

A nyelvi előfeldolgozást követi a koreferencia-viszonyok feldolgozása, mely az antecedens-jelölteket szűrő megszorítások és a fennmaradó jelöltek közül választó preferenciák módszerén alapul ([7]). A módszer minden lépése a feloldandó anaforikus elem típusától (tulajdonnév, határozott névelős köznév vagy (zéró)névmás) függő szabályokat tartalmaz. Az általános algoritmus a következő 4 lépésben működik:

1. *Előszűrés*: az anaforikusnak feltételezett, tovább feldolgozandó NP-ket azonosítjuk. A jelenleg nem kezelt visszautaló elemek mellett próbáljuk felismerni és kizárni azokat a formailag visszautaló, azonban a szövegből kiutaló, tehát szövegbeli előzménnyel nem rendelkező NP-ket is, melyek további feldolgozása zajként jelentkezne ([12]). Ebbe a lépésbe beépítettünk 5 olyan heurisztikát is, melyek célja a nyelvi elemző által nagy valószínűséggel hibásan elemzett, így a koreferencia-feloldásban is szükségképpen hibát okozó NP-k felismerése és kizárása, pl. töredék-elemzésekben 2 szónál többet nem lefedő, névszói állítmány VP alá eső zérónévmások kizárása.

2. *Az antecedens-jelöltek listájának előállítása:* ebben a lépésben az anafora típusától függően a szövegben megadott távolságtól visszakeresve kijelöljük azokat a korábbi, az anaforának megfelelő típusú NP-ket, amelyek antecedensként szóba jöhetnek. A kötéselmélettel összhangban az anaforához legközelebbi antecedens-jelölt sem eshet az anaforával egy VP alá (mivel jelenleg nem kezeljük a visszaható és kölcsönös névmásokat.)

3. *A jelöltek szűrése:* ebben a lépésben megpróbálunk kizárni minél többet az antecedens-jelöltek közül (a konkrét módszer az anafora tipusától függ, ld. később), illetve a jelöltekre is alkalmazzuk az 1. lépésben ismertetett elemzési hiba-felismerő heurisztikákat.

4. *Antecedens kiválasztása a fennmaradó jelöltek közül:* az anafora típusától függő módszer szerint. Bizonyos típusú anaforák esetében az algoritmusnak kötelező kiválasztani egy jelöltet, mások esetében nem (ld. később.)

Az alábbiakban ismertetjük az algoritmus konkrét lépéseit a különböző anafora-típusok esetében.

2.1 Tulajdonnevek

Előszűrés: jelenleg nincs előszűrés (minden, a szövegben előforduló tulajdonnevet feldolgozunk).

Jelöltek: a jelöltek listázásának hatóköre a teljes megelőző dokumentum, az összes tulajdonnév NP-t hozzáadjuk a listához az anaforát tartalmazó VP kezdetéig.

Szűrés: jelenleg nincs szűrés.

Antecedens kiválasztása: az anafora és az antecedens-jelölt normalizálása (a kezdő determinánsok elhagyása, a fej tövesítése) után kiszámítjuk közöttük a Levenshtein-távolságot ([11]), melyet a hosszabbik string hosszával normalizálunk. Az algoritmusnak nem kötelező az antecedens-jelöltek közül választania, így az anaforához legjobban hasonlító (a legkisebb Levnshtein-távolságot mutató) antecedens-jelöltet csak azok közül a jelöltek közül választjuk ki, amelyek egy paraméterben meghatározott küszöbérték alatti hasonlóságot mutatnak (amennyiben a lista nem üres.)

2.2 Határozott névelős köznevek

Előszűrés: a „szemantikus NP”-knek ([12]) nevezett, közös világismeretből azonosítható unikus obejktumokra referáló, tehát a szövegben antecedenssel nem rendelkező határozott névelős közneveket próbáljuk meg felismerni és kizárni a feloldás alól (pl. „*az amerikai elnök”*). Ehhez jelenleg egy külön, előre összeállított listát használunk.

Jelöltek: tulajdonnevek és köznevek (determináns típusától függetlenül) az anafora teljes megelőző bekezdésében, az anafora VP-jéig.

Szűrés: jelenleg nincs.

Antecedens kiválasztása: a jelöltek közül meghatározzuk az anaforához legközelebb eső, vele azonos fejű NP-t (ismétlés), vagy szinonimát, vagy hipo-/hipernimát.

A szinonimitás vizsgálatához mind az anafora, mind az antecedens-jelölt lehetséges jelentéseit kikeressük a Magyar WordNetben, és ha van olyan synsetet, ami mindkettőt tartalmazza, szinonimáknak tekintjük őket. Mivel nincs jelentés-egyértelműsítés, a módszer nyilvánvalóan nem lesz minden esetben helyes.

Az anafora és az antecedens-jelölt közötti hipernima-viszony meghatározására a Leacock-Chodorow szemantikai hasonlósági formulát alkalmazzuk ([5]), amely a visszautaló és a jelölt összes WordNet-beli megfelelőit összekötő, hipernima-reláció szerinti útvonalak közül a legrövidebb alapján számít ki egy, az útvonal hosszától függő pontértéket. Hipernima/hiponima jelölteket csak az anaforát megelőző mondatban fogadunk el akkor, ha a Leacock-Chodorow hasonlósági képlet meghaladt egy paraméter küszöbértéket, és csak akkor, ha nem találtunk azonos fejű, vagy szinonim antecedenst a bekezdésben. Jelentés-egyértelműsítés hiányában a lexikális többértelműségek nyilvánvalóan itt is fognak hibákat okozni.

2.3 Névmások

Előszűrés: csak a zérónévmásokkal, személyes névmásokkal, valamint az „az" mutatónévmással foglalkozunk, feltéve, hogy utóbbi a VP-jében alanyi szerepben áll, és nem egy alárendelt tagmondatra utal. Nem foglalkozunk az első, illetve második személyű, ún. deixikus névmásokkal és zérónévmásokkal ([9]).

Jelöltek: az anafora mondata előtti második mondattól kezdve (amennyiben az létezik a bekezdésben) választjuk ki az összes NP-t, az anaforát tartalmazó tagmondat határáig.

Szűrés: az anafora és az antecedens-jelölt számának, személyének és két szemantikai jegyének (+/-élő, +/-ember) egyezését vizsgáljuk. Utóbbiak értéke lehet alulspecifikált (zérónévmások, illetve az elemző szótárában többértelmű főnevek esetében), ezek minden lehetséges értékkel kompatibilisek.

Kizárjuk továbbá azokat a lehetséges antecedenseket is, amelyekre már koreferenciát állapítottunk meg a vizsgált anaforával egy tagmondatban szereplő valamelyik másik névmási vagy zérónévmási anaforára nézve (ld. kötéselmélet).

Antecedens kiválasztása: egy mondatban mindig először az alanyi szerepű névmási anaforát oldjuk fel, és utána a többit (ha van). Így az előbb említett, már kötött antecedensek kizárásának segítségével kizárásos alapon is sok nem alanyi szerepű névmási anafora feloldható.

A (tag)mondatában alanyi szerepű névmási vagy zérónévmási visszautaló antecedensének meghatározásában Pléh Csaba és munkatársainak a magyar mondatmegértés pszicholingvisztikája körében végzett kutatási eredményeire támaszkodtunk ([10]). A heurisztika a szerkezeti párhuzamosság feltételezéséből indul ki, mely szerint az alanyi helyzetű anafora az előzménymondat alanyára utal vissza. Ezt felülbírálhatja az alanyi szerepben álló „az" mutatónévmás, ami alanyváltást jelöl:

(2a) **Hugó$_j$** felhívta **Amáliát$_k$**. **(Ő$_j$)** elmondta **neki$_k$** a történetet.
(1b) **Hugó$_j$** felhívta **Amáliát$_k$**. **Az$_k$** elmondta **neki$_j$** a történetet.

Alanyváltást egyéb jelenségek is előidézhetnek (pl. a második mondat predikátuma szemantikailag inkább a nem alanyi vonzatot preferálja stb.), ezekkel jelenleg nem foglalkozunk. Amennyiben a megelőző tagmondatban a szűrés után nem maradt rendelkezésre álló alany, az algoritmus a jelöltek listájában továbblép az azt megelőző tagmondat alanyára (amennyiben nem megy túl a bekezdés határán.)

„Az" formájú alany esetén, amennyiben az előzménymondatban több, nem alanyi szerepű antecedens-jelölt NP is található, az alábbi szabályok alapján választunk:

1. Hozzáférhetőség: az oblikuszi hierarchiában (tárgyi vonzat < egyéb vonzat < szabad határozó) magasabb helyen álló NP-t választjuk.

2. Távolság: a mondatában az anaforához közelebb eső NP-t preferáljuk (az oblikuszi hierarchiában azonos szinten álló NP-k közül).

Nem alanyi pozícióban álló névmások, zérónévmások esetén több, az alannyal nem koreferens antecedens-jelölt közül szintén a fenti két szabály alkalmazásával választunk.

A koreferencia-feloldást először minden mondatban a tulajdonnevekre, határozott névelős köznevekre végezzük el, ez után következik a mondat névmási, zérónévmási anaforáinak feldolgozása. Reményeink szerint ezzel további segítséget adunk a névmási anaforák feloldásának a szűrési feltételekben leírt szabály alkalmazásával.

3 Kiértékelés

A koreferencia-feloldó modul pontosságának kiértékelése jelenleg is folyamatban van, így csak részleges eredményekről tudunk az alábbiakban beszámolni. A kiértékeléshez létrehoztunk egy kézzel annotált kiértékelő-korpuszt, amely 5 darab, általános iskolai történelemkönyvekből kiemelt szöveget tartalmaz (2. Táblázat). A szövegekben a MetaMorpho segítségével azonosítottuk a maximális NP-ket, majd annotáltuk közöttük a koreferencia-viszonyokat. Az automatikus annotációhoz hasonlóan a koreferencia-láncokban mindig az anaforához legközelebbi antecedenseket jelöltük be. A munkát egyetlen annotátor végezte.

Mivel a nyelvtani elemző nem minden NP-t ismert fel, illetve egy részüket hibásan, csak a jól felismert NP-ket tudtuk annotálni (és azokat is csak akkor, ha az antecedensük is helyesen volt bejelölve), így fedés(recall) kiértékelésére a korpusz jelenleg nem alkalmas.

2. Táblázat: a kiértékelő korpusz jellemzői

Szövegek száma	5
Bekezdések száma	79
Mondatok száma	652
NP-k száma	3115
Antecedenssel annotált NP-k száma	338

A koreferencia-feloldó algoritmust ezután lefuttattuk a korpusz szövegein, majd összevetettük az automatikus annotáció eredményeit a kézivel. A rendszer 145 NP-re adott eredményt, ezek közül 101 egyezett meg az annotációval (69 %-os átlagos pontosság.) Ezután megvizsgáltuk a különböző visszautalási típusok felismerésének pontosságát külön-külön is (3. Táblázat.) A kiértékeléskor nem állt rendelkezésre a tulajdonnevek koreferenciájának felismerése, így az erre vonatkozó adatok hiányoznak a táblázatból.

Szembeötlő a különbség a hipernimán alapuló és a többi módszer között. A hipernima-kereső módszer nélkül az algoritmus átlagos pontossága 80%-os lenne.

3. Táblázat: a különböző koreferencia-feloldó módszerek pontossága

Visszautalási típus	Automatikusan annotált NP	Helyesen annotált NP	Pontosság (%)
Ismétlés	21	20	97%
Szinonima	8	6	75%
Névmás	110	75	68%
Hipernima	6	0	0%

Kíváncsiak voltunk arra, hogy mennyiben befolyásolja a nyelvi elemző a névmási anafora-feloldás teljesítményét, ezért egy másik szövegen részletesen megvizsgáltuk a névmási anafora-feloldás hibáit. A szöveg 109 mondatot tartalmazott, a rendszer ezekben 521 db NP-t jelölt be, melyek közül 34 db névmási NP-hez azonosított antecedenseket. Az automatikusan azonosított antecedensek mindegyikét kézzel helyes vagy hibás eredményként értékeltük, és a hibás automatikus annotáció alábbi három esetét különítettük el

a) a hiba a helytelen nyelvi elemzés következménye (rosszul elemzett anafora és/vagy rosszul/nem elemzett antecedens) (*jelölés: KO_parser*)
b) az anaforának nem volt a szövegben antecedense (a program nem ismerte fel, hogy az elem nem utal vissza) (*jelölés: KO_noant*)
c) az anaforának volt a szövegben antecedense, de a program helytelenül azonosította (*jelölés: KO_cr*)

A helyes elemzések és a különböző hibatípusok arányait a 4. Táblázatban foglaltuk össze.

4. Táblázat: az anafora-feloldás hibatípusainak aránya

Eredmény típusa	előfordulása	%
OK	24	67%
KO_parser	7	20%
KO_noant	0	0%
KO_cr	3	8%

A táblázatból látható, hogy az automatikus annotáció hibájának nagy százaléka a nyelvi elemző hibájának következménye. Hibák nélküli szintaktikai elemzésre támaszkodva a névmási anafora-feloldás pontossága a jelenlegi algoritmussal akár 83%-os pontosságot is elérhetne.

A névmási anafora-feloldás az általunk vizsgált mintában nem azonosított antecedenseket olyan NP-khez, melyeknek nincs szövegbeli előzménye. Ennek oka, hogy a névmási visszautalások általunk kezelt fajtái mindig létező, szövegbeli antecedensre utalnak, hibát csak a helytelen nyelvi elemzésből származó, invalid NP-k okozhatnak, azonban az ezeket kiszűrő heurisztikák jól működtek.

4 További munka

Elsőként szeretnénk folytatni a jelenlegi rendszer teljesítményének kiértékelését. Ehhez a tulajdonnevek feloldásának kiértékelése, másrészt a fedés vizsgálati módszerének kidolgozása szükséges. Ezután természetesen a hibák részletes kategorizálása és elemzése következik, különös tekintettel a hiponimák azonosításának módszerére, amely lényegesen rosszabbul teljesített a többi módszerhez képest.

Szeretnénk egy baseline megoldásnak megfelelő algoritmust implementálni, amelyhez képest meghatározható a rendszerünk teljesítménye. Ehhez a Centering elméletre alapuló, a szakirodalomban jól ismert BFP-algoritmust ([1]) szeretnénk kipróbálni, melyet vizsgáltak már magyar szövegekkel is ([6]).

Szeretnénk létrehozni egy olyan, koreferenciával annotált kiértékelő korpuszt, ami mások számára is hozzáférhető, így a rendszerünk teljesítménye más hasonló rendszerekkel is összevethető lesz. Ehhez legalkalmasabbnak a frázis-annotációkat tartalmazó Szeged Treebank 2.0-s változata tűnik ([2]). A Szeged Treebank használatával nyelvi elemzőtől független, nagy pontosságú szintaktikai elemzésekre lehetne koreferencia-feloldó algoritmusokat építeni (ugyanakkor bizonyos jegyek, melyek a MetaMorpho kimenetében azonosíthatók, nem lesznek elérhetők.)

Az anafora-feloldás fedésének növelésére további főnévi anaforikus jelenségek kezelésére lesz szükség: visszaható és kölcsönös névmások, vonatkozó névmások, birtokos névmások valamint a komplex NP-k részegységeinek, a birtokos szerkezeteknek stb. elemzésére és koreferencia-kapcsolataik feltárására.

A pontosság növelése érdekében a tulajdonnevek felismerésére további karakterhasonlóságon alapuló módszereket és normalizációs eljárásokat mutat be [11]. A karakteres és a szemantikai hasonlóság-képletek számára a küszöbértékeket empirikus úton, korpuszpéldák segítségével lenne célszerű optimalizálni.

A tulajdonnevek és köznevek feloldásánál további, felhasználható információ lehet az anafora és az antecedenst egymástól való távolsága. A MetaMorpho lexikonjában tárolt szemantikai jegyek (pl. tulajdonnév-osztályok, szemantikai kategóriák stb.) egyezésének vizsgálata további szűrési feltételeket adhat.

Határozott névelős közneveknél felmerül a kérdés, hogy az azonos fejű, számban egyező, de eltérő módosítókat tartalmazó anafora–antecedens-jelölt párokat hogyan kezeljünk (pl. *a katonák–az út szélén elrejtőzött katonák*.)

A névmási anaforák kezelésénél további heurisztikák alapja lehet a Centering Elmélet, a diskurzustopik változásának figyelése ([1]), illetve [10] által leírt egyéb jelenségek modellezése (pl. a predikátum által preferált vonzatok korpuszstatisztikai vizsgálata.)

További lehetőség a zajt okozó, feloldást nem igénylő NP-k azonosítása további módszerekkel. Az egyik a szükségszerű/valószínű rész viszony, pl. „*Tegnap szerelőhöz vittem a* ***biciklim****, mert eltört a* ***pedál***.”. Ha rendelkezésre állna megfelelő adatbázis, az esetkeretből levezethető entitásokat is fel lehetne ismerni, pl „***A konferencia*** *véget ért.* ***A résztvevők*** *elégedetten távoztak.*”

Bibliográfia

1. Brennan, Susan E., Marilyn W. Friedman, Carl J. Pollard. A centering approach to pronouns. In Proceedings of the 25th Meeting of the Association for Computational Linguistics (1987), pp. 155-162.
2. Csendes D., Alexin Z., Csirik J., Kocsor A.: A Szeged Korpusz és Treebank verzióinak története. III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005) kiadványa, Szeged, december 8-9., pp. 409-412 (2005)
3. Hatvani Cs., Kocsor A., Miháltz M., Szarvas Gy., Szécsi K.: Főnevek a Magyar WordNetben. IV: Magyar Számítógépes Nyelvészet Konferencia, Szeged (2006), pp. 109–116.
4. Kenesei István: Az alárendelt mondatok szerkezete. In: Kiefer Ferenc (szerk.): Strukturális Magyar Nyelvtan, I. kötet, Mondattan. Akadémiai Kiadó, Budapest (1992)
5. Leacock, C., M. Chodorow: Combining Local Context and WordNet Similarity for Word Sense Identification. In C. Fellbaum (ed.): WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA (1998), pp. 265–285
6. Lejtovicz Katalin, Kardkovács Zsolt: Anaforafeloldás magyar nyelvű szövegekben. IV. Magyar Számítógépes Nyelvészet Konferencia, Szeged (2006)
7. Mitkov, Ruslan: Anaphora Resolution: The State of The Art. Working Paper, University of Wolverhampton, 1999.
8. Ng, Vincent: Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. Proceeding of the 43rd Annual Meeting of the Associaction for Computational Linguistics (1995)
9. Pléh Csaba, Radics Katalin: „Hiányos mondat”, pronominalizáció és a szöveg. In Általános Nyelvészeti Tanulmányok, XI, 261-277 (1976).
10. Pléh Csaba: Mondatközi viszonyok feldolgozása: az anafora megértése a magyarban. In: Pléh Csaba: Mondatmegértés a magyar nyelvben. Osiris Kiadó, Budapest (1998)
11. Uryupina, Olga: Evaluating Name-Matching for Coreference Resolution. In Proceedings of the 4th International Conference on Language Resources and Evaluation (2004)
12. Varasdi Károly: Koreferenciák feloldása. Projektdokumentum (2005)

V. Gépi tanulás

Eljárás radiológiai leletek automatikus BNO kódolására

Farkas Richárd[1] és Szarvas György[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
rfarkas@inf.u-szeged.hu
[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
szarvas@inf.u-szeged.hu

Kivonat: Cikkünkben egy amerikai kórházak és kutatóintézetek által, 2007 tavaszán rendezett nyílt verseny eredményeiről számolunk be. A verseny célja radiológiai leletek automatikus címkézése volt ICD-9-CM kódokkal (a Betegségek Nemzetközi Osztályozásával /BNO/ megegyező, számlázáshoz használt kódrendszer). A feladat érdekességét más, korábbi szövegfeldolgozási versenyekhez hasonlítva a szöveghez rendelendő kódok nagy száma, illetve a kódrendszer címkéi közti belső összefüggések adták (összesen 45 kód 96-féle különböző kombinációja fordult elő a korpuszban). A leletek automatikus osztályozását lehetővé tevő számítógépes eljárások fejlesztése létfontosságú, hiszen orvosi témájú szöveges dokumentumok kódolására, illetve a feladat során keletkező hibák javítására évi mintegy 25 milliárd dollárt fordítanak, pl. az Egyesült Államokban. A versenyre benyújtott rendszerek tanulsága, hogy a klinikai dokumentumok – emberi pontossághoz közelítő – eredményes feldolgozása nem lehetetlen célkitűzés a napjainkban rendelkezésre álló eszközökkel.

1 Klinikai dokumentumok feldolgozása (bevezetés)

A számítógépes ontológiák, valamint strukturált szótárak fejlesztése terén tapasztalható fejlődés ellenére a legtöbb kórházban az adatok tárolásának részben továbbra is folyó szöveg formájában történik. Ez a gyakorlat sok gigabájtnyi szöveges adatot eredményez, melynek – az adott beteg klinikai kezelésén túl – korlátozott a hasznosíthatósága, az adatok mennyisége és hozzáférhetősége folytán. Nyelvtechnológiai eszközökkel lehetséges e nagy mennyiségű, szöveges adatban a rejtett struktúra felfedezése, és a tárolt információ elérhetővé tétele. Az így kinyert strukturális információ felhasználható keresőalkalmazások, számlázás, minőségbiztosítás céljára, de gyógyszerkutatásokhoz is igen hasznos lehet (adott betegségek ill. tünetek milyen kórtörténetű páciensek, milyen lefolyással stb.). Korábbi munkákban [5] megmutattuk, hogy ez kivitelezhető akár olyan, összesített formában, mely nem sérti az egyes betegek személyiségi jogait.

A klinikai dokumentumok szövegei számos, a hétköznapi nyelvhez képest eltérő tulajdonságot mutatnak [2]. Ilyenek pl. a hiányos (pl. állítmány nélküli), rövid mondatok; a rövidítések gyakori használata; speciális írásjelezés; szokatlan metonímiák.

Ezek a jelenségek többnyire korlátozzák az általános célokra fejlesztett nyelvtechnológiai programcsomagok alkalmazhatóságát orvosi szövegeken.

1.1 Klinikai dokumentumok BNO-kódolása

Az ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification, magyarul Betegségek Nemzetközi Osztályozása, BNO) kódjait a kórházi ellátás során elvégzett vizsgálatok, kezelések dokumentálására használják az Egyesült Államokban. Az egészségbiztosítók kifizetései a klinikai dokumentumokhoz rendelt BNO címkék alapján történnek, melyeket a kezelés után, utólag rendelnek a kórházi dokumentációhoz. Maga a címkézési eljárás, melyet szakképzett munkaerő (pl. orvosok) végez, illetve az ezzel kapcsolatos hibák javításának költségét évente mintegy 25 milliárd dollárra becsülik [4] az USA-ban. Emiatt a címkézési eljárás automatizálása, illetve támogatása nyelvtechnológiai megoldásokkal intenzíven kutatott, piaci szempontból is fontos feladat.

Maga a címkézés hivatalos kódolási útmutatók alapján történik (pl. [3]), melyekben számos, számítógépes nyelvészeti szempontból is érdekes előírást találunk. Ilyenek pl.:

- Bizonytalan diagnózis semmi esetben sem kódolható (bizonytalanság, spekulációk, illetve tagadás azonosítása).
- Tünetek kódjait tilos a dokumentumhoz rendelni, amennyiben egy kapcsolódó betegség kódja már hozzárendelésre kerül (címkék közti összefüggések modellezése).
- Múltbeli betegségek, kezelések, illetve amelyek közvetlenül nem kapcsolódnak a kezelésekhez, nem szabad kódolni (múlt idő felismerése, alany meghatározása). Gyakran felsorolnak kórtörténeti utalásokat, illetve a család és tágabb környezet tagjainak problémáit.

Mivel a címkéket elsősorban könyvelési és dokumentációs célokra használják, a címkézés precizitása anyagi szempontból is fontos az egészségügyi intézetek számára. A dokumentumhoz hozzá nem rendelt kódok bevételkieséssel járnak a kórháznak, míg minden tévesen hozzárendelt kódért - ha a tévedésre fény derül - a számlázott összeg háromszorosával büntetik az intézményt (valamint a csalásnak egyéb jogi következményei is lehetnek).

1.2 A verseny leírása

A 2007 nyarán megrendezett nemzetközi versenyre[1] [4] a szervezők mintegy 2000 radiológiai leletet láttak el a megfelelő ICD-9-CM címkékkel. Az elkészült korpusz (melynek etalon címkézése három, a kódolást egymástól függetlenül elvégző szervezet többségi annotációja lett) felét adták ki a résztvevőknek tanító adathalmazként, míg a fennmaradó dokumentumhalmazon a szervezők végezték el a beérkező eredmények kiértékelését. A kiértékeléshez, illetve a beküldött rendszerek rangsorolására

[1] részletes leírása található a www.computationalmedicine.org/challenge oldalakon

címke szintű $F_{\beta=1}$ értéket használtak (a továbbiakban minden eredményt mi is eszerint közlünk).

2 Eredményeink

Az alábbiakban ismertetjük a versenyre benyújtott rendszerünk főbb jellemzőit, illetve annak fejlesztése és a későbbi kapcsolódó munkáink során nyert tapasztalatokat.

2.1 Nyelvi modell

Tapasztalataink szerint a feladat megoldásában szerepet játszó legfontosabb nyelvi jelenségek a következők voltak: tagadás (egy betegség vagy tünet nemléte nem kódolandó), feltételes/spekulatív nyelvhasználat (a kódolási útmutatók szerint bizonytalan diagnózist semmilyen esetben sem szabad kódolni), illetve múlt idő kezelése (olyan tünetek és betegségek jelennek meg a dokumentum címkézésében, melyeknek hatása van a dokumentum keletkezésekor végzett vizsgálatokra, a fennálló betegségre).

A versenyre benyújtott modellünkben a tagadás és a spekuláció kezelésére készítettünk szótáron és kötött illesztési szabályokon alapuló megoldást. Az időbeliséget a szakértők inkonzisztens címkekódolása miatt nem kezeltük (volt olyan betegség, melynek jelen és múltbeli lefolyásához külön kód tartozott, de a két címke használatában semmi szabályszerűséget sem találtunk a manuálisan annotált korpszban). A spekulatív vagy tagadó szövegelemek hatókörét az írásjelek segítségével állapítottuk meg. Ez a megoldás a tagmondathatárok azonosításának egy durva egyszerűsítése, ami a feldolgozandó szövegek speciális volta miatt itt hatékonynak bizonyult.

A feladat megoldása során döntő tényező volt a lényeges nyelvi jelenségek pontos kezelése (a tagadás szűrése 10.66%, a feltételes mód kezelése 9.6%, a kettő együtt 18.56% pontosságnövekményt eredményezett, míg a végleges rendszer hibáinak egy számottevő részét épp a nem kezelt múltbeli hivatkozások adják).

2.2 Sokosztályos dokumentumosztályozási feladat

A fenti nyelvi előfeldolgozási lépések után minden címke hozzárendelését önálló osztályozási feladatként oldottuk meg (45 db bináris osztályozási feladat). Az így óhatatlanul előálló érvénytelen kombinációkat nem kezeltük. Az egyes feladatokra felírt modellünk a kódolási útmutatók előírásait követő szabályalapú rendszer és a dokumentumok vektortérmodellben ábrázolt példáin tanított statisztikai osztályozók (melyek a szabályalapú rendszer hibás predikcióit igyekeztek modellezni) kombinációjából állt. Ezeket külön-külön részletesen bemutatjuk a következő fejezetekben.

2.3 Szabályalapú BNO-kódoló alkalmazás

A BNO-kódoláshoz használható többé-kevésbé strukturált útmutatók számos helyen elérhetők az Interneten[2]. Egy – ezek valamelyikének felhasználásával készített – automatikus BNO-kódoló, mely az útmutatóban talált kifejezések illesztésével rendel kódokat a leletekhez, minimális emberi beavatkozással előállítható.

1. Táblázat: Hivatalos útmutató konverziója BNO-kódoló szabályrendszerré

KÓDOLÁSI ÚTMUTATÓ	**GENERÁLT DÖNTÉSI SZABÁLYOK**
label *518.0*	**if document contains**
Pulmonary collapse	pulmonary collapse **OR**
Atelectasis	atelectasis **OR**
Collapse of lung	collapse of lung **OR**
Middle lobe syndrome	middle lobe syndrome
Excludes:	**AND document NOT contains**
atelectasis:	
congenital (partial) (770.5)	congenital atelectasis **AND**
primary (770.4)	primary atelectasis **AND**
tuberculous, current disease (011.8)	tuberculous atelectasis
	add label *518.0*

Egy ilyen egyszerű rendszer, melynek kifejlesztése címkézett példák meglétét sem igényli (nincs modelltanítási fázis), meglepően jó pontosságot ér el a leletek osztályozásában (a megfelelő nyelvi előfeldolgozás után). Az NCSH alapján készült szakértői modellünk a verseny tanító adatbázisán 84.07%, a teszt adatokon 83.21% pontosságot mutatott. A módszer legnyilvánvalóbb korlátai, hogy az útmutatókban található kifejezés-lista fedése nem teljes, valamint, hogy azokban nincs utalás a címkeközi (betegség-tünet) összefüggésekre, amiket így a modell nem kezel.

2.4 Címkeközi összefüggések feltárása

A címkeközi összefüggések elsősorban betegségek és tünetek kódjai közt állnak fent. Például a *tüdőgyulladás* (pneumonia, 486-os kód) szövegbeli megjelenése magával hozza a *köhögés* (kód 786.2) és a *láz* (kód 780.6) tünetek előfordulását is a szövegben, azonban ilyen esetekben csak a betegséget szabad címkézni. Az ilyen jellegű összefüggéseket nem tartalmazza a kódolási útmutató, így annak alkalmazása igen gyakran túlkódolja a dokumentumokat tünet-címkékkel.

Ennek orvoslására azt a gépi tanulási feladatot fogalmaztuk meg aminek outputja olyan szabályokat tartalmaz, amelyek bizonyos betegség kódolása esetén bizonyos tüntetet eltávolít a predikált címkehalmazból (hamis pozitív címkézések leválasztása).

---

[2] Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/research/umls/)
National Center for Health Statistics - Classification of Diseases, Functioning and Disability (http://www.cdc.gov/nchs/icd9.htm)
ICD9Data.com: Free 2007 ICD-9-CM Medical Coding Database (http://www.icd9data.com/)

Az így nyert címkeközi relációkkal (5 db. szabály) bővített modell körülbelül 1.5%-os javulást hozott, a tanító adatbázison 85.57%, a teszt halmazon 84.85%-os pontosságot elérve.

Megvizsgáltuk manuálisan is, hogy milyen címkeközi relációkat lehet érdemes felvenni. A manuális és a statisztikai módon nyert címkeközi szabályhalmaz megegyezett, ez a lépés tehát – ezen adatbázis alapján állíthatjuk – gépi tanulási módszerekkel kivitelezhető.

2.5 A nyelvi modellen alapuló adatreprezentáció

A statisztikai modellek alkalmazása nyelvtechnológiai problémákra a vizsgált szövegek gépi feldolgozásra alkalmas ábrázolását teszi szükségessé. Dokumentumosztályozási feladatok megoldására használt leggyakoribb adatreprezentációs modell az úgynevezett vektortérmodell. A vektortérmodell egy sokdimenziós vektortérben ábrázolja a dokumentumokat, a vektortér dimenziói pedig a dokumentumgyűjteményben előforduló egyedi szavak. Hátrány, hogy ebben a reprezentációban a szavak szövegen belüli pozíciójára és sorrendjére vonatkozó információ elvész. Ennek ellenére számos feladatra nagyon jól működő, vektortérmodelles megoldást fejlesztettek ki.

A BNO kódolási feladathoz szükség volt a fentebb ismertetett nyelvi előfeldolgozási lépések elvégzése a dokumentumokon. Azaz a nyelvi modellek által megjelölt tagadott vagy spekulatív szövegrészeket eltávolítottuk a szövegekből, és csak a maradék, tényszerű információkat tartalmazó szövegrészeket használtuk a dokumentumok ábrázolásához.

2.6 Vektortérmodellen alapuló statisztikai BNO-kódoló alkalmazás

A fenti szabályalapú osztályozóhoz hasonló feladatot ellátó statisztikai modell építhető címkézett példák segítségével a (nyelvi előfeldolgozás utáni) dokumentumok vektortérmodellbeli ábrázolását használva, gépi tanulási módszerek segítségével. Ennek a reprezentációnak, és a versenyen közzétett tanító adatbázisnak a használatával a tanító adatokon 88.20%, a tesztadatbázison 86.69% pontosságot mutató modellt kaptunk. Ennek a megközelítésnek a hátránya, hogy azokra a címkékre, amelyekre csak nagyon kevés példa állt rendelkezésre (22 olyan címke volt a 45-ből, aminek a tanító-adatbázisbeli gyakorisága 6 vagy az alatt volt) nem építhető megbízható statisztikai alapú modell, míg előnye, hogy az útmutatóban nem szereplő szinonimákat, rövidítéseket is fel tudja fedezni.

2.7 Statisztikai modell és szakértői szabályrendszer kombinálása

A további kísérletekben azokra a kérdésekre kerestük a választ, hogy a statisztikai alapú (tanító-adatbázison épített) modell és a külső szakértői tudásbázis (tanító adatbázistól független információ) milyen módszerekkel kombinálható és ezzel a hibrid

modellel milyen eredmények érhetőek el. A két modell előnyeinek egyesítésére számos lehetőség kínálkozik:

- A szabályalapú rendszer illesztési szabályainak bővítése, illetve finomítása a címkézett tanító adatok alapján: Ez a folyamat történhet a kiinduló szabályrendszer hibáira (hamis pozitív és negatív címkék) felírt gépi tanulási problémák megoldásával. Ezt a megközelítést döntési fa illetve Maximum Entrópia osztályozók használatával is teszteltük. C4.5 döntési fa osztályozó segítségével 90.22% és 88.92% pontosságot értünk el a tanító ill. tesztadatbázisokon, míg Maximum Entrópia modellel 90.26% és 88.93% pontosságot sikerült elérni.
- A gépi tanulási modell kibővítése a szabályalapú rendszer tudásbázisával: Ennek a megoldásnak a legegyszerűbb formája, a dokumentum vektortérmodellbeli ábrázolását további jellemzőkkel bővítjük, melyek a szabályalapú rendszer címkézését (kimenetét) írják le. Ekkor a gépi tanulási modellnek lehetősége van kiaknázni mind a kódolási útmutatóban rejlő tudást, mind a címkézett adatok vizsgálatával nyerhető mintákat. Az ilyen kombináció egyik nyilvánvaló gyenge pontja, hogy a statisztikai modell csak a szabályalapú rendszer megfelelő számú példával alátámasztott jelöléseit képes integrálni a tanult modellbe. Ez a módszer 90.62% és 87.92% pontosságot mutatott a tanító és tesztadatbázisokon.
- A szabályalapú modell, valamint a gépi tanulási modell együttes használata (kaszkád modell). Ekkor a modellek kombinációja helyett együttesen alkalmazzuk a két osztályozót, azaz a két modell által predikált címkehalmazok uniója lesz a rendszer végső kimenete. Ez a megközelítés 90.53% és 89.33% pontosságot ad.

A szabályalapú rendszer illesztési szabályainak bővítése a címkézett tanító adatok alapján (beleértve a címkeközi összefüggések kezelését is) történhet a leletek tanulmányozásával, manuális módon is. E módszer – noha nagyon jó modellt eredményez – munkaigényessége, illetve skálázhatósága miatt (több ezer BNO kódra kivitelezhetetlenné válik a munkaigénye miatt) gyakorlati szempontból nem ideális megoldás. Egy, ezt a megközelítést követő modell 90.02% valamint 89.41% pontosságot mutatott a tanító illetve tesztadatokon. Ezek az eredmények tekinthetőek a fenti három hibrid modell által elérhető elméleti felső korlátnak.

2. Táblázat: A különböző modellek, és pontosságuk

	tanító adatbázis	teszt adatbázis
45-osztályos statisztikai modell	88.20%	86.69%
Szabályalapú BNO-kódoló	84.07%	83.21%
Szabályalapú r. címkeközi összefüggésekkel	85.57%	84.85%
Hibrid1 (kiterjesztett szabályalapú)	90.26%	88.93%
Hibrid2 (kiterjesztett statisztikai)	90.62%	87.92%
Hibrid3 (szabályalapú + statisztikai kaszkád)	90.53%	89.33%
Kézileg fejlesztett szabályalapú modell	90.02%	89.41%

3 Értékelés

Az alábbiakban értékeljük az általunk adott BNO-kódolási modell hatékonyságát az emberi címkézéshez, illetve a versenyre benyújtott modellekhez viszonyítva.

3.1 Egyetértés az annotálást végző szervezetek és a modelljeink között

A fentebb ismertetett eredmények megközelítik azt a pontosságot, amit képzett szakemberek képesek elérni a címkézési eljárás végrehajtásában. Az etalon címkézés 3 független (BNO-kódolással a versenytől függetlenül is foglalkozó) egészségügyi szervezet jelöléseinek többségi szavazásával állt elő, azaz minden olyan címke szerepelt az etalon címkézésben, amit legalább két szervezet javasolt.

3. Táblázat: Az annotátorok egyetértési rátái egymással, az etalon címkézéssel valamint két modellünkkel.

	A1	A2	A3	GS	Szabály	Hibrid
Annotátor1	—	73.97/75.79	65.61/67.28	83.67/84.62	75.11/75.56	78.39/79.42
Annotátor2	73.97/75.79	—	70.89/72.68	88.48/89.63	78.52/78.43	83.60/83.14
Annotátor3	65.61/67.28	70.89/72.68	—	82.01/82.64	75.48/74.29	80.00/78.80
Gold Standard	83.67/84.62	88.48/89.63	82.01/82.64	—	85.57/84.85	90.53/89.33
Szabályalapú r.	75.11/75.56	78.52/78.43	75.48/74.29	85.57/84.85	—	—
Hybrid	78.39/79.42	83.60/83.14	80.00/78.80	90.53/89.33	—	—

Fontos megjegyezni, hogy a címkézést végző szervezeteknek nem volt hozzáférésük az etalon címkékhez, míg az adatbázison tanított statisztikai modellek közvetlenül az etalon címkézést modellezhették. Ennek megfelelően, ha a címkézést végző szervezeteknek lehetőségük lett volna a többségi címkézés jellegzetességeit tanulmányozni, várhatóan nagyobb egyetértési rátát lennének képesek. Másrészről viszont a 3 annotálást végző szervezetnek hatása volt az etalon címkékre, mivel azok az ő többségi szavazásukkal álltak elő. Ez a tény magyarázza, hogy mindhárom szervezet nagyobb egyetértési rátát mutat a többségi címkézéssel, mint a szervezetek címkézései egymással összehasonlítva. Fair összehasonlítást egy olyan, negyedik szervezet által a dokumentumokhoz rendelt címkézés és az etalon jelölés között lehetne tenni, akik előzetesen tanulmányozhatták az etalon címkéket, de arra nem volt kihatásuk. Ez utóbbi statisztika mutatná jól a feladatban elérhető elvi felső korlátot, azaz a képzett szakemberek teljesítményét a BNO-kódolási feladaton.

A különböző szervezetek jelölései között megfigyelhető, feltűnően alacsony egyetértési ráták arra engednek következtetni, hogy az egyes szervezeteknek saját BNO-kódolási stílusuk, szokásaik vannak. A táblázatban szerepeltetjük azon, szabályalapú modellünk és a szervezetek egyetértési rátáit is, mely a BNO-kódolási útmutatók (és nem a címkézett adatok) modellezésével készült. Ez a rendszer az útmutató kifejezéseit illeszti, és kezeli a címkeközi függéseket (az útmutatók általános előírásai szerint), azaz fair összehasonlítást ad a jelölést végző szervezetekkel. Az a tény, hogy a 3

szervezet kissé magasabb egyetértést mutat ezzel a modellel, mint egymással, azt sejteti, hogy a szervezetek sajátos kódolási szokásai a hivatalos BNO-kódolási útmutató eltérő értelmezéséből fakad; illetve, hogy a többségi címkézés jobban közelíti a hivatalos előírásokat, mint a szervezetek egyedi címkézései.

3.2 Összehasonlítás a versenyre benyújtott rendszerekkel

Összesen 50 résztvevő indult a versenyen, a beküldött modellek 89.08% és 15.41% közötti pontosságot mutattak, 76.7% átlagos érték mellett[3]. Összesen 21 rendszer ért el 80% feletti teljesítményt, mely – korlátozott mértékben – összevethető az emberi annotáció pontosságával. A legjobb modellek az emberi címkézés pontosságát közelítik, azaz a klinikai szövegek jó pontosságú gépi feldolgozása reális célkitűzés.

A versenyre beküldött, második helyezést elérő rendszer [1] 88.55% pontosságot ért el, azaz az általunk automatikusan készített legjobb modell (89.33%) megközelíti, vagy kissé meghaladja a más megközelítésekkel elért eredményeket. Ez alapján elmondható, hogy a cikkünkben ismertetett, szabályalapú és gépi tanulási modellek kombinációján alapuló megoldás versenyképes eredményt ad klinikai dokumentumok osztályozásában. Mivel a kézi szabályrendszer kifejlesztésének főbb, munkaigényes lépéseit nagyrészt sikerült automatikusan is reprodukálnunk, – az eredmények jelentős romlása nélkül –, a rendszerünk a versenyben használtnál jelentősen nagyobb számú BNO-kódra is megvalósítható, skálázható lenne.

3 Konklúzió

A kódolási útmutató előírásait alapul vevő szabályalapú szakértői rendszerek meglepően hatékonynak bizonyultak a radiológiai leletek BNO kódolásánál. Ezeken a külső szakértői tudást reprezentáló rendszereken további javítást tudtunk elérni statisztikai gépi tanulási modellek és a szabályhalmazok megfelelő összekapcsolásával. A versenyre beküldött rendszerünk 89.08% pontosságot ért el a leletek egészségügyi kódrendszer kategóriáiba való besorolásában.

A verseny fejlesztési és értékelési időszakát követően végzett, a modellünk skálázhatóságát célzó kutatásaink azt mutatták, hogy a kézileg kiélezett szabályrendszerek helyett (melyek kifejlesztése több ezer kódra időigényes, kivitelezhetetlen lenne) hasonló eredményt érhetünk el, ha a kódolási útmutatókból többé-kevésbé automatikus konverzióval egy kezdeti szabályrendszert állítunk elő, majd ezt a címkézett adatok felhasználásával gépi tanulási modellekkel teljesen automatikusan fejlesztjük tovább.

Hasonló szövegfeldolgozási problémák magyar nyelven való vizsgálatához jelenleg kutatási partnereket keresünk.

[3] Részletes kimutatás található a http://www.computationalmedicine.org/challenge/res.php oldalon.

Bibliográfia

1. Goldstein, I., Arzumtsyan, A., Uzuner, Ö.: Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports. Proceedings of the Fall Symposium of the American Medical Informatics Association (AMIA 2007), Chicago, IL, November 10-14, (2007)
2. Lang, D.: Natural Language Processing in the Health Care Industry, Consultant Report, Cincinnati Children's Hospital Medical Center, Cincinatti, Ohio, USA (2006)
3. Moisio M. A.: A Guide to Health Insurance Billing. Thomson Delmar Learning, USA (2006)
4. Pestian J. P., Brew C., Matykiewicz P., Hovermale D. J., Johnson N., Cohen K. B., Duch W.: A shared task involving multi-label classification of clinical free text, In Biological, translational, and clinical language processing, Prague, Czech Republic, 97–104 (2007)
5. György Szarvas, Richárd Farkas, Róbert Busa-Fekete: State-of-the-art anonymisation of medical records using an iterative machine learning framework. Journal of the American Medical Informatics Association, Volume 14, Issue 5, pp 574-580 (2007)

Magyar jelentés-egyértelműsített korpusz

Szarvas György[1], Hatvani Csaba[2], Szauter Dóra[1],
Almási Attila[1], Vincze Veronika[1] és Csirik János[1]

[1] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
{szarvas, csirik}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
szauter.dora@freemail.hu, vizipal@gmail.com,
{vinczev, hacso}@inf.u-szeged.hu

Kivonat: Az első magyar WSD korpusz elkészítéséhez 39 olyan szóalakot választottunk ki, melyek jó mintapéldák a jelentés-egyértelműsítés feladatának vizsgálatára. A kiválasztásnál a kritériumok között szerepelt, hogy az adott szóalak legyen gyakori a magyar nyelvben (ennek mérésére a *Magyar Nemzeti Szövegtár (MNSZ)* [8] gyakorisági adatait használtuk), illetve, hogy legyen több, használatában gyakorinak tekinthető jelentése. A korpusz szövegeit is az MNSZ-ből, annak Heti Világgazdaság (HVG) számaiból összeállított részkorpuszából válogattuk. Így minden egyes példához rendelkezésre áll a vizsgálat szempontjából releváns kontextus (teljes HVG-cikk), illetve automatikus tokenizálás, szófaji kódolás, szótőre vonatkozó információ.

1 Jelentés-egyértelműsítés

A jelentés-egyértelműsítés (Word Sense Disambiguation, WSD) problémája alatt a szövegekben előforduló többértelműségek (homonímia, illetve poliszémia) feloldásának feladatát értjük. A többértelműség feloldásának problémája egyidős a gépi szövegfeldolgozással, és a legtöbb nyelvtechnológiai alkalmazás (pl. szövegmegértés, ember-gép párbeszéd, gépi fordítás, információ-visszakeresés, illetve -kinyerés) számára fontos köztes feladat.

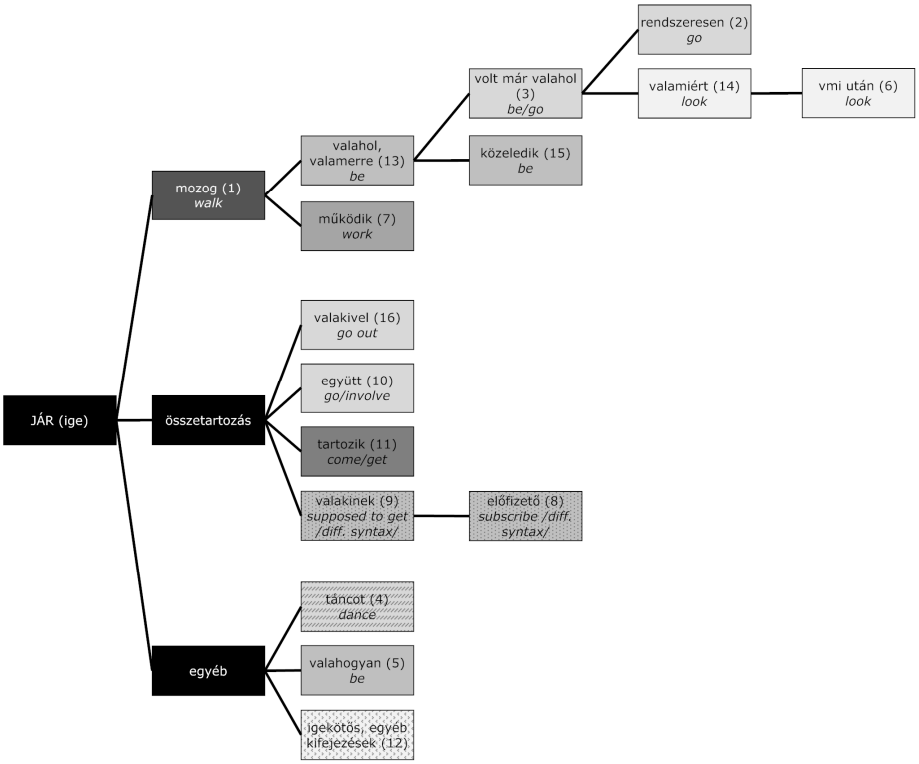
1.1 Kapcsolódó eredmények, áttekintés

Jelentés-egyértelműsítési kutatások más nyelveken
A kezdetben angolra, majd a későbbiekben más nyelvekre folytatott jelentés-egyértelműsítési kutatások nagyrészt kapcsolódtak az ACL-SIGLex által szervezett SensEval [4], [5] workshopokhoz. A 2006-ban megjelent Word Sense Disambiguation [1] című könyv, valamint a SensEval sorozat folytatásaként 2007-ben megrendezett SemEval workshop kiadványa [2] részletes áttekintést ad az eddigi eredményekről.

Jelentés-egyértelműsítési kutatások magyarra
Az angol–magyar, illetve magyar–angol fordítórendszerek fejlesztése kapcsán hosszú ideje foglalkoznak a jelentés-egyértelműsítési feladatokkal magyar nyelven, a fordítórendszer eredményének javítása érdekében [6], [7].

1.2 A jelentés-egyértelműsítési feladat

A jelentés-egyértelműsítő eljárások az alkalmazhatóságuk határai alapján és a jelentésmegkülönböztetés foka szerint két-két főbb csoportra oszthatók. Hatókör tekintetében a teljes szókincsre alkalmazható (all-words WSD) és előre megadott szóalakokon működő (lexical sample WSD) módszereket különböztethetünk meg, míg a jelentésmegkülönböztetés részletessége szerint aprólékos vagy finom (fine grained), illetve durva (coarse grained) szinteket különböztethetünk meg. Az alábbi ábra szemlélteti a *jár* ige jelentéseinek különböző felbontási szintjeit – minden címke egy önálló elkülöníthető jelentés, melyek azonban 3 nagyobb csoportba oszthatók, amiket a színekkel illusztráltunk:



A Magyar WordNet projekt során elkészült – itt ismertetett – korpusz kijelölt szóalakokra részletes felbontású jelentésannotációt tartalmaz (fine grained lexical sample corpus).

1.3 Címkézési elvek

A munka első fázisában a megadott 39 szóalak lehetséges jelentéseit adtuk meg. Az egyes jelentések meghatározásában segítségül hívtuk a Magyar Értelmező Kéziszótár papír és elektronikus változatát, valamint nyelvi intuíciónkat is. Két külön jelentésnek vettük azokat az eseteket, amelyek valamelyik szótári definíció szerint határozottan elkülöníthetők[1].

A nemzetközi gyakorlatnak megfelelően a korpusz példáinak címkézését két független annotátor (képzett nyelvész) is elvégezte. Független alatt azt értjük, hogy a munka elvégzése során tilos volt a nyelvészek közötti kommunikáció. A páros jelölés segítségével egyrészt lehetővé vált az adatbázis konzisztenciaszintjének mérése, másrészt az esetleges jelölési hibák konszenzussal javíthatók.

Fontos kitétel volt, hogy a szóalakokat csak egy adott szófaj keretén belül címkéztük. Pl. a *pont* szónak csakis főnévi jelentéseit címkéztük. Határozószóként - *pontosan* jelentésben – nem címkéztük, azaz poliszémiát vettünk figyelembe, homonímiát nem. Ugyanezen okból nem lett felvéve a *század* szó *századrész* jelentése, mivel ez nem főnév, hanem törtszámnév.

2 A korpusz bemutatása

Ebben a fejezetben ismertetjük az elkészült korpusz főbb technikai, illetve tartalmi jellemzőit.

2.1 A korpusz összetétele

A korpusz építésekor minden szóalakra 350-500 példa címkézését tűztük ki célul, ezzel a mérettel az egyes szóalakokra készülő részkorpuszok méretben összevethetők a más nyelvekre elérhető adatokkal. Az elkészült adatbázisban azonban benne hagytuk azokat a példákat is, melyek végül nem kerültek kézi egyértelműsítésre.
A kiválasztott szóalakok a következők:
melléknév: *anyagi, élő, erős, képes, pontos, szociális*
főnév: *család, élet, ház, helyzet, intézmény, iskola, kép, képviselő, kormány, nap, oldal, ország, perc, pont, program, század, személy, szervezet, tanár, világ, víz*
ige: *függ, hat, jár, kap, kerül, marad, rendelkezik, szerepel, tart, tartozik, tud, válik*

A 39 szóalak átlagos jelentésszáma igen magas (átlag 6 jelentés szóalakonként), melyből a korpusz anyagát képező szövegekben átlagosan csak 5 jelentés jelenik meg. Ha azonban nem vettük számításba az elhanyagolható mértékben (1-2%-ban) jelen levő jelentéseket, akkor az átlagosan megjelenő jelentések száma még kevesebb, 3,7 (mérsékeltebb a többértelműség). Külön érdekes a *tanár* szóalak, mely a vizsgált szövegben teljesen egyértelműnek bizonyult annak ellenére, hogy a szóalakok válo-

[1] A későbbiekben felmerült két újabb kritérium is, miszerint ha az adott jelentéseknek két külön szó feleltethető meg egy másik nyelvben, illetve ha a szóalak más-más vonzatkerettel fordul elő, akkor is külön jelentésnek vesszük fel.

gatásánál követelmény volt, hogy legyen több, a nyelvben gyakran használt jelentésük. Némely szó azonban a homogén nyelvhasználat ellenére is igen sokféle formában megjelenik, mint a *jár* ige, melynek 16 jelentéséből 14 előfordul a szövegben.

2.2 A korpusz formátuma

A korpusz építésekor a *SensEval/SemEval* (Association for Computational Linguistics által szervezett) nemzetközi konferencia workshopokon WSD-feladatokhoz készített korpuszok formátumát követtük. Ezzel a választással egyrészt egy meglévő XML-formalizmust vettünk át, így nem kellett az adatformátumot tervezni, másrészt a szabvány adattárolás remélhetőleg megkönnyíti a korpusz terjesztését is.

Egy példa a korpuszból:

```
<instance id="jár.V.mnsz.01" docsrc="press-hvg.1">
        <answer instance="jár.V.mnsz.01" senseid="jar_v_5_valahogyan"/>
        <context>
        Ez azonban a dolognak már csak a technikája és nem a tartalma volt . Az üzenet , amely az első
        forduló után az akkor még csak kvázigyőztesek szájából megfogalmazódott , és amely a jelek
        szerint a választáson részt vevők többségénél meghallgatásra talált , körülbelül így szólt : az
        MSZP az elmúlt négy évben meghatározó pozícióból folytatott kormányzati politikájával betöl-
        tötte a társadalmi-gazdasági átmenet időszakában neki osztott szerepet .
        Elvégezte azt az egyébként hálátlan feladatot , a stabilizációt , ami nélkül egyetlen kelet-közép-
        európai országnak sincs esélye a felemelkedésre . A polgári átalakulás további vezényletét vi-
        szont az abban érintettek többsége a legalábbis külsőre frissebbnek , dinamikusabbnak tűnő és a
        múlt relikviáitól , a már egyszer eldobott maníroktól nem ( vagy legalábbis kevésbé ) terhes poli-
        tikai erőkre és személyekre kívánja bízni .
        Ezért is tűnik e pillanatban irrelevánsnak annak felvetése , hogy nem <head>jártak</head> vol-
        na -e jobban a szocialisták , ha Horn Gyulát időben katapultálják a pártelnökségből , de leg-
        alábbis a miniszterelnök-jelöltségből , és mondjuk a külügyminiszteri tevékenységével és szemé-
        lyes karakterével , amellett fiatalabb korával a választóközönség számára esetleg vonzóbb Ko-
        vács László vezényletével vágnak bele a választási kampányba .
        Az MSZP a jelek szerint így is a lehetséges maximumot hozta ki magából , legalábbis ami az el-
        ső forduló eredményét illeti .
        </context>
</instance>
```

2.3 A korpusz főbb statisztikai jellemzői

Az alábbi táblázatban foglaltuk össze a korpusz főbb statisztikai jellemzőit. Részletesebb elemzések az interneten érhetők el.

1. Táblázat: A korpusz főbb adatai

	Szóalakok	Duplán annotált példa	Szimplán annotált példa	Annotáció nélkül
Melléknév	6	2087	462	688
Főnév	21	6853	2714	11459
Ige	12	3537	1898	13501
Összesen	39	12477	5074	25648

2.4 Elérhetőség

A korpusz első változata a *Magyar ontológia építése és alkalmazása információkinyerő rendszerekben* [3] projekt keretében készült el. A korpusz – kutatási és oktatási célokra – szabadon hozzáférhető, letölthető a www.inf.u-szeged.hu/hlt oldalról.

3 A korpusz értékelése

Ebben a fejezetben kiértékeljük a korpusz jelentés-annotációinak konzisztenciáját, ismertetjük a két párhuzamos jelölés eltéréseinek egységesítése során követett protokollt, valamint egy egyszerű vektortérmodellen alapuló osztályozó segítségével megmutatjuk, hogy a jelentés-egyértelműsítés kihagyása esetén az alkalmazások által használható leggyakoribb jelentésnél jobb eredmények érhetők el.

3.1 Annotátorok egyetértési rátája (konzisztencia-ellenőrzés)

A korpusz készítésekor első lépésben a megkülönböztetni kívánt jelentések halmazát definiáltuk minden szóalakra, melyeket rövid szöveges leírással (definícióval) láttunk el. A Magyar WordNet állományát is bővítettük a korpuszban használt, a HuWN állományából még hiányzó jelentések synsetjeivel. A nemzetközi gyakorlatnak megfelelően a korpusz példáinak címkézését két képzett nyelvész is elvégezte, egymástól függetlenül.

Az annotátorok egyetértési rátája, azaz a címkézési feladat szakértők általi elvégzésének pontossága azokban az esetekben alacsonyabb, amikor a leggyakoribb jelentés részaránya nem túl magas. Ezekben az esetekben az egyértelműsítés igen nehéz feladat (hisz az egyetértést képzett nyelvészek közt mértük). Másrészt a nyelvtechnológiai alkalmazások, mint a gépi fordítórendszerek, információkinyerő rendszerek stb. éppen ezekben az esetekben profitálnának leginkább egy hatékony WSD-megoldásból, a leggyakoribb jelentés választása helyett. Az annotátori egyetértési ráta a teljes korpuszra nézve 84,78%-os volt.

Az egyik legnehezebb feladat az volt, hogy az annotátor következetességét a szóalak címkézése során végig megőrizze. Ha egy kérdéses esetben az annotátor egy adott jelentés mellett döntött, akkor ugyanazon címkével lássa el a szóalakot egy későbbi, hasonló kontextusban történő előfordulásakor is. Pl. ha X annotátor a BUX tőzsdeindex *pontját* címkézte a *pont_2: az értékelés egysége* jelentéssel első előfordulásakor, akkor ugyanígy kellett eljárnia az összes esetben, még akkor is, ha az egyes esetek egymástól „nagy távolságra” helyezkedtek el[2].

A gyakorlatban a jelentések nem mindig lettek a legprecízebbek, nem mindig tükrözték az elméleti vagy szótári jelentés-megkülönböztetést. Sokszor túl finom, nehezen megkülönböztethető jelentésárnyalatok is fel lettek véve, ezzel tovább romlott az

[2] Az annotátor még önmagán „belül” sem mindig konzisztens, nemhogy másokkal összehasonlítva. Épp ezért bizonyos esetekben következetlenségek előfordulhatnak a korpuszban. Az egyes távoli esetek fejben tartása a kézi annotálás egyik nagy nehézsége.

egyetértési mutató, és az annotátor saját következetessége is csorbát szenvedhetett. A nagyobb következetesség elérése céljából a rendszer tovább finomítható.

Bizonyos esetekben a morfológiai elemző nem megfelelően kategorizálta a szóalakot. Pl. a *vált* jelen idejű igealakot a *válik* múlt idejének elemezte, így felkínálta címkézésre a *válik* jelentései közé. Ebben az esetben a szóalakot nem címkéztük.

A szövegek tematikájából adódóan bizonyos jelentések sokkal gyakrabban fordultak elő a többinél. Mivel a korpusz HVG-szövegekre épül, például a *kormány* szó előfordulásainak túlnyomó többsége a 'politikai kormány' jelentést hordozza. Ha azonban a korpuszban lennének például autókról szóló szövegek, az 'irányító szerkezet' jelentés aránya rögtön megnőne.

Külön problémát jelentettek a kollokációk és a szólások, közmondások, mert sok esetben lehet tudni, hogy a kifejezésen belül milyen jelentésben szerepel az adott szó. Például a *sok víz lefolyik a Dunán addig* szólásban a víz jelentése pontosan azonosítható: *víz_2: a föld felszínének valamely részét borító folyadéktömeg*, kérdés azonban, hogy ezt a címkét kapja-e, vagy pedig *egyéb* címkével lássuk el, mivel kifejezés része.

3.2 A korpusz címkézésének véglegesítése

A javítási munkaszakaszban egy harmadik független annotátor nézte át azokat az eseteket, amikor a két annotátor címkézésében eltérés mutatkozott, és véglegesítette a problémás esetek címkéit. Így a korpusz azon részének címkézése, melyre duplán rendelkezésre állt jelentésannotáció, a lehetőségekhez mérten konzisztens. Azon példákat, ahol csak egy címkézés készült el, nem ellenőríztük.

Az eltérések nagy többsége abból adódott, hogy az annotátorok máshogy értelmeztek bizonyos, egymást részben fedő jelentéseket (ez egyben utalás is arra, hogy e jelentések szétválasztása talán nem teljesen indokolt). Például: *jár_6: valami után jár, megszerzésén fárad* és *jár_14: valamiért több helyre is elmegy*. A legtipikusabb eltérés amikor az egyik annotátor egy adott szóalakot egy adott kontextusban *egyéb* címkével látott el, míg a másik annotátor úgy érezte, hogy a szó adott előfordulása még belefér egy pontosabban meghatározott jelentéstartományba. Például a *kap* igénél gyakori volt, hogy az egyik annotátor a *jogdíjat kap* kifejezésben (és hasonló esetekben) a *kap_1: valamit adnak neki* címkét jelölte meg, míg a másik az *egyéb* címkét választotta.

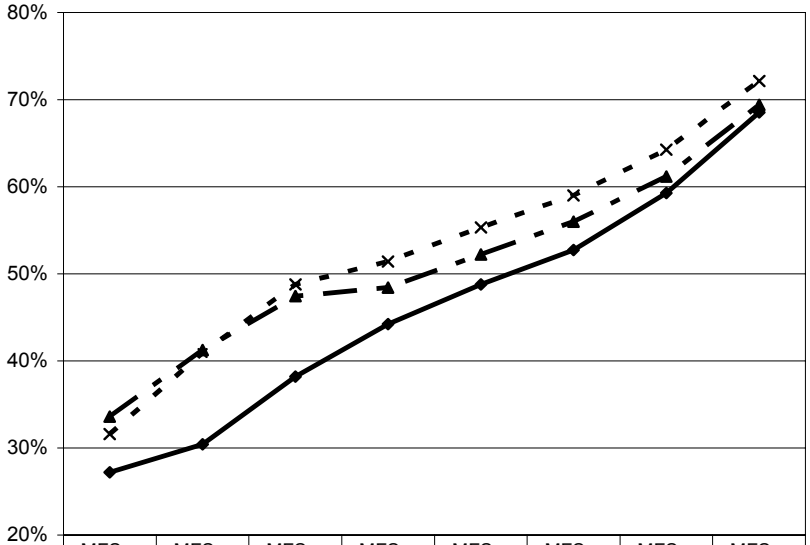
Előfordult néhány olyan eset is, amikor a harmadik annotátor nem értett egyet a két meglévő címkézés egyikével sem. Ilyenkor – a lehetőségekhez mérten – a három annotátor együttes megbeszélése és egyeztetése alapján alakult ki a végleges címke. Egy példa: *a víz felforralása vagy – erre szolgáló filtereken való – átszűrése* kontextusban a két annotátor a *víz_1: élethez nélkülözhetetlen folyadék*, illetve a *víz_2: a föld felszínének valamely részét borító folyadéktömeg* jelentéseket adták meg, a harmadik annotátor viszont a *víz_3: ivóvíz, fürdővíz* jelentésre szavazott. (A végső döntés a *víz_3* címke lett.)

Az eltérések jelentősen kisebb hányada nyilvánvaló tévesztésből fakadt: az egyik annotátor véletlenül a szomszédos címkére kattintott, vagy kifelejtette (azaz jelöletlenül hagyta) az adott szóalakot.

3.3 Baseline mérések, C4.5, naiv Bayes osztályozók

A jelentés-egyértelműsítő eljárások értékelésekor baseline pontosságként a leggyakoribb jelentés részarányát célszerű tekinteni (továbbiakban MFS), hiszen ez a triviálisan elérhető pontosság. Egy eljárás által adott címkézés (egyértelműsített szóelőfordulások) akkor tekinthető értékelhetőnek, ha a leggyakoribb jelentés részarányánál nagyobb hányadban rendeli a szóalakokhoz a megfelelő jelentést.

A felügyelt tanulási modellek építéséhez szükséges a feladat példáinak a tanuló algoritmus számára kezelhető formátumra való konvertálása. Kísérleteink során a példák leírására csak a címkézett szóalak közvetlen környezetét (egyetlen bekezdés) használtuk fel, illetve ábrázolására a jól ismert *vektortérmodell*t használtuk. A jelentés-egyértelműsítési feladat esetében ez a reprezentáció nyilvánvalóan túlzott egyszerűsítés, hiszen a célszó közvetlen környezete, a mondattani szerepek sokszor kiemelten fontosak az egyértelműsítésben, és ez az adat itt elvész. Eredményeinket a korpusz terjesztéséhez mint összehasonlítási alap szántuk.



	MFS < 30%	MFS < 40%	MFS < 50%	MFS < 60%	MFS < 70%	MFS < 80%	MFS < 90%	MFS < 100%
MFS	27,20%	30,42%	38,21%	44,23%	48,80%	52,72%	59,28%	68,55%
BAYES	33,60%	41,23%	47,44%	48,40%	52,21%	55,98%	61,18%	69,40%
C4.5	31,60%	40,91%	48,78%	51,42%	55,31%	59,00%	64,27%	72,14%

1. ábra: Az MFS, C4.5 és naív Bayes osztályozók pontosságai a leggyakoribb jelentés részarányának függvényében.

Tapasztalataink azt mutatják, hogy a statisztikai modellek jobban szerepelnek az alkalmazások által jelenleg használt *leggyakoribb jelentés* heurisztikánál (lásd 1. ábra). A pontosságbeli különbség jelentős a korpuszon a bonyolultabb célszavak esetén, ahol a *leggyakoribb jelentés* jelentősen elmarad az elvi maximális pontosságtól, melyet az annotátorok egyetértési rátájával mértünk.

4 Javítási és finomítási lehetőségek

A nagyobb következetesség elérése céljából a rendszer tovább finomítható az azonos jelentések csoportos átnézésével:

- Végig kell nézni egy adott jelentés címkéjével ellátott összes szóalakot és összevetni őket egymással, hogy valóban következetes-e a címkézés az adott jelentéstartományon belül.
- Ezt minden jelentésnél érdemes elvégezni!
- Ami kilóg a sorból, azt újra kell címkézni!

A jelentésárnyalatok finomságát, a különféle jelentések megkülönböztetését érdemes lehet újragondolni (l. a *jár* példája). Más témájú szövegek annotálása is hasznos lehet a jelentések gyakoriságának meghatározásában.

A teljes szókincsre alkalmazható (all-words) WSD létrehozása nehéz feladat, óriási energiabefektetést igényel, hiszen a teljes magyar szókincsre ki kellene dolgozni a lehetséges jelentéseket. Tovább nehezíti a feladatot, hogy időnként a szókapcsolat együtt hordoz bizonyos jelentést a kontextusban, és a szóalakról önmagában nehéz eldönteni, hogy hordozza-e az adott jelentést – például az *A védőoltások növekedése terén pedig erős képzelőerő kell a tényleges kormányzati cselekvési mező megtalálásához.* mondatban a kontextus egésze hordozza a túlzó, ironikus jelentést, nem kizárólag az *erős* szóalak.

Bibliográfia

1. Agirre, E., Edmonds, P.: Word Sense Disambiguation – Algorithms and Applications, In Ide, N. and Véronis J., editors: Text, Speech and Language Technology Series, Volume 33, Springer, Dordrecht, The Netherlands (2006)
2. Agirre, E., Márquez, L., Wicentowski, R.: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic (2007)
3. Hatvani, Cs., Kuti, J., Miháltz, M., Szarvas, Gy.: GVOP 3.1.1-2004-05-0191/3.0 – Magyar ontológia építése és alkalmazása információkinyerő rendszerekben, projektzáró összefoglaló jelentés, Technical Report, Szeged, Hungary (2007)
4. Kilgariff, A.: Proceedings of Senseval 2: Second International Workshop on the Evaluating Word Sense Disambiguation Systems, Association for Computational Linguistics, Toulouse, France (2001)
5. Mihalcea, R., Edmonds, P.: Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics, Barcelon, Spain (2004)
6. Miháltz, M.: Towards A Hybrid Approach To Word-Sense Disambiguation In Machine Translation. In Proceedings Modern Approaches in Translation Technologies Workshop at RANLP-2005, Borovets, Bulgaria (2005)
7. Miháltz, M., Póhl, G.: Javaslat szemantikailag annotált többnyelvű tanítókorpuszok automatikus előállítására jelentés-egyértelműsítéshez párhuzamos korpuszokból. In Proceedings of III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary (2006)
8. Váradi, T.: Szótár, Korpusz – Magyar Nemzeti Szövegtár. In Gecső, T., editor: Lexikális jelentés, aktuális jelentés. Segédkönyvek a nyelvészet tanulmányozásához IV. Tinta Kiadó, Budapest, Hungary (2000)

Részben felügyelt tanulási módszerek a tulajdonnév felismerésben

Farkas Richárd[1]

[1] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
rfarkas@inf.u-szeged.hu

Kivonat: Az általános gépi tanulás egyik paradigmája a részben felügyelt tanulás az elmúlt években ismét előtérbe került. A módszer célja a jelöletlen adatban rejlő összefüggések kihasználásával javítani a pusztán jelölt adatokat használó tanuló algoritmusokon. Más szemszögből ezen technikák alkalmazásával kevesebb annotált adattal, így kevesebb emberi élőmunka felhasználásával ugyanolyan (vagy közel ugyanolyan) pontosságú modellek építhetőek, mint a nagyobb jelölt adatbázist használó modellekkel. A számítógépes nyelvészet statisztikai megközelítéseiben a megfelelő méretű és minőségű annotált tanítókorpuszok megléte alapfeltétel. Cikkünkkel arra szeretnénk felhívni a figyelmet, hogy a részben felügyelt technikák alkalmazása mellett jelentősen kisebb méretű korpuszok kézi annotálása is elégséges. Ezt az állítást empírikusan is alátámasztjuk magyar és angol tulajdonnév-felismerési korpuszok felhasználásával.

1 Bevezetés

Korábbi munkánkban [5] bemutattuk korpusz alapú tulajdonnév-felismerő modulunkat amelyet három különböző adatbázison (amelyek mind nyelvét, mind témáját tekintve különböztek) értékeltünk ki. Akkori munkákból azt a következtetést vontuk le, hogy ha rendelkezésre áll megfelelő méretű, manuálisan jelölt tanító adatbázis akkor gépi tanulási módszerek felhasználásával igen jó pontosságú automatikus taggelő rendszert lehet építeni.

Azonban minden egyes új tématerület új, kézileg jelölt adatbázis építését követeli meg, ami igen költséges és időigényes feladat. Ezzel szemben az esetek többségében rendelkezésre áll nagy mennyiségű, de címkézetlen szöveg, gondoljunk csak az Internetre. A felügyelt és felügyelet nélküli gépi tanulási paradigmák közt félúton helyezkednek el a részben felügyelt tanulási technikák (semi-supervised learning) [24] amelyeknek a célja a jelöletlen adatban rejlő információ felhasználása a jelölt adatok alapján történő modell-építés folyamán.

Részben felügyelt technikákkal számos általános gépi tanulási feladatban sikerült a felügyelt modellen jelentékenyen javítani (például [14]), valamint nemzetközi szinten megtörtént ezek adaptációja a számítógépes nyelvészeti problémákra [16][18]. Cikkünk elsődleges küldetése, hogy felhívja a figyelmet ezen technikák hasznosságára. Munkánk folyamán magyar illetve angol tulajdonnév korpuszokon teszteltünk néhány, a gépi tanulásban sztenderdnek számító technikát illetve tárgyaljuk a számító-

gépes nyelvészeti problémák sajátosságait kiaknázó technikák lehetőségét, majd ezek közül empirikusan is tesztelünk néhányat.

2 Részben felügyelt tanulási megközelítések

Az alábbiakban ismertetjük a főbb általános részben felügyelt tanulási megközelítéseket, majd feszegetjük azt a kérdést, hogy milyen számítógépes nyelvészet specifikus módszereket lehet érdemes használni.

2.1 Részben felügyelt tanulási technikák

A részben felügyelt tanulásnál rendelkezésre áll egy kis méretű jelölt adathalmaz és egy nagy méretű de jelöletlen adathalmaz. A cél jelöletlen adatokból nyerhető mintázatok hasznosítása a címkézett adatbázison történő modell-építés folyamán, azaz jobb modell építése. Az általános gépi tanulásban használt részben felügyelt technikákat három nagy csoportba sorolhatjuk. Ezt a három megközelítést nagyon röviden bemutatjuk ebben a fejezetben, részletesebb tárgyalás található például [24]-ben.

A legkorábbi megközelítések az ún. **generatív módszerek**kel kapcsolatosan születtek [1]. A generatív modellek (pl. Hidden Markov Models [11]) a címkézés feltételes valószínűségét közvetlenül próbálják meg leírni feltéve az inputot (esetünkben szavakat és jellemzőiket). A 'közvetlenül' azt jelenti, hogy feltételezünk valamilyen eloszlást, ami általában több elemi eloszlás kombinációja (mixture model), és annak paramétereit a tanító adatbázisból becsüljük. A nagy mennyiségű jelöletlen adat segítségével az összetett eloszlás komponensei meghatározhatóak. A generatív modellek közé sorolhatjuk a klaszterezés alapú módszereket is (ahol először a jelöletlen adatot klaszterezzük, az egyes klaszterekhez címkét rendelünk és ezzel bővítjük az eredeti adatbázist) hiszen minden klaszterező algoritmus csak akkor működhet helyesen, ha felfedezi a mintát generáló eloszlást. A generatív modellek gyakorlati használhatóságát éppen az a tény teszi nehézkessé, hogy ismernünk kell a input adat eloszlását [3].

Az ún. **bootstrapping** megközelítések nem feltételezik speciális generatív módszer alkalmazását, esetükben tetszőleges felügyelt gépi tanulási algoritmus alkalmazható. Itt a tanító adatbázist automatikusan címkézett egyedekkel (iteratívan) bővítjük. Az ön-tanulás (self-training) [23] folyamán egy gépi tanulási módszer a címkézett adatbázis alapján épített modell segítségével felcímkézi a jelöletlen adathalmazt, majd a legmegbízhatóbb, automatikusan jelölt adatokkal bővíti a tanító adatbázist és megismétli a modellépítési műveletet. Az együtt-tanulásban (co-training) két (vagy több) osztályozó címkézi fel a jelöletlen halmazt és a megbízható, automatikusan címkézett egyedekkel egymás címkézett adatbázisát bővítik (ezen keresztül „tanítják egymást"). Az osztályozók származhatnak különböző algoritmus osztályokból [7], de használhatjuk ugyanazt a módszert is különböző jellemző-készlettel [13].

A legfiatalabb részben felügyelt tanulási módszerek a **vágás az alacsony sűrűségű területeken** (low density separation) irányelvet követik. Ezeknél a módszereknél a kiértékelő adatbázist is felhasználjuk, mint jelöletlen adat. A cél tulajdonképpen ezeknek az adatoknak a jelölése (transzduktív megközelítés) és nem az új, ismeretlen

példákon predikáló modell fejlesztése (induktív tanulás). A legismertebb ilyen módszer a Transductive Support Vector Machine (TSVM) [19] – az SVM egy kiterjesztése - estében a jelöletlen pontokat felhasználjuk a kernel térbeli maximális margójú vágás megtalálására (elkerüljük a vágással azokat a régiókat ahol a jelöletlen pontok sűrűsége nagy).

A gráf-alapú megközelítések az utóbbi években kerültek előtérbe [2]. Itt a címkézett és címkézetlen egyedeket (beleértve a kiértékelési adatbázist is) egy gráf csúcsainak képzeljük el, ahol két csúcs közti él súlya a két egyed hasonlóságával arányos (a gyakorlatban csak a legközelebbi szomszédokat kötjük össze éllel). Egy megfelelő hasonlósági metrika és a gráf vizsgálatával könnyen számolhatunk lokális sűrűségi értékeket a gráfban. A gráfban ezek után az alacsony sűrűségű helyeken (élek mentén) kell a vágást elvégeznünk. A vágás után kapott klaszterekben a kiértékelendő egyedeket a klaszterben szereplő címkézett csúcsoknak megfelelően jelöljük.

Ezek az alacsony sűrűségű régiókban vágó módszerek elméletileg jól alátámasztottak, azonban a gyakorlatban csak kis adatbázisokra alkalmazhatóak (mind tár, mind időigényük igen magas). Még azok a letölthető megoldások is amelyek magukat nagy adatbázisokon működőnek írják le (large scale solutions) nem adnak megoldást 100 jellemző mellett 30 ezer egyedre egy héten belül[1], pedig a tulajdonnév felismerési problémában 300 jellemzővel és 3 millió egyeddel kell dolgoznunk.

2.2 Részben felügyelt tanulás a számítógépes nyelvészetben

Ha számítógépes nyelvészeti problémákra fókuszálunk lehetőség nyílik azok specifikumainak kiaknázására. Úgy gondoljuk, hogy ez a terület még nincs kielégítően körüljárva. Itt mindössze a számítógépes nyelvészeti problémák két speciális tulajdonságát tárgyaljuk röviden, melyek kiaknázásas nem lehetséges a sztenderd részben felügyelt technikákkal.

A természetes nyelv szekvenciális tulajdonsága lehetővé teszi, hogy összetettebb statisztikákat (szabályosságokat) fedezzünk fel a jelöletlen szövegekben. Ilyen statisztikák a szó és karakter n-grammok, szógyakoriságok (ahol megkülönbözethetünk kis és nagy kezdőbetűs vagy mondat eleji előfordulásokat is [8]) valamint a nyelvmodellek, amibe beleértünk minden olyan modellt ami a nyelv szabályosságait valószínűségi alapon próbálja meg modellezni (tehát nem csak a szűk értelemben vett $P(w_t|w_{t-1})$ feltételes valószínűséggel leírható nyelvi modellt). Az ilyen jellegű statisztikákat felhasználhatjuk a felügyelt tanulási modell jellemző-terének konstrukciójának folyamán.

Egy másik érdekes tulajdonsága az emberi szövegekkel kapcsolatos problémáknak, hogy az Interneten szinte korlátlan mennyiségben fordul elő folyó szöveges információ. A World Wide Web-nek, mint jelöletlen korpuszt azonban nem kezelhetjük ugyanúgy, mint az egyéb offline korpuszokat (nem tudjuk például egy szó összes előfordulásán végigiterálni), azt csak a kereső-motorok (pl. Google, Yahoo) segítsé-

[1] Két ilyen programcsomagot töltöttünk le és teszteltünk:
http://www.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html és
http://www.learning-from-data.com/te-ming/semil.htm

gével tehetjük meg effektíven. Ehhez meg kell fogalmaznunk kéréseket (query) majd a találatul kapott oldalakat letölthetjük (és feldolgozhatjuk), de a találatok becsült száma és a keresőszavak alapján relevánsnak vélt szövegkörnyezetek (snippet) is hordoznak igen hasznos információkat. Létezik már néhány megoldás egyszerű számítógépes nyelvészeti problémára melyek hasznosítják a WWW-et [15][18]. A tulajdonnév-felismerés problémájához legközelebb álló ilyen megoldások egy bizonyos név-osztályba tartozó listákat próbálnak az Internetről automatikusan összegyűjteni (ilyen például a Google Sets szolgáltatás vagy [4][17]). Úgy gondoljuk , hogy a jövőben ez a terület jóval nagyobb figyelmet fog kapni és egyre mélyebb elemzést végző alkalmazások is inputként fogják kihasználni az WWW-et.

3 Empirikus eredmények

Az előző részben bemutatott technikák közül az ön-tanulás, az együtt-tanulás és néhány Web alapú módszert magyar és angol tulajdonnév-felismerési adatbázisokon teszteltük. A kísérletek paramétereit, valamint az elért eredményeket mutatjuk be ebben a fejezetben.

3.1 Magyar és angol tulajdonnév-felismerési adatbázisok

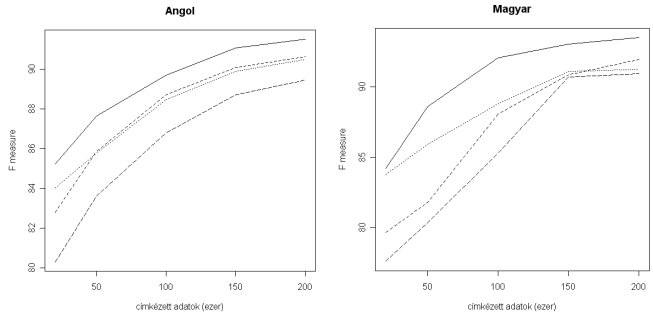
A tulajdonnevek azonosítása (és kategorizálása) folyó szövegben meghatározó fontosságú számos számítógépes nyelvfeldolgozó alkalmazás során. Példaként tekinthetjük a különböző információkinyerő rendszereket, ahol a tulajdonnevek általában jelentős, információt hordozó szerepet töltenek be a szövegben, vagy a gépi fordítási alkalmazásokat, ahol értelemszerűen más módon kell kezelni emberek, szervezetek neveit, mint a szöveg többi részét. Magyar és angol nyelvre a gazdasági témájú tulajdonnév-felismerési feladaton teszteltük a részben felügyelt tanulási modelleket. Itt a feladat egy szöveg minden egyes szavához a *szervezet/helység/személy/egyéb/nem-tulajdonnév* címkék valamelyikének hozzárendelése.

A CoNLL által kiírt nyílt versenynek 2003-ban [22] volt feladata ez a típusú klasszifikáció. Az adatbázis Reuters híreket[2] tartalmazott 1996-ból, amelyek felöleltek sport, politikai és gazdasági témákat egyaránt. Az akkori verseny szervezői azt szerették volna elérni, hogy a rendszerek hasznosítsák a jelöletlen adatból nyerhető információt is, ezért a címkézett adatbázisok mellé és egy közel 18 millió szavas jelöletlen korpuszt (szintén Reuters hírek) is elérhetővé tettek. Ezt a jelöletlen adathalmazt használtuk fel mi is kísérleteink folyamán. A 2003-as versenyre nem küldtek be olyan rendszert ami a címkézetlen adatokat hasznosította volna. A CoNLL testB adatbázisa hordoz néhány olyan speciális tulajdonságot ami a tanító adatbázistól megkülönbözteti, ezért ebben a munkánkban a testA halmazra közlünk eredményeket.

Magyar nyelvre a Szeged Korpusz 200 ezer szóból álló, gazdasági rövidhíreket tartalmazó szegmensét (SzegedNE korpusz) [21] használtuk, mint címkézett és kiértékelési adatbázis. Jelöletlen adatbázisnak gazdasági témájú újsághíreket (nem rövid-

[2] http://www.reuters.com/researchandstandards/

híreket!) próbáltunk meg alkalmazni, azonban ez nem vezetett eredményre. A magyar adatbázisra közölt eredményeink a transzduktív megközelítést követték, azaz a kiértékelési adatbázist használtuk fel, mint címkézetlen korpusz.



1. ábra: Felügyelt tanulási modell által elért eredmények különböző jellemzőterek használata mellett a címkézett korpusz méretének függvényében

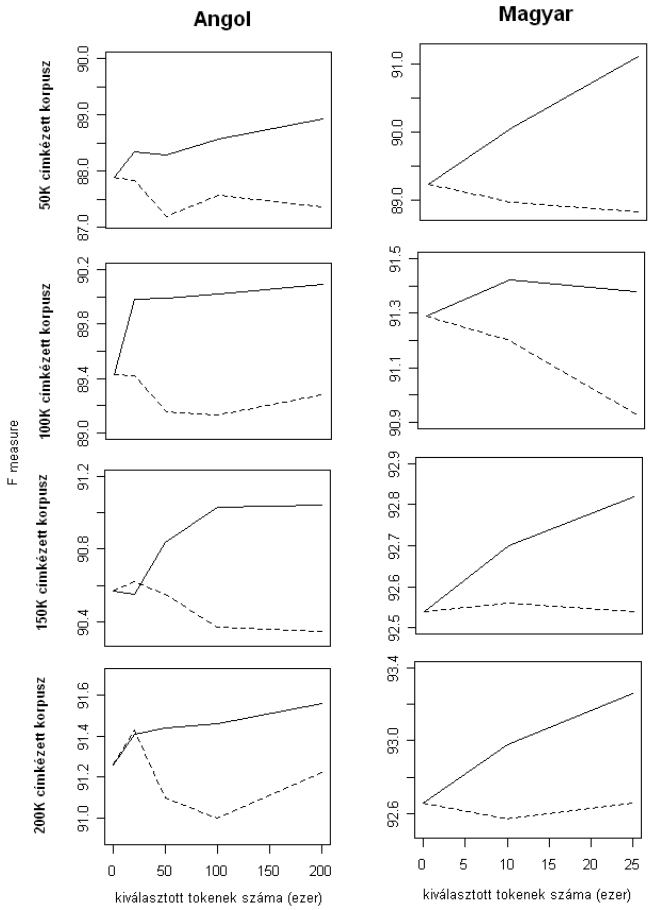
A modellépítés folyamán felhasznált jellemzőkészlet igen változatos volt [20]. A következő kategóriákba sorolhatjuk a jellemzőket (a magyar és angol adatbázison ugyanazokat a jellemzőket használtuk fel):

- Felszíni jellemzők: kis/nagy kezdőbetű, szóhossz, tartalmaz-e számot, van-e nagybetű a szó belsejében, arab/római szám-e stb., illetve legyűjtöttük a tanuló halmaz legjellegzetesebb két-, hárombetűs szórészleteit.
- Frekvenciainformációk: token előfordulási gyakorisága, kis- és nagybetűs előfordulások aránya, mondat eleji előfordulások és nagybetűs előfordulások aránya.
- Környezeti jellemzők: mondatbeli pozíció, megelőző szavakra modell által javasolt tulajdonnévi címke (online kiértékelés), zárójelben, idézőjelek közt van-e; a tanító halmazból legyűjtöttük, hogy a megelőző/rákövetkező szavakból melyek azok, amelyek az egyes osztályokat implikálhatják.
- Egyértelmű tulajdonnevek listája: Felvettük egy-egy listába azokat a szavakat és többszavas kifejezéseket, amelyek a tanító halmazon legalább ötször előfordultak, és az esetek legalább 90 százalékában ugyanabba az osztályba tartoztak.
- Tulajdonnév szótárak: magyar és angol keresztnevek, vállalattípusok (mint pl. kft., rt.), nagyvárosok és országok, stb. Összesen nyolc angol és négy magyar listát alkalmaztunk.

3.2 A jelöletlen korpuszból származtatott jellemzők hozzáadott értéke

A kísérletekben elsősorban a Conditional Random Fields (CRF) [10] nevű osztályozó algoritmusra (MALLET implementáció [12]) támaszkodtunk, ami számos szekvenciális jelölési probléma megoldásában bizonyított és az utóbbi néhány évben a state-of-the-art-nak számít. Első kísérleteinkben a címkézett adatbázis méretének hatásait és a jelöletlen korpuszokból származtatott jellemzők hozzáadott értékét vizsgáltuk meg.

Ilyen címkézetlen korpuszból származó jellemzőcsoport a frekvenciainformációk és a szótárak. Előbbit több milliárd szavas Webkorpuszokból számítják ki. Angolra a Gigaword korpuszt, míg magyarra a Szószablya Gyakorisági Szótárat [8] használtuk. A tulajdonnév szótárak szintén az Internetről összegyűjthető listák. Egyes kategóriákhoz (tulajdonnév osztályokhoz) tartozó listákat gyűjthetünk automatikusan, keresőmo-



2. ábra: Együtt-tanulás (folytonos) és ön-tanulás (szaggatott) a jelölt és jelöletlen adatbázis méretének függvényében

torok és egyszerű szintaktikai keretek illesztésével [4], de a legalapvetőbb listák összeállítva elérhetőek, azokat legfeljebb csak szűrni és normalizálni kell. Az itt használt listákat az utóbbi módon gyűjtöttük, körülbelül egy embernapnyi ráfordítással.

Az 1. ábrán látható 4-4 görbe a teljes jellemzőtér (folytonos vonal), a frekvencia típusú jellemzők (pontozott vonal), a szótárak (szaggatott vonal) illetve mindkét jellemzőcsoport mellőzésével („teli-üres” vonal) nyert eredményeket mutatják a címké-

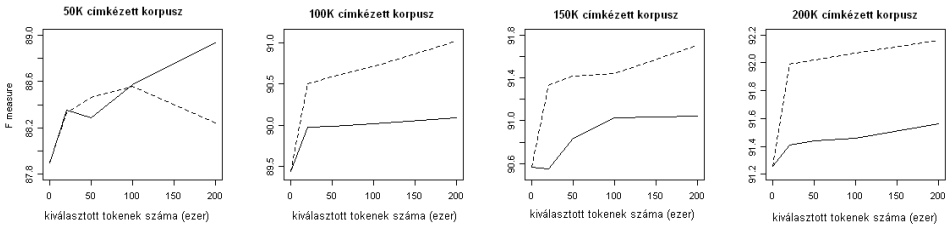
zett tanító adatbázis méretének függvényében. Az a tendencia egyértelműen megfigyelhető, hogy a szótárak megvonását egyre kevésbé érzi meg a modell ha nő a tanító adatbázis. Tehát a szótárak kis méretű adatbázisnál rendelkeznek komoly hozzáadott értékkel, nagyobb halmazokon ugyanezen információt meg tudják szerezni a statisztikai modellek a címkézet adatból is. A frekvencia jellemzők alkalmazásával átlagosan a hibák 19% eliminálhatóak, míg a szótárakkal 15%, együttes alkalmazásukkal 28%.

3.3 Bootstrapping módszerek

Az ön-tanulási és az együtt-tanulási algoritmusokat szimultán vizsgáltuk. Az öntanulásánál a CRF önmagát tanította a jelöletlen adatokkal, míg az együtt-tanulásnál egyetlen iterációban a korábbi munkáinkban alkalmazott és jó eredményeket elérő AdaBoostM1+C4.5 modellünk [20] által jelölt szövegekkel bővítettük a CRF tanító adatbázisát. Az alkalmazott két osztályozási modell teljesen másképp közelíti meg a címkézési problémát. Egyrészt a döntési fa alapú módszer az egyes szavakat egymástól függetlennek tekinti (a környezetre vonatkozó információk a jellemzőtérbe vannak beépítve), míg a CRF az egész mondatot (szekvenciát) egyben jelöli be. Másrészt a CRF az egyes jellemzők együttes eloszlása felett épít exponenciális modellt (logisztikus regresszió), míg a C4.5 algoritmus mohó módon választ minden lépésben egy jellemzőt információ elméleti metrikák alapján. Ezt a diverzitást tudja együtttanulás kihasználni.

A 2. ábrán szaggatott vonallal jelöltük az ön-tanulással, míg folyamatos vonallal az együtt-tanulás által elért eredményeket különböző méretű címkézett adatbázis mellett a jelöletlen adatbázis méretének függvényében (az x tengely 0 pontjánál lévő függvényértékek megegyeznek az 1. ábra értékeivel). Láthatjuk, hogy az együtt-tanulással minden esetben tudtunk javítani a felügyelt modellhez képest, míg az ön-tanulással nyert automatikusan címkézett példák csak összezavarták a modell-építést.

Az eredményeket javíthatjuk, ha nem minden automatikusan címkézett mondatot adunk hozzá a tanító adatbázishoz, hanem azok közül csak a legmegbízhatóbbakat. A 3. ábrán láthatjuk az együtt-tanulással elért eredményeket miután a döntési fa által jelölt mondatok közül csak azokat használjuk fel amelyek bizonytalansága kisebb, mint 10^{-3} illetve 10^{-10}. Természetesen minél alacsonyabb ez a küszöbérték annál több jelöletlen adat felhasználására van szükség, hogy szignifikáns módon bővíthessük a címkézett adatbázist. A 10^{-10} küszöb mellett 3 millió szövegszónyi nyers szöveget használtunk fel, 23407 „megbízható" mondatot kiválasztva. Együtt-tanulással sikerült hajszállal jobb eredményeket elérnünk 100 ezer jelölt adat felhasználásával (91,28% F érték), mint 200 ezer jelölt adattal jelöletlen adatok nélkül (91,26% F érték).



3. ábra: Performancia a jelölt és jelöletlen adatbázisok méretének függvényében különböző kiválasztási küszöbértékek mellett

A szép eredmények ellenére el kell mondanunk, hogy magyarra is hajtottunk végre jelöletlen gazdasági szövegek felhasználásával (a transzduktív megközelítés helyett) ön- illetve együtt-tanulást. Azonban ezzel – a CoNLL B adatbázisához hasonlóan – nem sikerült szignifikáns javítást elérnünk a felügyelt modellhez viszonyítva. Ez minden bizonnyal a jelöletlen adathalmaz és a kiértékelési adatbázis eltérő jellegéből fakad (hasonló következtetést vont le [9] is).

3.4 A WWW hasznosítása mint külső szakértői tudásbázis

Számos lehetőség kínálkozik arra, hogy az Internetről információt gyűjtsünk számítógépes nyelvészeti problémák megoldásához. [6] munkánkban három heurisztikát mutattunk be melyekkel az angol tulajdonnév-felismerő rendszerünk hibáit próbáltuk meg eliminálni a GoogleAPI és a Wikipedia felhasználásával.

A tulajdonnév-felismerő rendszerek hibáinak egy szignifikáns része abból fakad, hogy a rendszer nem jól találja meg a hosszabb frázisok határát (elejét vagy végét). Ezért megvizsgáltunk minden olyan egyedet ahol a címkézett frázis előtt (vagy után) közvetlenül nagybetűs szó állt vagy legfeljebb két stopword ékelődött nagy kezdőbetű szó és a jelölt egyed közé. A hipotézisünk az volt, hogy ha az ilyen módon kiterjesztett tulajdonnév előfordulási gyakorisága összemérhető még az eredeti jelöltével akkor a kiterjesztés végrehajtandó. Ennek eldöntésére Google keresést hajtottunk végre a címkézett tulajdonnévre és a kiterjesztett frázisra és ha a találatok számának aránya 0,1%[3] felett volt elfogadtuk a kiterjesztést.

A második heurisztikát arra a hipotézisre építettük, hogy a tulajdonnevek leggyakoribb jelentése (osztálya) statisztikailag hasznos információ. Ezért ha a rendszer nem tudott megbízható döntést hozni egy felismert tulajdonnév osztályáról akkor az Interneten megkerestük annak leggyakoribb szerepét és azt adtuk a frázis címkéjének. Módszerünk néhány kérdést küldött minden tulajdonnévhez (1. táblázat), hogy annak kategóriáját megtudjuk. Ezeket a Google snippetjeinek elemzéséből (főnévi csoportok azonosítása) nyertük ki.

1. táblázat: Felhasznált kereső-kifejezések

Angol	Magyar
NP such as NE	NE egyike NP
NP including NE	NE és más NP
NP especially NE	NE és egyéb NP
NE is a NP	NE vagy más NP
NE is the NP	NE vagy egyéb NP
NE and other NP	
NE or other NP	

[3] Ezt az értéket nem a kiértékelési adatbázis, hanem a fejlesztési adatbázis alapján határoztuk meg, az előbbi így ismeretlen maradt.

Azt hogy melyik kategória melyik tulajdonnév osztályba tartozik a tanító adatbázison egyértelmű egyedekre futtatott ugyanezen Google keresések eredményéből nyertük ki. Azt, hogy a címkézés megbízható-e különböző algoritmusok egyetértési rátájával mértük (comitee based learning).

A harmadik heurisztikánál az egymást követő, azonos típusú tulajdonnevek (pl. „Taleban | Míg-19") problémáját igyekeztünk orvosolni (ilyen esetben a tulajdonnevek határát B- kezdetű címke jelöli). A legtöbb ilyen esetben valamilyen írásjel választja el az egyedeket azonban például beszédfelismerés eredményeként előálló szövegben ezek nincsenek jelen. Az ilyen esetek kiszűrésére a Wikipediát alkalmaztuk: minden legalább kettő hosszúságú frázisra megnéztük, hogy létezik-e a nevet egy az egyben lefedő Wikipedia oldal, és ha létezett írásjelek nélkül elfogadtuk azt egy tulajdonnévnek. Ellenkező esetben a jelölt frázis darabjaira kerestünk, illetve elvégeztük az írásjelek menti vágást. Ily módon sikerült például elválasztanunk a „Golan Heights | Isreal"-t.

A [6] publikáció empirikus eredményei bizonyítják a WWW-ről, mint külső tudásbázisból, inputként gyűjtött információ felhasználásának hasznát. Azonban ezek magyarra történő adaptálásánál problémákba ütköztünk. Először is nem tudtunk minden kérdést lefordítani (a létigét a magyarban nem tesszük egyes szám harmadik személyben), újakat kellett kigondolnunk. De az igazán komoly gondot az okozta, hogy a kérdések mintegy 70%-ára nem érkezett Google találat vagy nem létezett Wikipedia oldal. A magyar Web (a *site:.hu* kifejezést használtuk) és a magyar Wikipedia (aminek mérete az angolhoz viszonyítva 3,5%) nem elég nagy ilyen jellegű feldolgozáshoz.

4. Összegzés

Ennek a cikknek az elsődleges küldetése az volt, hogy rávilágítson a jelöletlen korpuszok felhasználásában (részben felügyelt tanulási modellek) rejlő potenciálra. Eredményeket magyar és angol tulajdonnév-felismerési problémákra közöltünk. A különböző elméleti bázissal rendelkező felügyelt tanulók együttes-tanulásával sikerült 100 ezer szónyi címkézet szöveg és 3 millió szónyi jelöletlen korpusz felhasználásával ugyanolyan eredményeket elérnünk, mint 200 ezer szövegszónyi címkézett adatbázissal. De azt is bemutattuk, hogy a sztenderd részben felügyelt tanulási technikák vagy nem alkalmazhatóak (alacsony sűrűségnél vágás) a nagy méretű problémákra (ami általában a helyzet a számítógépes nyelvészetben) vagy nagyon körültekintő szövegválasztást igényelnek (jelöletlen adat és a kiértékelő adatbázis jellegében meg kell, hogy egyezzen). Ezért javasoljuk speciálisan nyelvtechnológiai problémákban alkalmazható módszerek alkalmazását.

Külön megvizsgáltuk azoknak a jellemzőknek a hozzáadott értékét melyeket jelöletlen korpuszokból származtattunk. Ezek alkalmazása átlagosan mintegy 28% relatív hibacsökkenést vontak maguk után. Végül három, WWW-en alapuló heurisztikát ismertettünk, melyekkel bizonyítottuk, hogy – annak ellenére, hogy a Web-en sokszor találkozhatunk elírással, valótlan információval, azaz zajjal – számítógépes nyelvészeti problémák megoldása során igen hasznos segítség lehet, a világ legnagyobb jelöletlen korpusza a WWW.

A jövőben szeretnénk a specifikusan számítógépes nyelvészeti problémák megoldására testreszabott részben felügyelt tanulási technikákat tovább vizsgálni, elsősorban olyan megoldásokat megcélozva, amelyek a WWW-et, mint külső szakértői tudást effektívebben és sokkal általánosabban tudják felhasználni nyelvtechnológiai problémák megoldása közben.

Bibliográfia

1. Baluja S.:Probabilistic modeling for face orientation discrimination. Learning from labeled and unlabeled data. In Neural Information Processing Systems (1998)
2. Chapelle, Olivier; Alexander Zien: Semi-Supervised Classification by Low Density Separation. In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, 57-64 (2005)
3. Cozman, F.; I. Cohen; M. Cirelo:Semi-supervised learning of mixture models. ICML-03, 20th International Conference on Machine Learning. (2003)
4. Etzioni, Oren; Michael Cafarella; Doug Downey; Ana-Maria Popescu; Tal Shaked; Stephen Soderland; Daniel S.Weld; Alexander Yates.Unsupervised named-entity extraction from the web: an experimental study. In Artificial Intelligence Volume 165, Issue 1, 91-134 (2005)
5. Farkas Richárd, Szarvas György: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. Magyar Számítógépes Nyelvészeti Konferencia (2006)
6. Richárd Farkas, György Szarvas and Róbert Ormándi: Improving a State-of-the-Art Named Entity Recognition System Using the World Wide Web. Lecture Notes on Computer Sciences Vol. 4597. pp 163-172 (2007)
7. Goldman, S.; Y. Zhou: Enhancing supervised learning with unlabeled data. In Proceedings 17th International Confrence on Machine Learning. San Francisco, CA: Morgan Kaufmann (2000)
8. Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón: A Szószablya project. I. Magyar Számítógépes Nyelvészeti Konferencia (2003)
9. Ji, Heng; Ralph Grishman: Data Selection in Semi-supervised Learning for Name Tagging In Proceedings of ACL'06 Workshop on Information Extraction Beyond Document, Sydney (2006)
10. Lafferty, John; Andrew McCallum; Fernando Pereira: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of 18th International Conference on Machine Learning (2001)
11. Manning, Chris; Hinrich Schütze: Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. (1999)
12. McCallum, Andrew K: MALLET: A Machine Learning for Language Toolkit url: http://mallet.cs.umass.edu (2002)
13. Mitchell, T: The role of unlabeled data in supervised learning. In Proceedings of the Sixth International Colloquium on Cognitive Science. San Sebastian, Spain (1999)
14. Ke Lu, Jidong Zhao, Deng Cai: An algorithm for semi-supervised learning in image retrieval. Pattern Recognition, Vol. 39, No. 4. pp. 717-720 (2006)
15. Pasca, Marius; Dekang Lin; Jeffrey Bigham; Andrei Lifchits; Alpa Jain: Organizing and Searching the World Wide Web of Facts - Step One: the One-Million Fact Extraction Challenge. In Proceedings of American Association for Artificial Intelligence (2006)
16. L Rigutini, M Maggini: A semi-supervised document clustering algorithm based on EM. Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference pp. 200-206 (2005)

17. Shinzato, Keiji; Satoshi Sekine; Naoki Yoshinaga; Kentaro Torisawa: Constructing Dictionaries for Named Entity Recognition on Specific Domains from the Web. In Proceedings of ISWC'06 Workshop on Web Content Mining with Human Language Technologies (2006)
18. Sumita, Eiichiro; Fumiaki Sugaya: Using the Web to Disambiguate Acronyms. In Proceedings of NAACL '06 (2006)
19. Vapnik, V: Statistical learning theory. Springer (1998)
20. Szarvas György; Richárd, Farkas; András, Kocsor:. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In Lecture Notes on Artificial Intelligence Vol. 4265, 267-278 (2006)
21. Szarvas György; Richárd, Farkas; László, Felföldi; András, Kocsor; János, Csirik:. A highly accurate Named Entity corpus for Hungarian. In Proceedings of International Conference on Language Resources and Evaluation (2006)
22. Tjong, Erik F.; Kim Sang; Fien De Meulder: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of CoNLL-2003 (2006)
23. Yarowsky, D: Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 189–196 (1995)
24. Zhu, Xiaojin:Semi-Supervised Learning Literature Survey. Computer Sciences, University of Wisconsin-Madison, #1530 (2005)

VI. Fordítás és korpusz

A MetaMorpho projekt 2007-ben – a sorozat vége

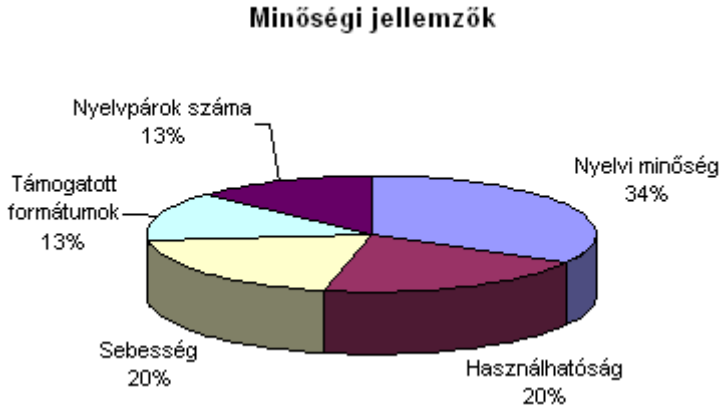
Tihanyi László

MorphoLogic
1126 Budapest Orbánhegyi út 5
tihanyi@morphologic.hu

Kivonat: Ez a fordítóprogram fejlesztéséről szóló sorozat utolsó előadása. A 2000-ben indult fejlesztésről az első MSZNY konferencián, 2003-ban számoltam be először. Az elért eredményeket az idén a programok minőségi jellemzőinek ismertetésével foglalom össze.

1. A fordítóprogramok nyelvi minősége

A MetaMorpho rendszert a fordítóprogramok minőségi jellemzőin keresztül vizsgáltam meg. A minőséget meghatározó szempontok között a nyelvi minőség a legfontosabb, de a működő rendszereknek további igényeknek is meg kell felelnie, amelyeket megfelelő súlyozással [5] szokás figyelembe venni.



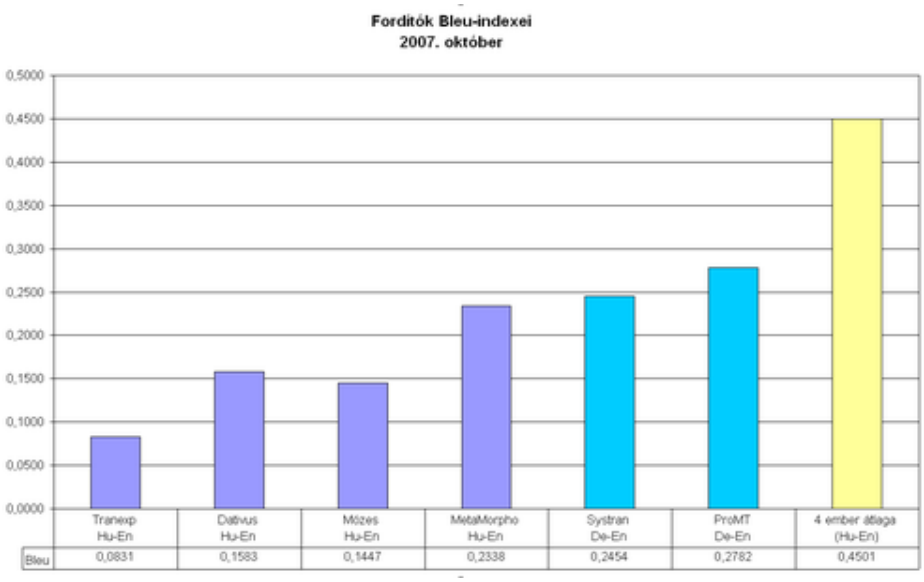
A fordítóprogramok minőségét és annak változását leggyakrabban a Bleu-értékkel mérik. A mérési eljárás lényege röviden, hogy a forrásszövegről több emberi fordítóval referenciafordítást készítenek, majd a gépi fordítást ezekkel a referenciákkal hasonlítják össze. A Bleu-index a gépi fordításban lévő olyan 1-től 4-ig terjedő hosszúságú szósorozatok mértani közepe, amelyek megtalálhatók valamelyik referenciában.

Az összehasonlíthatóság érdekében a mérésekhez mi is a legtöbbek által használt NIST implementációt használtuk, és háromreferenciás méréseket végeztünk. A fordítás minőségének vizsgálatát a magyar–angol anyagon januárban kezdtük, és a mérést minden fejlesztés után elvégeztük. A Bleu-indexet elsősorban ellenőrzési céllal készítjük, de ezúttal háromféle összehasonlító mérést is elvégeztünk.

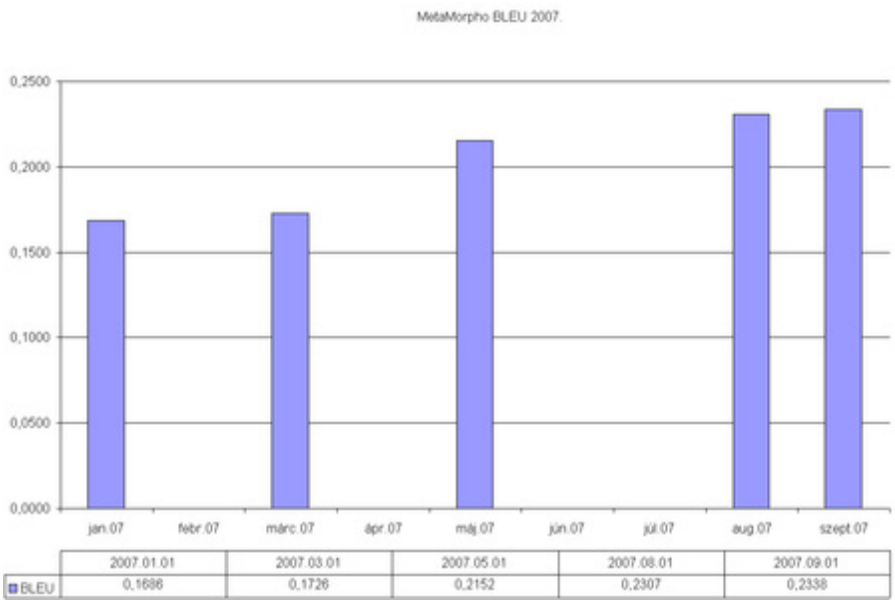
Elsőként összevetettük eredményeinket emberi fordítók teljesítményével. Ez gyakorlatban az adott szövegen elérhető maximális Bleu-érték meghatározását jelenti. A méréshez készítettünk egy negyedik referenciafordítást, majd mind a négy emberi fordítás Bleu-indexét, a másik hármat referenciaként használva, meghatároztuk. A négy fordítás számtani középértéke 0,45-re adódott. Az emberi fordítások Bleu-indexének eltérése igen alacsony volt, 0,021-es szórással.

Ugyancsak újdonság volt, hogy a magyar–angol Bleu-indexet más nyelvű fordítók indexeivel is összevetettük. A kísérlethez a magyar forrásszöveget németre fordítottuk, majd ezt a német forrást különféle programokkal angolra. Azonos tartalmú szövegek, azonos nyelvű fordításait összemérhetőnek gondoljuk. A vizsgálatba a legismertebb Systrant, és a legjobb minőséget adó @promt fordítókat vontuk be. Fordítóprogramunk megközelítette a Systran színvonalát (MetaMorpho: 0,2338, Systran 0,2454). Fontos észrevenni, hogy a magyar–angol fordítót két germán nyelv közötti fordító teljesítményével vetettük egybe.

Megmértük más magyar-angol fordítók teljesítményét is: Tranexp: 0,0831, Dativus: 01583. Az eredmények alapján megállapítottuk, hogy rendszerünk jobb, és gyorsabban is fejlődik, mint a többiek. Egészen friss a Moses statisztikai rendszerrel készített magyar-angol gépi fordító első eredménye: 0,1447. A program a Hunglish korpusz irodalmi részén lett betanítva.



A következő ábrán a magyar-angol MetaMorpho modul ez évi minőségjavulása látható:



2. A fordítóprogramok használhatósága

A használhatósággal kapcsolatban két fontos kérdésre kell válaszolnunk: kik és mire tudják használni a fordítóprogramot? A mai fejlettségi szinten a program a megértés segítésére szolgál, és elsősorban az anyanyelvre történő fordításkor használható.

Az egyéni, céges és internetes felhasználói köröknek különféle programtermékeket készítettünk. Mindamellett különféle megoldásokra van szükség az egyes alkalmazásokhoz és fájltípusokhoz, az internet böngészéséhez, a dokumentum és egyéb szövegtípusokhoz. Ezeknek kiszolgálására az alábbi termékszerkezetet alakítottuk ki:

	weblap	**dokumentum**	**egyéb szöveg**
egyének	MorphoWeb MoBiCAT	MorphoWord MoBiCAT	-
egyének, csomag		MorphoWord Plus	
Cégek		MorphoWord Pro	
Internet	webforditas/web	-	webforditas/szöveg

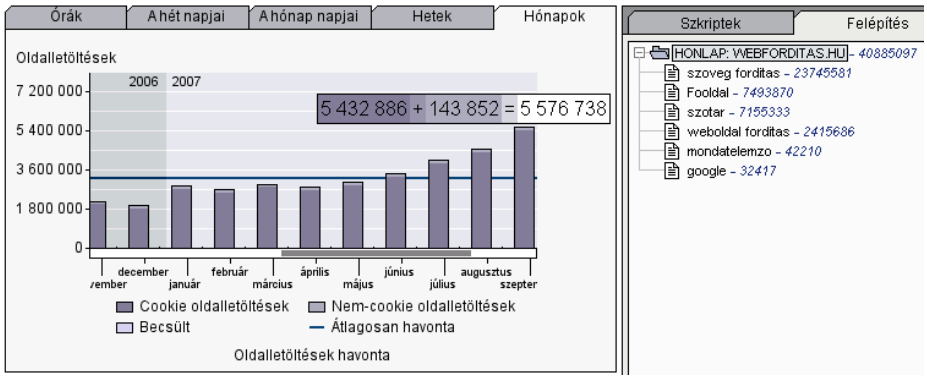
A cégeknek szánt **MorphoWord Pro** változat egy kliens/szerver megoldás, az egyfelhasználós kombinált változat neve: **MorphoWord Plus**. A céges változathoz egy terminológiakivonatoló programot is kifejlesztettünk. A **MoBiCAT** fordító kiegészítésként szolgál azok számára, akik csak egy-egy mondat fordítását szeretnék megkapni, de a szöveget eredeti formájában kívánják olvasni.

A program használhatóságának legfőbb bizonyítéka a használat. A különböző termékek értékesítési adatait nem közölhetjük, de a webforditas.hu oldal látogatottsági értékei is érdekesek. A webforditas.hu ingyenes szolgáltatás éppen a cikk írásakor (2007. október 22.) lett egyéves, így teljes évi adatokkal szolgálhatunk.

A webforditas.hu weboldalról egy év alatt összesen mintegy 40 millió oldalt töltöttek le. 23 millió rövidebb szöveg-, és 2,4 millió weboldal-fordítási kérést szolgáltunk ki. A szótárból további 7,1 millió szó jelentését kérdezték le.

Ez a forgalom szövegmennyiségben is kifejezhető. Egy év alatt 6,3 GB szöveget, azaz 6,3 milliárd karaktert fordítottak le. Ez 1800 karakteres oldalmérettel számolva 3,5 millió oldalnak felel meg, ami többszöröse a teljes hazai emberi fordítás mennyiségének, amelyet az irodák forgalmából néhány százezer oldalra becsülhetünk.

A webforditas.hu 2007. szeptemberében több mint 220 000 látogatót szolgált ki, akik 720 000 látogatást tettek. Az induló forgalom az év végére megháromszorozódott. A napi látogatók száma ez év októberében 20 000 fölé nőtt, így a www.webforditas.hu az ötven leglátogatottabb magyar weboldal közé került.



3. A fordítóprogramok sebessége

A fordítóprogramok sebességi és minőségi szempontjai ellentétesek. A nyelvi elemzésre annyi időt tervezhetünk, amennyit az elfogadható sebességi érték megenged. Az egyre jobb nyelvészeti algoritmusok és egyre gyorsabb számítógépek a közeljövőben újabb minőségjavító eljárásokat tesznek majd lehetővé.

A sebességi adatok egyébként a számítógép memóriaméretétől, a processzor sebességétől és terheltségétől, valamint a mondatok bonyolultságától függően jelentősen eltérhetnek, ezért az adatok csak tájékoztató jellegűek: angol–magyar irányban 400 karakter/s, magyar–angol irányban 250 karakter/s. A számítógép paraméterei: P4 2,8 GHz, 1 GB RAM.

4. Nyelvpárok

A szakmai és laikus vélemények egybehangzóan állítják, hogy egy fordítóprogram annál jobb, minél több nyelvet ismer. Mi szakítottunk ezzel a véleménnyel, és kizárólag az angol-magyar nyelvpárral foglalkozunk. Az alábbiakban ismertetem, miért tesszük ezt, és hogyan gondolom megoldani a soknyelvűségi problémát.

Mindnyájan ismerjük a nyelvi poligont. Ha *N* nyelv között akarunk fordítani, akkor az *N* oldalú poligon átlóinak és éleinek összegével azonos összeköttetést kell létesítenünk közöttük. Valójában az irányítottság miatt kétszerannyi, vagyis *n*(n-1)* megoldásra lesz szükségünk. Ha kiválasztunk egy közvetítő nyelvet, akkor elég azt a poligon csúcsaival irányítottan összekötnünk, vagyis 2**n* nyelvi modul szükséges.

Vizsgáljuk meg az Európai Uniót, melynek 2007. január elseje óta 27 tagországa és 23 hivatalos nyelve van. Ez 23*22=506 közvetlen kapcsolatot jelentene. Ha a csúcsokat a középponttal kötjük össze, és a közvetítő nyelves megoldást választjuk, a szükséges kapcsolatok száma 46 lesz. Csak az utóbbi megoldás megvalósítása tűnik reálisnak.

A probléma az, hogy bár a közvetítő nyelvre sok éve várunk, mégsem született meg. Ennek számos szakmai és egyéb oka is van. Mára az interlingva kérdése a gyakorlatban eldőlni látszik. A fordítóprogramok -- a természetes nyelvű fordítókhoz hasonlóan -- természetes nyelveket választanak közvetítő nyelvül. Ez általában (bár nem kizárólagosan) az angol nyelv.

Értékeljük a kialakult helyzetet: a mesterséges közvetítőnyelvekkel kapcsolatban felmerülő, a nyelv kidolgozottságát érintő problémák kiküszöbölését egy világnyelv esetén a beszélők sokasága biztosítja. Az élő nyelvek területileg és időben változnak ugyan, de ez kezelhető mértékű. Nem jelentkezik a nyelvalkotóktól való függőség, hiszen a mű, embermilliók közös kincse, szabadon felhasználható szellemi termék. Ugyancsak nem probléma a fejlesztők előzetes nyelvi képzettsége, mert pl. angol nyelvtudással a fejlesztők általában rendelkeznek. Az általános érvény és az esélyegyenlőség szempontjai viszont nem teljesülnek. A közvetítő nyelv és a vele rokon nyelvek előnyösebb helyzetbe kerülnek, de ezek a szempontok alulmaradni látszanak.

Vizsgáljuk meg az angolt, mint számítógépes közvetítőnyelvet. Az angol nyelv ismeretét tekintve osszuk az embereket három csoportra: angol anyanyelvűek, angolul tudók és angolul nem tudók.

	Európai Unió (2006)	**Világ (2006)**
Angol anyanyelvűek	13%	5%
Angolt nem anyanyelvként beszélők	38%	8,2%
Angolul tudók összesen	51%	13,2%

Felhasznált adatok: [6], [7]

Mit történik akkor, ha fordítóprogramok jönnek létre az angol és az Európai Unió más nyelvei között? A 13% anyanyelvű számára az előny nyilvánvaló, mert közvetlenül érthetik meg az összes többi nyelvet. Az angolt második nyelvként beszélő 38% számára is elérhetővé válik a többi nyelv, mert az angol és a saját anyanyelvük közötti fordítást nyelvismeretük alapján elvégzik. Például, egy magyar–angol fordító nemcsak az angoloknak jelent megoldást, hanem segítségével például az angolul jól beszélő németek is el tudják olvasni a magyar weblapokat. A fordítások célja elsősorban a megértés, amelynél nincs szükség arra, hogy az anyanyelven írt szöveg valóban létre is jöjjön. És mit jelent ugyanez a népesség felét kitevő angolul nem beszélő uniós polgárok számára? Számukra ez a megoldás az angol szövegek megértésének lehetőségét teremti meg. Azaz mindhárom csoport számára hasznos az angol központú nyelvi megoldás.

	English	French	German	Spanish	Russian	Italian	Swedish
EU25	**77%**	**33%**	**28%**	**19%**	**3%**	**2%**	**0%**
BE	88%	50%	7%	9%	0%	1%	-
CZ	89%	9%	66%	4%	9%	0%	-
DK	94%	13%	62%	13%	0%	0%	0%
DE	89%	45%	3%	16%	6%	2%	-
EE	94%	6%	22%	1%	47%	0%	1%
EL	96%	34%	50%	3%	0%	6%	-
ES	85%	44%	14%	4%	0%	1%	-
FR	91%	2%	24%	45%	0%	6%	-
IE	3%	64%	42%	35%	1%	4%	0%
IT	84%	34%	17%	17%	0%	0%	-
CY	98%	49%	19%	2%	4%	4%	0%
LV	94%	6%	28%	1%	42%	0%	0%
LT	93%	6%	34%	2%	43%	0%	0%
LU	59%	83%	43%	2%	0%	1%	-
HU	85%	4%	73%	3%	2%	2%	-
MT	90%	24%	13%	2%	-	61%	-
NL	90%	22%	40%	21%	0%	0%	-
AT	84%	29%	2%	10%	4%	11%	-
PL	90%	7%	69%	1%	10%	1%	-
PT	90%	60%	8%	7%	-	0%	-
SI	96%	6%	69%	3%	0%	12%	0%
SK	87%	7%	75%	3%	6%	1%	0%
FI	85%	10%	24%	3%	10%	0%	38%
SE	99%	17%	35%	31%	1%	0%	1%
UK	5%	71%	34%	39%	1%	3%	-
BG	87%	13%	49%	5%	14%	1%	-
HR	82%	5%	69%	2%	0%	14%	-
RO	64%	34%	17%	7%	2%	8%	-
TR	72%	12%	52%	1%	2%	1%	-

= First language = Second language

Várható, hogy az Európai Unió lakosságának nyelvtudása a jövőben intenzíven növekedni fog. 2002-ben elhatározás született arról, hogy minden uniós állampolgárnak lehetőség szerint két nyelvet kell majd megtanulni.

Az előző táblázat azt mutatja, hogy egy felmérés szerint melyek az első és második helyen választott nyelvek. Az angol vált a természetes közvetítő nyelvvé, és a fenti szempontok miatt most mint számítógépes közvetítőnyelv is egyre nagyobb tért hódít.

A gépek nyelvtudása azonban még az emberek nyelvtudásánál is gyorsabb ütemben fejlődik. Egy adott fejlettségi szint után a gépi fordítások összeláncolhatóvá válnak, mert a kétszeres fordítás is elfogadható minőséget fog adni. Ezt a lépést a minőségre kevesebbet adó on-line internetes fordítók már egy ideje meg is tették. Ekkor pedig az angolul nem tudók számára is megteremtődik az egyéb nyelveken való kommunikáció lehetősége.

A fentiek alapján, mi a többnyelvűségre való törekvés helyett a közeljövőben inkább megvalósított angol-magyar és magyar-angol rendszereink továbbfejlesztésén fogunk dolgozni.

5. A jövő

A fejlesztés következő, harmadik fázisának célja, hogy a fordítóprogram valóban a mindennapok eszközévé váljon. Ehhez a nyelvi minőség és a funkciók további tökéletesítésére lesz szükség. Ezekről azonban már nem ebben a sorozatban, hanem a problémával foglalkozó konkrét előadások formájában számolunk be.

6. Hivatkozások

1. Tihanyi László: A MetaMorpho projekt története. In: Alexin Zoltán; Csendes Dóra (szerk.) *Az 1. Magyar Számítógépes Nyelvészeti Konferencia előadásai*, 247−253. SZTE, Szeged (2003)
2. Tihanyi László: A MetaMorpho projekt 2004-ben. In: Alexin Zoltán; Csendes Dóra (szerk.) *A 2. Magyar Számítógépes Nyelvészeti Konferencia előadásai,* 85−87. SZTE, Szeged (2004)
3. Tihanyi László: A MetaMorpho fordítóprogram projekt 2005-ben. In: Alexin Zoltán; Csendes Dóra (szerk.) *A 3. Magyar Számítógépes Nyelvészeti Konferencia előadásai*, 99−107. SZTE, Szeged (2005)
4. Tihanyi László, Merényi Csaba: A MetaMorpho fordítóprogram projekt 2006-ban. In: Alexin Zoltán; Csendes Dóra (szerk.) *A 4. Magyar Számítógépes Nyelvészeti Konferencia előadásai*, SZTE, Szeged (2006)
5. Hans-Udo Stadler, Ursula Peter-Spörndli: The Quest for Machine Translation Quality at CLS Communication. *Proceedings of the MT Summit*, Copenhagen (2007)

Adatok:

6. Az angol anyanyelvűek (2006): 326 millió, angolul tudók: 860 millió http://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population
7. A világ lakossága (2006): 6,5 milliárd: http://en.wikipedia.org/wiki/World_population
8. A beszélt nyelvek száma és lefedettségük: az 374 nyelv, amelyet 1 milliónál többen beszélnek (és amely az összes 6912 nyelv 5%-a) lefedi a lakosság 94%-át. http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

Főnévicsoport-azonosító módszerek főnévicsoport-szinkronizációs célokra

Pohl Gábor

Pázmány Péter Katolikus Egyetem Információs Technológiai Kar
1083 Budapest, Práter utca 50/A
pohl@itk.ppke.hu

Kivonat: A MorphoTM frázisalapú kísérleti fordítómemóriában alapvető fontosságú a főnévi csoportok (NP) automatikus azonosítása és szinkronizációja (*alignment*). Az előző években bemutattunk egy gyors, szótári és szófaji információkra építő NP-szinkronizáló módszert, most csak a főnévi csoportok azonosításával foglalkozunk. A MorphoTM rendszerben a forrásoldali NP-ket minden esetben mély szintaktikai elemzéssel választjuk ki, ebben a cikkben a célnyelvi NP-azonosítás lehetséges módszereit hasonlítjuk össze. A mérési eredmények azt mutatják, hogy az NP-ket fordításaik alapján meghatározó sekély elemzést használó módszerünk megfelel az elvárásoknak, illetve hogy a pontosság növelhető, ha a kiválasztott NP-jelölteket mély szintaktikai elemzővel (de a teljes mondat elemzése nélkül) ellenőrizzük.

1 Bevezetés

Cikkünkben főnévi csoportok (NP) azonosításának módszereit főnévicsoport-szinkronizációs alkalmazásban való felhasználhatóságukat vizsgálva hasonlítjuk össze. Az NP-azonosítás módszereinek szinkronizációs feladatban való vizsgálatát a MorphoLogic-nál fejlesztett MorphoTM [1] angol-magyar kísérleti fordítómemória fejlesztése motiválja.

A MorphoTM rendszer két fő ponton tér el a piacon termékként elérhető nagyrészt nyelvfüggetlen fordítómemóriáktól. Egyrészt a keresett forrásmondathoz az adatbázisban leginkább hasonlót nem csak karakteralapú, felszíni karaktersorozatokat összehasonlító hasonlósági mértéket használva keresi, a hasonlóság számítása során morfológiai és szintaktikai információra is épít. Másrészt nem csak teljes szegmenseket (mondatokat) keres az adatbázisában. Hasonló NP-ket illetve lehetséges mondatvázakat (a mondatban az NP-ket szimbolikus NP helyekre cserélve kapott struktúrát nevezzük így) is keres, majd a leghasonlóbbak tárolt fordításaiból a megfelelő morfológiai alakokat generálva épít javasolt fordítást.

A fordítómemória fedését növeli, hogy eredetileg különböző mondatok NP-it illetve mondatvázait kombinálva is képes fordításokat javasolni. (Azaz a MorphoTM valójában egy egyszerű minta-alapú gépi fordítórendszer.) A különböző emberi fordításokat automatikusan kombináló módszer hibás fordításokat eredményez, ha az NP-k fordítása a mondat vázától vagy más mondatbeli NP-től függ, ugyanakkor a javasolt

fordítás részei még ilyenkor is segíthetik a fordítómemóriát használó fordítót. Úgy gondoljuk, hogy a teljes szegmenseket tároló fordítómemóriáknál magasabb fedés még akkor is hasznos, ha az elérhető pontosság biztosan alacsonyabb.

Az adatbázisban tárolni kívánt NP-k azonosítását és fordításaikkal való összerendelését (szinkronizációját, párhuzamosítását) nem bízhatjuk a fordítómemóriát használó fordítóra, mert az NP-k megjelölésére és összerendelésére fordított munkaidő a későbbiekben valószínűleg nem térülne meg a fordítómemória fedésének növekedése révén. Az NP azonosítás és szinkronizáció feladatát tehát gépi algoritmusokra bízzuk.

Az előző években kidolgoztunk egy gyors, lexikai jegyeket használó NP szinkronizációs módszert [2], amely párhuzamosított korpuszon tanított osztályozót alkalmaz [3], így az NP-k kiválasztására alapvetően különbözően viselkedő (hibázó) módszereket is választhatunk, csak egy kézzel NP-szinkronizált tanítókorpuszt kell készíteni a választott NP-azonosító módszerhez. Cikkünkben a korábban kidolgozott NP-szikronizációs módszerünkkel csak érintőlegesen foglalkozunk, célunk a MorphoTM rendszerben az NP-k azonosítására leginkább alkalmas módszer mérésekkel történő kiválasztása, pontosabban az NP-k fordításbeli azonosítására legalkalmasabb módszert keressük, mivel a forrásoldali mondatokban mindenképp teljes mondatelemzéssel választjuk ki a főnévi csoportokat.

2 Az NP-azonosítás hibái

Automatikus módszerekkel, a vizsgált mondat megértése nélkül (szemantikai információk nélkül) az NP-k helyes azonosítása elvileg sem lehetséges, ahogyan erre az alábbi 1. példa is rámutat.

[I] saw [the man] in [the street].
[I] know [the girl in the garden]. (1. példa)

1. példa: Az automatikus NP-azonosító módszerek általában nem tudják helyesen kiválasztani a fenti angol példamondatok maximális méretű főnévi csoportjait. A maximális főnévi csoportokat zárójelekkel jelöltük.

Az automatikus NP-azonosító módszerek minimalizálni próbálják a hibákat, természetesen a teljes siker reménye nélkül. A fő kérdés, hogy egy adott alkalmazásra mely módszer a legalkalmasabb, illetve hogy az egyes módszerek hibái milyen következményekkel járnak az adott alkalmazásban. A hibákat tehát nem az NP azonosítás közvetlen kimenetében, hanem az NP-k azonosítására építő alkalmazás kimenetében kell majd keresnünk.

3 Az NP-azonosító és szinkronizáló módszerekkel szembeni elvárások a MorphoTM rendszerben

A MorphoTM rendszerben az NP-azonosítás és szinkronizáció minősége és sebessége egyaránt fontos. A minőség mérésére a szokásos fedés és pontosság értékeket használjuk.

A pontosság azért fontos, mert az adatbázisban tárolt hibásan azonosított és/vagy szinkronizált NP-párok, újra és újra megjelennek a javasolt fordításokban, amíg kézzel el nem távolítjuk őket az adatbázisból (jelenleg a MorphoTM rendszerben semmilyen automatikus megoldás nincs az adatbázis tisztítására).

Az NP-azonosító és szinkronizáló módszerek fedése több okból is fontos. Egyrészt minél több NP-párt tárolunk a memóriában, annál nagyobb lesz a fordítómemória fedése, másrészt minél több NP-párt azonosítunk egy mondatpárban, annál általánosabb lesz a párosított NP-ket szimbolikus NP-helyekre cserélve kapott mondatváz.

Fordítómemória alkalmazás esetében a sebesség különösen fontos. Az új mondatpárokat kevesebb, mint 1 másodperc alatt úgy kell tárolni a memóriában, hogy a frissen tárolt fordítás már a következő mondat javasolt fordításában is megjelenhessen. A hagyományos fordítómemóriák gyorsan tárolják a fordításokat, így a fordítók egy fejlettebb fordítómemóriától is jogosan elvárják ugyanezt. Az NP-k szinkronizációjára kifejlesztett módszerünk [2, 3] elég gyors (hosszú mondatpárokon is kevesebb, mint 10 ms alatt lefut egy 4 éves PC-n), azonban ez az NP-azonosításra használt módszerekről már nem mondható el, utóbbiak futtatása csak nehezen vagy egyáltalán nem fér be az egy másodperces időkeretbe.

4 Célnyelvi NP-azonosító módszerek

A MorphoTM rendszerben a forrásnyelvi mondatok NP-it a teljes mondat mély elemezésével választjuk ki. A tárolt fordítás NP-inek azonosítására több lehetőségünk is van, ezeket hasonlítjuk most össze. Megvizsgáljuk a teljes mondat mély szintaktikai elemzésének lehetőségét, az NP-ket fordításaik alapján sekély elemzővel meghatározó módszerünket valamint a két módszer kombinálásának lehetőségeit.

4.1 A célnyelvi mondatok mély elemzése

A mély elemzés során a teljes mondatot lefedő elemzési fát keres az elemző, vagy ha ilyet nem talál, a mondat kiesebb részeit külön fákkal fedi le. A mély elemzés memória- és számításigényes. A módszer előnye, hogy amennyiben a teljes mondatot lefedő elemzési fát talál, nagy pontosságú NP-azonosításra képes. Sajnos azonban számos hátránnyal is számolnunk kell:

- A MorphoTM rendszerben a magyar mondatok teljes elemzése túl sokáig (>1s) tartana. (A forrásnyelvi angol mondatok elemzése is 1 másodperc körüli időt vesz el, így nagyon gyors módszer kellene a magyar elemzéshez.)

- A teljes mondatot fedő elemzési fát sajnos a mondatok többségénél még nem talál az elemző, így a fedés még nem elég magas.
- A gyakori elemzési hibák csökkentik az NP-azonosítás pontosságát. A teljes mondatot lefedni nem képes elemzési fák gyakran tartalmaznak hibásan azonosított NP-ket.
- A forrás és célnyelvi mondatok elemzésére használt elemzők különböző hibákat ejtenek, ami megnehezíti a pontos szinkronizációt.

Az előbbiek miatt sajnos a teljes mondatelemzés továbbra sem valódi lehetőség a MorphoTM rendszerben.

4.2 Fordítás által támogatott sekély elemzés

A korábbi években kifejlesztettünk egy speciális NP-azonosító módszert, amely párhuzamos szövegekben a mondatpárok forrásoldalán megbízható módszerrel (a gyakorlatban a mondat teljes elemzésével) meghatározott NP-k párjelöltjeit jelöli meg a fordításban.

Módszerünk a megbízhatóan kiválasztott forrásnyelvi NP-k szavait leképezi a célnyelvi mondat szavaira. Tövesített szótári megfejeletetést, szófaji megfeleltetést alkalmazva, illetve hasonló alakú szavakat keresve [4] minden egyes nem grammatikai funkciót betöltő szó összes lehetséges párját megkeressük a célnyelvi mondatban. A pusztán grammatikai funkciót betöltő szavakhoz (pl. névmás, névelő) nem keresünk párt. Mivel egy forrásnyelvi szó több megfelelője és akár többször is előfordulhat a célnyelvi mondatban, a lehetséges találatok közül azt választjuk ki, amelynek szavai a lehető legrövidebben illeszkednek a célnyelvi mondatra. A legrövidebb illeszkedés szavait NP-váznak nevezzük. Természetesen a találatok között más szavakat is tartalmazhat az NP-váz. Az NP-vázat ezek után egyszerű szintaktikai szabályok (sekély nyelvtan) szerint, a forrásnyelvi főnévi csoport le nem fedett szavainak szófaját is figyelembe véve teljes célnyelvi főnévi csoporttá bővítjük. A bővítés során először a célnyelvi mondatban az illeszkedő szavak közötti meg nem feleltetett szavakat próbáljuk szófajuk alapján a forrásnyelvi NP meg nem feleltetett szavaival összerendelni, majd balra, illetve szükség esetén jobbra bővítjük a célnyelvi főnévi csoportot. (A bővítés preferált iránya nyelvfüggő, illetve függ attól, hogy a le nem fedett szavak között hány főnév van.)

Az algoritmus több előnyös tulajdonsággal bír:

- Az NP-azonosítás nagyon gyors (<10 ezredmásodperc egy 4 éves PC-n futtatva).
- Az algoritmus szinte teljesen nyelvfüggetlen, a nyelvpárfüggő szófajmegfeleltetési tábla, illetve a sekély nyelvtani szabályok meghatározása egy új nyelv esetében kevesebb, mint egy nap alatt megvalósítható.
- Nagy kétnyelvű vagy automatikusan gyűjtött szótárral magas fedés érhető el.

A módszer gyengéje megkérdőjelezhető pontossága. A hibásan kiválasztott (hibásan bővített) NP-ket a szinkronizáló algoritmusnak kell elvetnie.

4.3 NP azonosító módszerek kombinációi

Az NP-ket a fordításuk alapján sekély elemzéssel meghatározó módszerünk pontosságának növelése érdekében megpróbáltuk a módszert úgy kombinálni mély elemzéssel, hogy a mély elemzés előnyét (a nagy pontosságot) átvegyük, de a számításigény (azaz a futási idő) alacsony maradjon. Alapötletként megvizsgáltuk, hogyan lehetne csökkenteni az elemzési időt a mély nyelvtan módosítása nélkül. A rendelkezésünkre álló MetaMorpho elemezőt [5] vizsgálva azt tapasztaltuk, hogy legfeljebb 10 szavas bemenet esetén az elemzés megfelelően gyors. A kérdés ezek után csak az, mennyire várhatók pontos eredmények, ha nem a teljes mondatot elemezzük, hanem csak NP jelölteket ellenőrzünk a mély elemzővel (azaz azt vizsgáljuk, hogy az egész elemzett mondatrész lehet-e főnévi csoport).

Két lehetséges kombinációját próbáltuk ki az NP-k fordításalapú meghatározásának és mély nyelvtannal való ellenőrzésének:

- az NP-váz bővítésekor a mély elemzőt használva,
- a mély elemzőt csupán a fordításalapú sekély elemzés eredményének ellenőrzésére használva.

Az első esetben megvizsgáltuk, hogy balra, illetve jobbra maximum két szóval bővítve az NP-vázat a mély elemző által elfogadható NP-t kapunk-e. A bal oldali bővítést részesítettük előnyben, ha a baloldali bővítés eredményes volt, jobboldali bővítést nem kíséreltünk meg.

A második esetben csak azokat a sekély nyelvtannal kiválasztott NP jelölteket választottuk ki, amelyeket a mély elemző elfogadott.

5 Kiértékelési módszer

Tavaly az NP-azonosító módszerek elméleti összehasonlítását már elkezdtük [3], azonban még nem állt módunkban méréseket végezni. A puszta elméleti fejtegetésnél mérnökként sokkal fontosabbnak tartjuk a legegyszerűbb mérést is, persze csak akkor, ha a mérés során azt mérjük, amit kell és ahogyan kell.

Három NP-azonosító módszert hasonlítunk most össze (a fordításuk alapján sekély nyelvtannal NP-ket meghatározó módszerünket, az előbbi eredményét mély elemzővel ellenőrző módszert, illetve az NP-váz bővítését mély elemzőre bízó módszert). Az összehasonlítás célja annak megállapítása, hogy melyik módszer alkalmasabb NP-szinkronizációs alkalmazásban való felhasználásra, így elsősorban nem a nyers kimenetüket vizsgáljuk, hanem azt hogy milyen szinkronizációs eredmények érhetők el velük.

NP-szinkronizáló módszerünk többféle skaláris lexikai jegyet (*feature*) határoz meg minden egyes összehasonlított NP-párhoz, majd a jegyértékek egyszerű normalizálása után egy korpuszon tanított osztályozó dönti el, hogy a vizsgált NP-párjelöltet fordításként tároljuk-e a fordítómemória adatbázisában. Az osztályozási elfogultság (*bias*) elkerülése érdekében minden vizsgált NP-azonosító módszerhez külön be kell tanítanunk az osztályozót, így minden vizsgált módszerrel az elérhető legjobb eredményt fogjuk kapni.

6 Mérés

A mérésekhez egy kisméretű, 100 mondatpárból álló párhuzamos korpuszt választottunk. A korpusz mondatpárjai különböző szövegekből származnak, kiválasztásuknál nem vettünk figyelembe különleges szempontokat. A korpusz angol oldalán a mondatok átlagos hossza 14 szó.

Az angol mondatok NP-it a MetaMorpho angol elemzővel [5] azonosítottuk, majd a mondatpárok magyar oldalán a három vizsgált NP-azonosító módszert alkalmazva három angol-magyar párhuzamos korpuszt készítettünk.

A magyar főnévi csoportok vázának meghatározásakor egy 116000 szó- és kifejezéspárt tartalmazó angol-magyar szótárat használtunk. Mély elemzőként a MetaMorpho magyar elemzőt használtuk.

A főnévi csoportok automatikus azonosítása után a három párhuzamos korpuszban kézzel szinkronizáltuk a kiválasztott főnévi csoportokat angol fordításaikkal, így három tanító- és tesztkorpuszt készítetünk az NP-szinkronizáló módszerünk osztályozójának tanítására és tesztelésére.

A kézi szinkronizálás során a következőképp jártunk el:

- Csak teljesen megfeleltethető NP-ket rendeltünk egymáshoz, ha az egyik NP fordítása csak része volt a másik NP-nek, nem rendeltük őket egymáshoz.
- Amennyiben az NP-k az adott mondatpárban egymás fordításai voltak, akkor is rögzítettük őket, ha a mondatváztól függően más mondatpárban nem lehettek volna egymás fordításai.
- Egymáshoz rendeltük azokat az NP-ket, amelyek kis mértékben, de csak grammatikai funkciót betöltő szavakban különböztek, és a mondatokat egészben vizsgálva egymás fordításának tekinthettük őket (pl. *this family – a család*).

Az automatikus NP szinkronizáció minőségét a három tesztkorpuszon 10-szeres keresztkiértékeléssel mértük. A szinkronizáló módszerünkben a tavalyi méréseik szerint legmegfelelőbb logisztikus regressziós osztályozót alkalmazva mértük a szinkronizáció pontosságát.

7 Mérési eredmények

Mindhárom vizsgált NP-azonosító módszer esetében megvizsgáltuk a kiválasztott NP-jelöltek számát, a helyesen kiválasztott NP-jelöltek számát, az adott NP-azonosító módszerrel készített korpuszon tesztelve az NP-szinkronizáló algoritmus döntési pontosságát (a helyes döntések számát, a hamis pozitív, illetve hamis negatív NP-párokat), valamint a fordítómemória adatbázisába kerülő helyes illetve hibás NP-párok számát. Természetesen az utóbbi két érték függ az előzőektől, de segítenek a MorphoTM rendszerben legmegfelelőbb NP-azonosító módszer kiválasztásában.

Az I. táblázat a nyers NP-azonosítási eredményeket mutatja, a II. táblázatban az NP-szinkronizáció eredményeit rögzítettük, a III. táblázat pedig az egyes módszereket alkalmazva a fordítómemória adatbázisába helyezett helyes illetve hibás NP-párok számát mutatja.

Az eredményekből tisztán kiolvasható, hogy az NP-vázat mély elemzővel bővítő módszer rosszabb az NP-ket fordításaik alapján sekély elemzővel meghatározó módszernél, mind nyers NP-azonosítási eredményeiben, mind a vele elérhető NP-szinkronizációs eredményeket tekintve.

Az NP-ket fordításaik alapján sekély elemzővel meghatározó módszer NP-jelöltjeit mély elemzővel ellenőrizve magasabb pontosság érhető el, azonban lényegesen alacsonyabb fedés árán.

Az I. és II. táblázatban az NP-azonosítás pontosságát valamint a helyes szinkronizációs döntések számát összevetve a szinkronizáló algoritmus futtatásának haszna is megfigyelhető. A szinkronizációs algoritmus eredményesen veti el a rosszul kiválasztott NP-ket, de futtatásának haszna csökken, ha a bemenetén a helyesen kiválasztott NP-k aránya magas. (Más kérdés, hogy ha nem olyan NP-azonosító módszereket használnánk, amelyek eleve egy fordításbeli NP-hez keresnek párt, akkor az NP-szinkronizáló algoritmusra mindenképp szükség lenne. Jelen esetben csak szűrőnek használjuk az egyébként általánosan használható módszert.)

I. táblázat: Nyers NP-azonosítási eredmények

Módszer	Kiválasztott NP-pár	Helyes párok	Pontosság
TGSNPP	**228**	186	82%
NPS+DP	196	139	71%
TGSNPP+DP	130	115	**88%**

TGSNPP = NP-ket fordításaik alapján sekély elemzővel azonosító módszer, NPS+DP = az NP-vázat elemzővel bővítő módszer, TGSNPP+DP = TGSNPP eredményének ellenőrzése mély elemezővel. A TGSNPP fedése nagy (=sok kiválasztott NP-párjelölt), A TGSNPP+DP módszer pontossága nagy. Az NPS+DP módszer pontossága meglepően alacsony.

II. táblázat: Szinkronizációs eredmények

Módszer	Helyes döntés	Hamis pozitív	Hamis negatív
TGSNPP	86%	11%	3,1%
NPS+DP	83,7%	9,7%	6,6%
TGSNPP+DP	**90%**	**7,7%**	**2,3%**

A TGSNPP+DP módszer magas szinkronizációs pontosságot eredményez. Az NPS+DP módszer mindkét másiknál rosszabbul szerepel. (A módszerek azonosítóit lásd az I. táblázatnál.)

III. táblázat: adatbázisba került NP-párok

Módszer	Helyes párok	Helytelen párok	Pontosság
TGSNPP	**179**	25	87,7%
NPS+DP	126	19	86,9%
TGSNPP+DP	112	10	**91,8%**

A TGSNPP+DP kombinált módszer érte el a legnagyobb pontosságot, de fedése viszonylag alacsony. Az NPS+DP módszer a mély elemzőt nem használó TGSNPP-nél is rosszabb eredményt ért el. (A módszerek azonosítóit lásd az I. táblázat ismertetőjében.)

8 Összegzés

Korpuszalapú méréseket végeztünk annak érdekében, hogy megtaláljuk az NP-szinkronizációs feladatra legalkalmasabb NP-azonosító módszert.

A mérések ismét igazolták, hogy a fordítás alapján sekély nyelvtannal NP-ket azonosító módszerünk akár önmagában is megfelelő módszer NP-szinkronizációs feladatokra. A mérések alapján kijelenthetjük, hogy a forrásnyelvi NP-kből a célnyelvi mondatra leképezett NP-vázakat mély elemzővel bővítve, a fordítás által támogatott sekély elemzőt használó módszernél rosszabb eredményeket érünk el, ugyanakkor a mély elemzőt a pontosság növelése érdekében használhatjuk a sekély elemzés által kiválasztott NP-jelöltek ellenőrzésére.

Az NP-jelöltek ellenőrzésére a MetaMorpho magyar elemzőt használva a fordítómemória adatbázisába hibásan felvett NP-párok arányát 12,3%-ról 8,2%-ra sikerült csökkenteni, ami már lényeges különbség, főképp mivel eddig nem dolgoztunk ki megoldást az adatbázis automatikus tisztítására.

Sajnos a pontosság növelését csak a fedés jelentős csökkenése árán tudtuk elérni. Az adatbázisba felvett helyes párok száma 37,4%-kal csökkent. Az alacsonyabb fedés okozta problémára megoldás lenne, ha a mély elemző által el nem fogadott NP-jelölteket is felvennénk az adatbázisba, viszont a fordítások ajánlásakor az elemző által elfogadott párokat részesítenénk előnyben.

A mérések azt is megmutatták, hogy a jelenleg csak a párjelöltek szűrésére használt NP-szinkronizáló módszerünk [2, 3] minden esetben jobb eredményt ért el a minden párjelöltet elfogadó baseline módszernél.

Bibliográfia

1. Hodász G., Pohl G.: MetaMorpho TM: a linguistically enriched translation memory. In *International Workshop, Modern Approaches in Translation Technologies* (szerk.: Hahn, W.; Hutchins, J.; Vertan, C.), Borovets, pp. 26-30, 2005.
2. Pohl G.: English-Hungarian NP-alignment in MetaMorpho TM. In *Proceedings of EAMT 2006 (11th Annual Conference of the European Association for Machine Translation)*, pp. 69-74, 2006.
3. Pohl G.: A MorphoTM főnévicsoport-szinkronizáló módszereinek továbbfejlesztése és vizsgálata. In *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2006)*, pp. 190-201, 2006.
4. Simard, M., Foster, G. & Isabelle, P. (1992): Using Cognates to Align Sentences in Bilingual Corpora. In: *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine translation*, (TMI92), Montreal, pp. 67-81, 1992.
5. Prószéky, G.: Translating While parsing. In *A Man of Measure* (szerk.: M. Suominen et al.), The Linguistic Association of Finland, Turku, pp. 449-459, 2006.

Élő vagy élettelen?

Sass Bálint

MTA Nyelvtudományi Intézet és PPKE ITK MMT Doktori Iskola
e-mail: joker@nytud.hu

Kivonat Hogyan lehet megállapítani az igei keretek alanyi pozíciójának élő vagy élettelen voltát? A kidolgozott módszer az igei személyragok eloszlását, valamint az élőre és élettelenre utaló vonatkozó névmások arányát veszi tekintetbe. Az élettelen alanyú keretek 70%-át megtalálja, miközben szinte sosem határoz meg élő alanyú keretet élettelenként. A nyerhető igelistát egy magyar-angol fordítórendszer lexikai erőforrásába építve arra használjuk, hogy a pro-drop magyar mondatok fordításakor a „semmiből" megfelelő testes névmást generáljunk az angol oldalon.

Kulcsszavak: élő, élettelen, gépi fordítás, pro-drop

1. Bevezetés

Hogyan fordítanánk angolra az alábbi két magyar mondatot?

1. *Alszik.*
2. *Elromlott.*

Valószínűleg legtöbben a következő angol megfelelőket tartanák természetesnek, legalábbis abból a szempontból, hogy automatikusan az ige szemantikájának megfelelő élőre illetve élettelenre utaló névmást használnák:

1. *He/she is sleeping.*
2. *It has gone wrong.*

Általánosan fogalmazva arról a kérdésről van tehát szó, hogy a gépi fordítás során mit tehetünk olyan esetekben mikor a forrásnyelv nem specifikál bizonyos tulajdonságokat, jegyeket, a célnyelv viszont ugyanazon a ponton elvárja a tulajdonság egy konkrétan megadott értékét. Az egyik lehetőség, hogy dinamikusan megkíséreljük kitalálni a szövegkörnyezetből az elvárt értéket, a most bemutatandó másik lehetőség pedig az, hogy a lexikonba bekódolt alapértelmezett értékeket használunk. Egyértelmű esetekben ez a módszer hibátlan megoldást ad futásidejű számítási igény nélkül. A javasolt eljárás tehát leegyszerűsítve az lesz, hogy nagyméretű korpuszban mért gyakoriságok alapján megbecsüljük a jegy alapértelmezett értékét, rögzítjük a lexikonban, és ezt az értéket használjuk akkor, ha nincs információnk a jegy aktuális értékéről, esetünkben az alany élő vagy élettelen voltáról.

2. Az élőségi skála jelentősége

Az *élőségi* (vö: *animacy*) skála (vagy élő/élettelen skála) a nyelvi prominenciaviszonyokat meghatározó egyik tényező, sok esetben valamely elem élő illetve élettelen volta szerint választunk két nyelvi forma között [1]. A megértés szempontjából központi szerepe van, lehetővé teszi, hogy a dialógusban követni tudjuk, hogy éppen melyik szereplőről van szó [2]. Univerzálisan kimondható, hogy az egyes szereplők élőségi skálán elfoglalt helye arányos az aktuális esemény befolyásolására való képességükkel [3].

Az élőségi skála a természetesnyelv-feldolgozásban kisebb figyelmet kapott, az alapkérdéssel – főnevek élő illetve élettelen voltának megállapításával – foglalkozó tanulmányok csak az utóbbi időben jelentek meg [4,5]. Éppen a gépi fordítás generálás fázisa az a terület, ahol az élőség fontossága nyilvánvaló [1]. A szemantikai szelekció az igék természetes tulajdonsága, ennek egy esete, hogy bizonyos igék élő ill. élettelen szereplőt várnak el az alanyi pozícióban. A fent felvetett kérdésnek, hogy ti. adott konstrukció adott pozícióját betöltő szóosztályról állapítsuk meg az élőségi értékét, a számítógépes kezelésével nem találkoztam az irodalomban.

Az univerzális *ember* > *állat* > *élettelen* skálán a különböző nyelvek különböző pontokon húznak határvonalakat [3]. A magyar és az angol is az *ember* kategóriát választja el az összes többitől, ennek megfelelően, amikor a továbbiakban élő és élettelen kategóriákról lesz szó, akkor az állatokat nyelvi szempontunk alapján (vö: *ami*-vel és *it*-tel hivatkozunk rájuk) az élettelenek közé soroljuk.

3. A konkrét kérdés

Az angollal ellentétben a magyar pro-drop nyelv, a személyes névmást semleges mondatban nem tesszük ki. Egyes szám harmadik személyben mindkét nyelv elkülöníti az élőre ill. az élettelenre utaló névmást. Probléma akkor merül fel, mikor az egyes szám harmadik személyű magyar mondatban nincs kitéve a névmás, az angol oldalon pedig el kell döntenünk, hogy a „semmiből" élő vagy élettelen testes névmást generáljunk.

Általános megállapítás, hogy az alany hajlamos élő és ágens lenni [3,5]. Ennek tudatában megtehetjük, hogy minden esetben *he/she*-t generálunk (a nemek közötti különbségtétellel jelen dolgozatban nem foglalkozunk). Kiértékeléskor ezt a primitív – azonban meglehetősen jó eredményeket adó – módszert fogjuk baseline-nak tekinteni. Felmerült egy másik baseline módszer lehetősége is, miszerint a tárgyas igék alanya alapértelmezésben élő, a tárgyatlanoké pedig élettelen. Ezt elvetettük, mert a fenti egyszerűbb „mindig élő" baseline rendszeresen jobb eredményt adott.

A fordítórendszer alapértelmezés szerint valóban *he/she*-t generál, így a kidolgozandó módszer felé az az elvárás, hogy lehetőleg soha ne tévedjen abban az irányban, hogy élő helyett élettelent javasol.

4. Módszerek, kiértékelés

4.1. Nyersanyag

A vizsgálatokhoz a Magyar Nemzeti Szövegtár egyvonzatkeretes egységekre bontott változatát [6] használtam. Ezek az egységek egy igét, és a mellette álló bővítményeket tartalmazzák. Így lehetőség van arra, hogy ne csak puszta igékkel, hanem igei keretekkel is dolgozzunk (pl. *tudomásul vesz vmit*, *kiderül vmiről vmi*, *rendben van vmi*), az igék különféle kereteit külön kezeljük. Hiányosság, hogy amikor adott keret megjelenéseit kérdezzük le a korpuszból, akkor csak azt lehet megadni, hogy mely bővítmények szerepeljenek az ige mellett, azt nem lehet meghatározni, hogy mi ne szerepeljen. Következésképpen a *megy* igére vonatkozó lekérdezés az ige bővítményeit különféle variációkban tartalmazni fogja, ezért jóval zajosabb lesz, mint a *nyilvánosságra hoz vmit* keretre vonatkozó.

Az MNSZ gyakoribb igei kereteiből válogattam a mintáimat: konkrétan azok közül a keretek közül, amik 925-nél többször fordulnak elő a Szövegtárban. Mindvégig *type* alapon dolgoztam, azaz egy igei keretet tekintettem egy egységnek, szemben azzal a felfogással, mikor egy adott előfordulás, mondat a vizsgálati egység.

4.2. Előzetes: a *3sz%* módszer

Komlósy megállapítja, hogy bizonyos igék csak egyes szám 3. személyben használatosak, és ezeknek az igéknek „az alanyi vonzata nem jelölhet személyt” [7, 335.o.]. Az 1. és 2. személy tehát élő alanyra utal, sőt valójában mindig élő alanyt jelent, míg a 3. személy jelenthet élőt és élettelent is. (Ennek megfelelően nem véletlen, hogy sok nyelv csak a 3. személyű névmásokban különíti el az élőt és élettelent [3].) Ezen a megfigyelésen alapul a *harmadik-személy% (3sz%)* módszer, mely szerint ha az ige túlnyomó többségében 3. személyben fordul elő, akkor alanya élettelen, különben élő.

1. táblázat. Néhány jellemzően élő ill. élettelen alanyú ige *3sz%*-értéke

ige	*élőség*	*3sz%*-érték
néz	élő	65,4%
alszik	élő	64,0%
megtörténik	élettelen	99,9%
tartalmaz	élettelen	99,9%

Néhány jellemzően élő ill. élettelen alanyú ige manuális vizsgálata (1. táblázat) után az alábbi szabályt állítottam fel:

3sz%-módszer: 3. személy aránya $> 90\% \Rightarrow$ élettelen az alany

Ezt a kiinduló módszert egy 68 véletlenszerűen kiválasztott igei keretből álló kis korpuszon teszteltem, a kereteket előzőleg annotáltam az alany élősége szerint. Az eredményeket a 2. táblázat tartalmazza. A baseline nagyon magas: pusztán azáltal, hogy minden alanyt élőnek veszünk, az igék négyötödét helyes kategóriába soroljuk. A *3sz%* módszer ezt kis mértékben meghaladja, de a teljesítménye nem kielégítő.

2. táblázat. A *3sz%* módszer kiértékelése ($n = 68$). Mértékek: A – megfelelőség (vö: *accuracy*), azaz hogy milyen arányban döntött helyesen a módszer; valamint: P_I – élettelen pontossága, R_I – élettelen fedése, P_A – élő pontossága, R_A – élő fedése.

	A	P_I	R_I	P_A	R_A
3sz%	84%	57%	86%	96%	83%
baseline	79%				

A módszer főleg a kellemetlenebb irányba hibázott, azaz élő helyett élettelennek határozott meg bizonyos alanyokat. A hibák elemzésekor körvonalazódott egy olyan igecsoport, ahol annak ellenére, hogy ezek az igék lényegében kizárólag egyes szám harmadik személyben fordulnak elő, az alany egyértelműen élő (pl. *nyilatkozik*, *vélekedik*, *aláír*, *tárgyal vmiről*). Komlósy fenti állítása tehát ezen az empirikus alapon cáfolhatónak tűnik, a módszert pedig valamilyen módon finomítani szükséges.

4.3. A *k3sz%* módszer

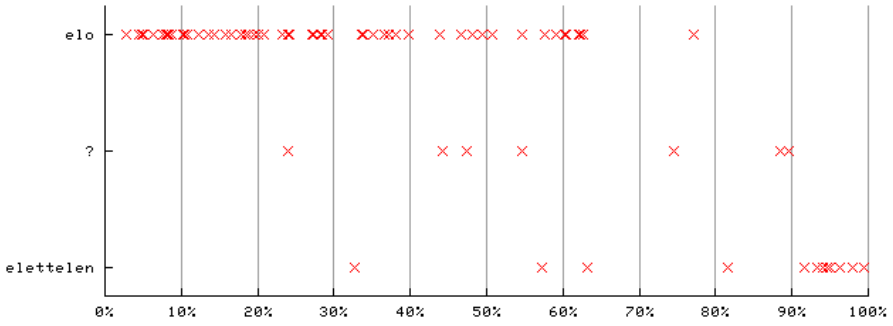
Mint említettük, az 1. és 2. személyű ragozás egyértelműen élő alanyt jelez, a továbbiakban a harmadik személyű mondatokkal foglalkozunk, itt kell megbecsülnünk az élő és élettelen alanyok arányát. Az alapötlet a következő: vannak olyan szópáraink, melyek funkciójukban azonosak, kizárólag abban különböznek, hogy az élő/élettelen jegy beléjük van kódolva: ilyen a speciális *aki/ami* vonatkozó névmás pár. Adott helyen pontosan vagy az egyik vagy a másik szerepel, és hogy melyik, az csakis a referált entitás élőségétől függ.

Ha egy pozíción nagy többségben van az *aki* névmás, akkor valószínűsíthetjük, hogy élő jegyű pozícióról van szó, másként fogalmazva az *aki/ami* arány értékes információval szolgálhat a pozíció élő/élettelen arányáról, annak közelítéseként fogható fel. Megjegyzendő, hogy ezen a ponton hallgatólagosan feltételeztük, hogy élő és élettelen dolgokra ugyanolyan arányban szoktunk vonatkozó névmással hivatkozni. A magyar nyelv sajátosságának megfelelően az *ami*-hoz

hozzá kell vennünk az *amely*-t és a *mely*-t, erre a háromelemű halmazra fogok egyszerűen *ami*-ként hivatkozni, ez fog szemben állni az *aki*-vel.

A *korrigált harmadik-személy% (k3sz%)* módszerben tehát az élettelen alanyok arányának becslését úgy finomítjuk, hogy a 3. személyű alanyok közül csak az *ami* összes alany pozícióban előforduló vonatkozó névmáshoz viszonyított arányának megfelelő számút tekintünk élettelennek, azaz az alábbi mértéket fogjuk alkalmazni:

$$\text{3. személy aránya} \cdot \mathit{ami\%} = \text{3. személy aránya} \cdot \frac{ami}{ami+aki}$$



1. ábra. A *k3sz%* értékek eloszlása a tanulókorpuszon. Minden pont egy igét jelöl. A felső sorban az élő, az alsó sorban az élettelen alanyú igék helyezkednek el. A középső sor azokat az igéket ábrázolja, melyek élő és élettelen alannyal is előfordulnak.

A már említett 68 igei keretet tartalmazó korpuszt tanulókorpuszként használtam fel, és ábrázoltam, hogy milyen a *k3sz%* értékek eloszlása az egyes kategóriákban (1. ábra). Az ábrán egyértelműen elkülönülnek az igék az alany élősége szerint: az élő alanyú igék lényegében 65% alatt, az élettelen alanyú igék lényegében 90% fölött helyezkednek el, a két érték között egy szinte üres sáv van, ahol csak néhány ige található. A 65 és 90%-ot döntési szabályként alkalmazva 5 ige esetén hibáznánk: a *kitesz vmit*, a *feltűnik*, a *kimarad vmiből* a *repül* illetőleg a *megváltoztat vmit* esetében is valójában olyan igei keretekről van szó, melyek természetes módon elképzelhetők élő és élettelen alannyal is. Ennek kapcsán felmerül az annotált korpusz megfelelőségének kérdése.

Ennek a „kézi” tanulási szakasznak a feladata az, hogy a *k3sz%* értékekhez döntési szabályt rendeljünk. Mivel semmiképp nem szeretnénk, hogy élő alanyt élettelenként osztályozzunk, a küszöbértéket magas értéken: 90%-ban állapítottuk meg. A 82% körül lévő élettelen alanyú igei keret outliernek tekinthető, a

küszöbérték leszállítása 80%-ra valószínűleg túltanuláshoz vezetne. A végső szabály tehát a következő:

k3sz%-módszer: 3. személy aránya · *ami%* $> 90\% \Rightarrow$ élettelen az alany

A *tanuló*korpuszon a módszer a 3. táblázatbeli eredményt adja. A módszer jelentősen túllépi a baseline-t, a kívántnak megfelelően csak abban az irányban téved, hogy élettelent néha élőnek mond (azaz a P_I és R_A értékeket 100%-on tartja), emellett az élettelen alanyok nagy részét (71%-át) felismeri. Az előző rész végén említett, lényegében kizárólag egyes szám harmadik személyben előforduló, mégis élő alanyú igéket a módszer helyesen osztályozza.

3. táblázat. A *k3sz%* módszer kiértékelése a *tanuló*korpuszon ($n = 68$). (Mértékeket ld: 2. táblázat)

	A	P_I	R_I	P_A	R_A
k3sz%	94%	100%	71%	93%	100%
baseline	79%				

4.4. A *k3sz%* módszer kiértékelése

Az éles teszteléshez egy nagyobb és megbízhatóbb korpuszt készítettem. Két független annotátor osztályozta a 383 véletlenszerűen kiválasztott igei keretet, a tanulókorpuszhoz hasonlóan három lehetőségből választhattak: az alany élő, az alany élettelen, az adott keret élő és élettelen alannyal egyaránt megfelelő. A 4. táblázat mutatja a különféle annotációk gyakoriságát.

4. táblázat. A tesztelőkorpusz annotációinak gyakorisága. Az annotátorok egyetértése $296/383 = 77\%$ volt.

db	*annotáció*
246	egyértelműen élő
59	élő ↔ mindkettő
18	egyértelműen mindkettő
22	élettelen ↔ mindkettő
32	egyértelműen élettelen
6	élő ↔ élettelen (azaz ellentmondás)

Az egyértelműen élőnek vagy élettelennek megjelölt kereteken lefuttatott tesztelés eredménye a 5. táblázatban látható. Az eredmény hasonló a tanulókorpuszon nyújtott teljesítményhez (vö: 3. táblázat), a baseline itt még magasabb. Egy esetben történt olyan hiba, hogy élő alany helyett élettelen jött ki: a tárgy nélküli *jelent* ige volt ez, a hibát egyértelműen az okozta, hogy a korpuszlekérdezésben az ige élettelen dominanciájú tárgyas formái elfedték a ritkább tárgyatlan változatot (ld: 4.1 rész).

5. táblázat. A *k3sz%* módszer kiértékelése ($n = 278$). (Mértékeket ld: 2. táblázat)

	A	P_I	R_I	P_A	R_A
k3sz%	95%	95%	63%	95%	100%
baseline	88%				

A meg nem talált 12 élettelen alanyú keret a következő: *sért vkit*, *minősül vminek*, *működik vmiben*, *rendben van vmi*, *emelkedik*, *készül vmiben*, *jut vkinek*, *jelentkezik vmiben*, *lesz vmikor*, *kiderül vkiről*, *elpusztul*, *sejtet vmit*. Az első 7 *k3sz%* értéke 80% fölötti, a *működik vmiben* keretet valószínűleg a *közre működik vmiben* élő alanyú keret fedte el. Az másik 5 keret pedig lehet, hogy ténylegesen élő alanyú (pl *lesz vki vmikor vhol*, *elpusztul*).

A megtalált 20 élettelen alanyú keret a következő: *vezet vmihez*, *kezdődik*, *kell vmihez*, *történik vkivel*, *következik vmiből*, *csökken*, *múlik vmin*, *megvalósul*, *létre jön vmi*, *véget ér vmi*, *épül vmire*, *kezdődik vmivel*, *szolgál vmire*, *irányul vmire*, *zajlik*, *keletkezik*, *kialakul vmiben*, *növekedik*, *fennmarad*, *zajlik vmiben*. Ezek valóban kizárólag élettelen alannyal állhatnak.

Gyakorlati célunk az egyértelműen élettelen alannyal járó keretek kiválasztása volt. A magyar-angol fordítórendszerben arra a számos igére is kénytelenek vagyunk meghagyni az alapértelmezett *élő* értéket, amelyek rendesen élő és élettelen alannyal is előfordulnak (pl. *kimarad vmiből*, *feltűnik*, *repül*, *megváltoztat-t*). Ilyen értelemben kettéosztva az igéket az egyik oldalra kerülnek az az egyértelműen élettelen alannyal járók, a másik oldalra pedig az összes többi. Ezzel a felosztással a teljes tesztelőkorpuszon a következő eredményt kaptam (6. táblázat).

A baseline szélsőségesen magas értéke abból adódik, hogy szinte minden igét élő alanyúnak vettünk (kivéve egyedül azt a 32 darabot, amit mind a két annotátor élettelen alanyúnak jelölt). Rosszabbnak tűnő értékeket kaptunk, de mindössze arról van szó, hogy 5 esetben „élő helyett" élettelen alanyt jósolt az osztályozó. A következő igékről van szó: *befolyásol vmit*, *előír vmit*, *sugall vmit*, *tilt vmit*, *erősödik*. Látható, hogy mindegyik természetszerűen járhat élettelen alannyal, ha éppen nem ez a gyakoribb használatuk.

6. táblázat. A *k3sz%* módszer kiértékelése ($n = 383$). (Mértékeket ld: 2. táblázat)

	A	P_I	R_I	P_A	R_A
k3sz%	95%	77%	63%	97%	98%
baseline	92%				

5. Összefoglalás, továbbfejlesztési lehetőségek, alkalmazás

Az ismertetett *k3sz%* módszer alkalmas az élettelen alanyú igei keretek nagy részének kiválasztására, miközben lényegében sosem téved abban az értelemben, hogy élő alanyú igét élettelennek határozna meg.

A módszer kiegészíthető egyéb jegyek vizsgálatával: élő alanyra utal például a felszólítómód használata. Szükséges azonban elválasztani az azonos alakú kötőmódtól, például egyszerűen a *hogy*-gyal kezdődő tagmondatok kiszűrésével. Míg a *megy* ige felszólítómódú alakjainak 75%-a, a *működik*-nek mindössze 10%-a van valódi felszólító tagmondatban.

Kézenfekvő, de jóval bonyolultabb módszer lenne az egyes szám harmadik személyű mondatok alanyi pozícióján megjelenő szavak kimerítő gyűjtése és élő/élettelen kategóriákba sorolása például a WordNet segítségével [4] vagy a szavak élőségének gépi tanulásával [5]. Éppen azt szándékoztam bemutatni, hogy erre nincs szükség, mert a fenti kevesebb erőforrást igénylő módszer is kielégítő eredményt ad.

A módszer minden bizonnyal egyéb nyelvekre is alkalmazható. Az első-második illetve a harmadik személy elkülönítése közvetlenül, az *aki/ami* párnak megfelelő szópárt pedig nyelvspecifikusan kell keresni, angolban a *who/what* megfelelőnek tűnik.

A módszerrel az igék tárgyának ill. egyéb bővítményeinek élőségi értéke is megállapítható. Hasonlóan kezelhető a predikatív melléknév alanya, esetleg birtok birtokosa is, ami magyarban szintén elmaradhat. Az élő alanyok azonosítása esetleg szemantikus taggelés alapját adhatja, amennyiben ez az ágens jó közelítése.

Az *aki/ami* arány mintájára bizonyos esetekben a nemek elkülönítése is megvalósítható: itt két kézzel kialakított szóosztály gyakoriságait lehetne vizsgálni. Illusztrációképpen a *lány,nő/fiú,férfi* arány a *megnősül* esetében 1/20, a *férjhez megy* esetében 108/2. Némely nem ennyire egyértelmű esetben is határozott eltolódás van az egyik nem irányába, a *zokog* esetén a fenti arány 25/9.

A leírt módszerrel megállapított alapértelmezett értékek a MetaMorpho magyar-angol fordítóprogram [8] lexikonjába kerülnek be. A rendszer szabadon elérhető, kipróbálható a `http://www.webforditas.hu` oldalon.

A kutatást a Magyar Tudományos Akadémia *Elnöki kerete* támogatta. Köszönet Munkácsy Dorottyának az annotálás elvégzéséért.

Hivatkozások

1. Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M.C., Wasow, T.: Animacy encoding in English: why and how. In: Proceedings of ACL Workshop on Discourse Annotation, Barcelona (2004)
2. Dahl, Ö.: Animacy and the notion of semantic gender. (1996)
3. Frawley, W.: Linguistic Semantics. Lawrence Erlbaum Associates (1992)
4. Orăsan, C., Evans, R.: Learning to identify animate references. In: Proceedings of ACL Workshop on CoNLL. (2001)
5. Øvrelid, L.: Towards robust animacy classification using morphosyntactic distributional features. In: Proceedings of EACL Student Research Workshop, Trento, Italy (2006)
6. Sass, B.: Igei vonzatkeretek az MNSZ tagmondataiban. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2006), Szeged (2006) 15–21
7. Komlósy, A.: Régensek és vonzatok. Strukturális magyar nyelvtan I. Mondattan (1992) 279–529
8. Tihanyi, L., Merényi, C.: A MetaMorpho fordítóprogram projekt 2006-ban. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2006), Szeged (2006)

VII. Pszichológiai vonatkozású fejlesztések

Az intencionalitás modul fejlesztése és alkalmazása történelmi szövegeken

Ferenczhalmy Réka[1], László János[2]

[1] Pécsi Tudományegyetem Pszichológia Doktori Iskola,
7624 Pécs, Ifjúság útja 6.
ferenczhalmy@gmail.com
[2] MTA Pszichológiai Kutatóintézete,
1132 Budapest, Victor Hugo u. 18-22.
laszlo@mtapi.hu

A tanulmány az NKFP 6/074/2005 számú pályázat támogatásával készült.

Kivonat: Kutatócsoportunk a NOOJ integrált nyelvelemző környezetben olyan algoritmusok fejlesztésével foglalkozik, melyek segítségével narratív pszichológiai tartalomelemzést végzünk. Az előadásban részletes ismertetésre kerül az intencionalitás modul felépítése és működése, találati pontossága és az alkalmazásánál felmerülő nehézségek természete. Ezt követi a modullal végzett vizsgálat bemutatása, melyet általános- és középiskolás tankönyvek szövegrészein végeztünk, melyek jelentős pozitív és negatív történelmi események leírását tartalmazták. Ezeket mint szociális reprezentációkat vizsgáljuk, melyek elképzelésünk szerint relevánsak a nemzeti identitás alakulásának és fenntartásának szempontjából. A vizsgálat során a saját és a másik csoportnak tulajdonított intenciót néztük, illetve ezen belül a siker és kudarc megjelenítését. Eredményeink a pozitív és negatív események eltérő intencionális mintázatát mutatják.

1 Bevezetés

Kutatócsoportunk pszichológiai narratív tartalomelemzéshez szótár alapú lokális grammatikákat fejleszt a NOOJ integrált nyelvelemző környezetben. Ennek célja olyan nyelvi kódok azonosítása és megragadása a szövegben, melyek pszichológiailag releváns implicit tartalmakat közvetítenek.

Megközelítésünk egyik kiindulópontja Bruner [1] elképzelése, aki a gondolkodás két alapvető formáját különítette el: a paradigmatikus, logikai-tudományos és a narratív gondolkodást. Előbbi absztrakt fogalmak és logikai műveletek használatán alapul, tartalmának meg kell felelnie a verifikáció kritériumának. Míg utóbbi a történetekben, elbeszélésekben való gondolkodást jelenti, melynek igazolásában az életszerűség és a valószínűség a döntő. A két forma más-más információk felfogására és szervezésére alkalmas, tehát vannak jelenségek, összefüggések, melyeket logikailag, és vannak, amiket történetekbe ágyazva tudunk megragadni. A nyelv – mint eszköz és forma –

használata több (pl. kulturális, társadalmi, egyéni) szinten meghatározott, ami szerepet játszik abban, ahogyan az egyén vagy egy csoport megalkotja saját világát. A hangsúly a konstrukció folyamatán van. Az információk az elbeszélés két különböző síkján jelennek meg: az első a deskriptív szint, ami a történések puszta leírását tartalmazza; a második pedig a pszichológiai szint, ami már az elbeszélő perspektíváját, viszonyulását és értékelését is közvetíti, azaz megjelennek a pszichológiai aspektusok: vágyak, intenciók, emóciók, stb., ami az események egyfajta interpretációját jelenti. Kutatásunk eme utóbbi feltárására irányul. Feltételezzük, hogy azon jellegzetességek azonosítása alapján, melyek mentén az egyén vagy egy csoport az elbeszélés által leképezi és megkonstruálja a maga valóságát, azaz ahogy a különböző aspektusokat megjeleníti vagy mellőzi ebben a folyamatban, információval szolgál az egyén személyiségére, pszichés működésére, illetve a csoportra nézve. [1]

1.1 Narratívum és identitás

A *Ki vagyok én?* illetve a *Kik vagyunk?* kérdések megválaszolása alapvetően, természetéből adódóan, csak narratívumok által képzelhető el. Az identitás tehát felfogható úgy is, mint az (élet)történetek folyamatosan újraszerkesztett összessége, amelyben a múlt (származás, gyökerek, életesemények) és jelen történetei alapján kirajzolódik a lehetséges jövő is. A személyes identitás az én stabilitását és folytonosságát jelenti az állandó változások hátterében, ennek egyik igen meghatározó elmélete Erikson (1968) nevéhez fűződik. Emellett beszélhetünk szociális identitásról is, ami Tajfel (1981) megközelítésében az egyénnek másokkal, egy csoporttal való azonosságára vonatkozik. Azáltal, hogy a személy azonosul a csoporttal, azaz átveszi a normáit és értékeit, a csoport tagjává válik, így az önmegerősítést és biztonságot nyújt számára. [3]

Jelen kutatás a szociális identitásra irányul, megközelítésünkben a történelemkönyveket mint a szociális identitás alakulásában meghatározó szociális reprezentációkat vizsgáljuk. Ezek a reprezentációk Moscovici (1961) elmélete alapján egy konstruktív folyamat által jönnek létre és meghatározzák a társadalmi valóságot (kulturális, társadalmi és egyéni szinten is), azaz hogy mi kerül be a társadalmi diskurzusba, illetve mi milyen jelentéssel bír. Az információ így nemcsak objektív ismeretekre terjed ki, hanem magába foglalja a jelentéstulajdonítást, kijelöli a lehetséges értékelési és viszonyulási módokat, szoros összefüggésben az aktuális csoportfolyamatokkal, érdekekkel és célokkal. A szociális reprezentációk tehát igen jelentősek a szociális identitás szempontjából. Ide kapcsolható Halbwachs teóriája is, aki kollektív reprezentáció helyett kollektív emlékezetről beszél, szintén kiemelve tehát a szociális meghatározottságot, illetve a konstrukció folyamatát. Moscovici megközelítésében fontos továbbá a familiarizáció folyamata, ami arra irányul, hogy hogyan válnak a tudományos ismeretek (az ő vizsgálatában pl. a pszichoanalízis, esetünkben a történettudomány) a mindennapi tudás részévé. [3]

A szociális identitás és reprezentációk összefüggését Assmann (1999) az emlékezet felől közelítette meg, megkülönböztetve kommunikatív és kulturális emlékezetet. Utóbbi a csoport eredetéig nyúlik vissza, felöleli tehát a csoport múltját, melyet történetek formájában hagyományoznak a csoport tagjai, kialakítva és fenntartva ezáltal a

csoport folytonosságát, identitását. A kommunikatív emlékezetben pedig, ami nagyjából nyolcvan évre, 3-4 generációra terjed ki, a közelmúlt és jelen eseményei jelennek meg, melyekben a kortársak osztoznak. Assmann szerint pl. egy trauma feldolgozásának folyamatához nagyjából nyolcvan év szükséges. A kommunikatív emlékezetben azonban nemcsak a közelmúlt emlékei konstruálódnak, hanem olykor a régmúlt történetek is előtérbe kerülnek és felülíródnak, akár mert valami új információ lát napvilágot, ami átszervezi a róluk való tudást, akár mert a csoport jelenében válik aktuálissá egy adott esemény, ami (akár a csoport aktuális szükségletei mentén) szintén vezethet a történet újraírásához. [3] „A történelmi események „affordanciái" (Liu-Liu, 2003), vagyis az általuk potenciálisan hordozott szimbolikus tartalmak és érzelmi azonosulási minőségek, az identitáspolitika és a társadalomban reálisan élő identitásszükségletek együttesen határozzák meg a történelem jelenbeli szociális reprezentációit."[1]

2 Az intencionalitás-modul

2.1 A modul ismertetése

Az intencionalitás tágabb értelemben mentális állapotok és szándékok tulajdonítását jelenti. A modul ezen belül kizárólag a szándéktulajdonításra terjed ki. A kiindulási pontot az igék jelentették. Bizonyos értelemben minden aktív, cselekvést leíró ige intencionális (pl. *Cikket írok.*), azonban a modul szempontjából csak azokat az eseteket vizsgáljuk, amikor az intenciót elsődlegesen jelöljük a szövegben. Ilyen értelemben ez lehet egy intencionális ige (pl. akar, tervez, szándékozik, stb.; *Cikket akarok írni!*), vagy lehet az igének egy intencionalitást implikáló esete (pl. a feltételes mód, ami nem konkrét cselekvést, hanem szándékot, vágyakat, stb. jelöl; *Úgy írnék egy cikket!*), illetve egyéb nyelvi kódok, melyeket alább részletesen ismertetek. Egy érdekes és ilyen tekintetben köztes csoportot képviselnek a beszédaktusok, mivel átmenetet képeznek a konkrét cselekvés és a pszichológiai sík között, ezeket a modulba nem vettük be. A célunk tehát nem a szándékos cselekvés, hanem magának a szándéknak a megragadása a szövegben.

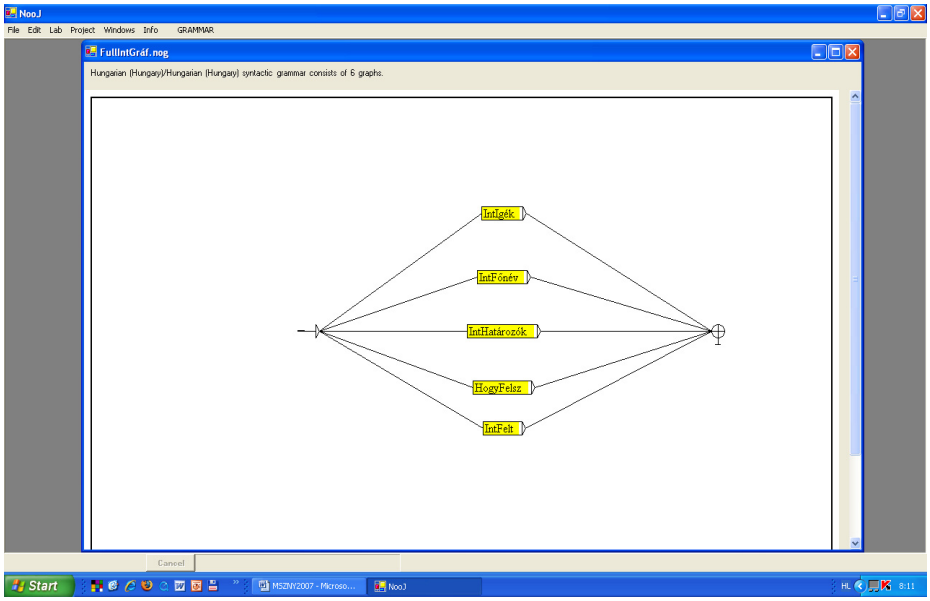
A modul kialakítása során a Todorov-féle igei transzformációkból indultunk ki, melyek közül az intencionalitás és az eredmény kapcsolódik ide. Az eredmény tekintetében nem minden eset tartalmaz intencionalitást, valami, például egy cikk megírása, nem várt eredményekhez is vezethet, melyek tehát cél és szándék hiányában nem tekinthetőek intencionálisnak. A modul szempontjából a cél elérésére vonatkozó esetek relevánsak, ilyen értelemben valaminek a sikere vagy kudarca már feltételezi a szándék meglétét.

A modul fejlesztése során a NOOJ integrált nyelvelemző környezettel dolgoztunk, ami szótárak és lokális nyelvtanok segítségével teszi lehetővé a kódok szövegbeli azonosítását.

[1] László, J. (2005) 188.old.

2.2 A modul felépítése

Az alábbiakban ismertetem a teljes intencionalitás gráfot, illetve az ezt alkotó algráfokat.

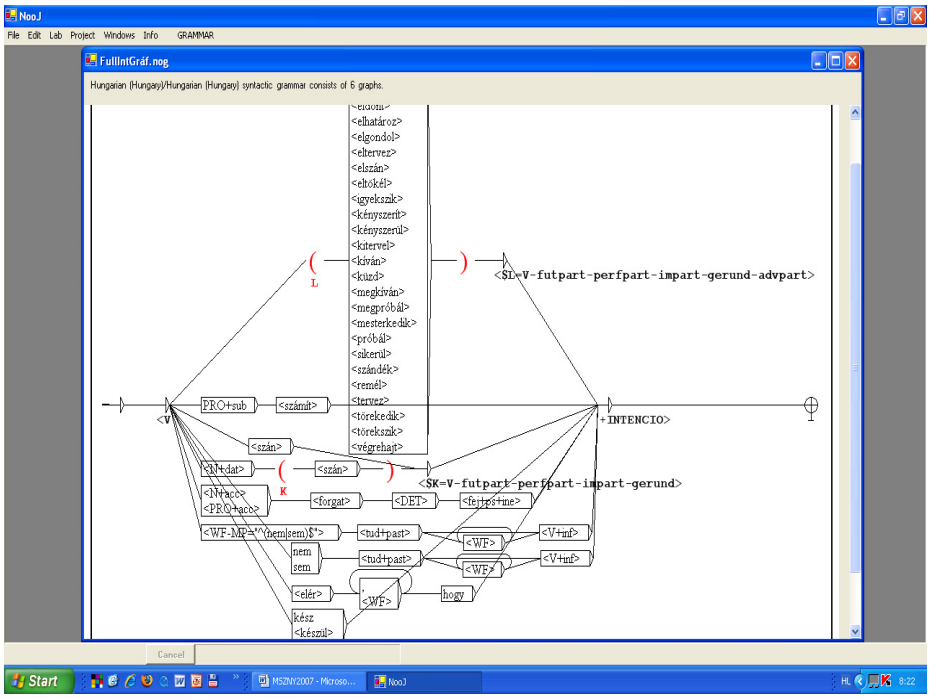


1. ábra Teljes intencionalitás gráf

Amint látható, a főgráf öt algráfból áll: 1, intencionális igék; 2, intencionális főnevek, 3, intencionális határozószók, melléknevek és névutó; 4, célhatározói mellékmondatok és 5, feltételes mód.

2.3 Az intencionális igék lokális nyelvtana

Az alábbi gráf tartalmazza az intencionális igék szótárát, illetve olyan idiómatikus és egyéb igei szerkezeteket, melyek bizonyos alakban intencionalitást fejeznek ki, pl. valaki a fejébe vesz valamit, vagy valamit viccnek vagy bántásnak szán valaki, az eredményekre vonatkozóan pedig hogy valamit meg tudott vagy nem tudott megcsinálni valaki, stb..

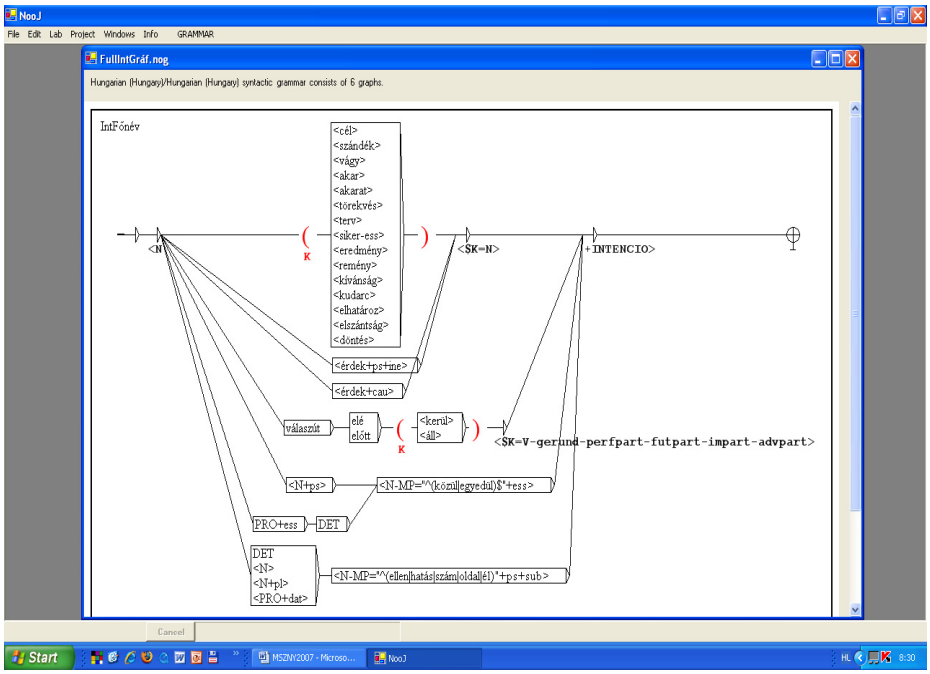


2. ábra Intencionális igék gráfja

2.4 Az intencionális főnevek lokális grammatikája

A gráf tartalmazza az intencionális főnevek listáját, illetve olyan szerkezeteket, melyek intenciót implikálnak. Láthatjuk például, hogy az '*érdek*' önmagában nem intencionális, pl. *nekem nem érdekem, hogy harcoljak* (ettől még lehet, hogy akarok), de ha *valaminek az érdekében tesz valaki valamit*, vagy *közös érdekekért harcolunk*, az már intencionális, mert a célt jelöli.

Hasonlóan, ha valaminek a jeléül tesz valaki valamit, pl. *hódolata jeléül átnyújtott egy csokor rózsát*, vagy *tiltakozása jeléül kirohant a szobából*, már megjeleníti az intenció tulajdonítását, egy értelmezése a helyzetnek. Ki kell zárni azonban a túl sok téves találatot eredményező *közül, egyedül* eseteket, melyeknek köre természetesen az a különböző szövegek elemzése során bővül. Az általunk elemzett történelemszövegekben ezek vezettek sok nem kívánt találathoz (ezt eredményezték, pedig nem akartam, tehát ez nem intencionális mozzanat). Ezt jelen esetben morfológiai szintű kizárással lehet megoldani.



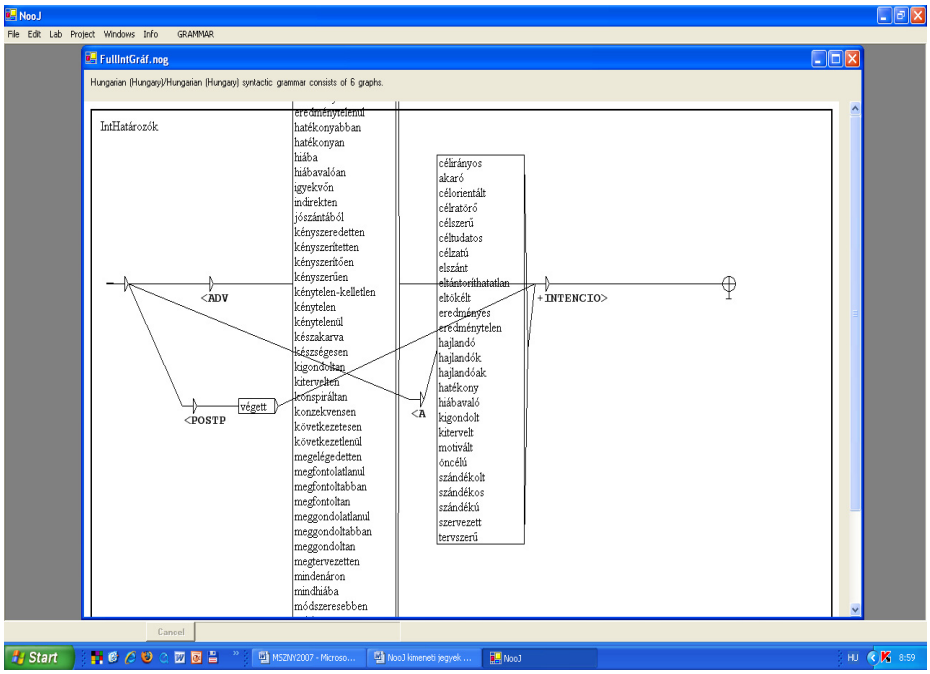
3. ábra Intencionális főnevek gráfja

2.5 Az intencionális módhatározószók, melléknevek és névutó lokális grammatikája

A gráf két szótárból ál, a módhatározószókéból és mellékneveké̋ből. Módhatározószók használata mellett bármely egyszerű cselekvést kifejező ige intencionálissá válhat. Pl. önszántából vagy kényszerűen, direkt vagy véletlenül, szándékosan, tudatosan, módszeresen, stb. tesz-e valaki valamit.

A melléknevek a főnevekhez kapcsolódóan hordozhatják ezt az információt. Pl. hogy egy cselekvés szándékos, kitervelt vagy véletlen, illetve például hogy sikeres, eredményes vagy eredménytelen volt-e.

Ebben a gráfban található a *végett* névutó, mely szintén a szándékot jeleníti meg, pl. *A kutatás végett írom ezt a cikket.*.

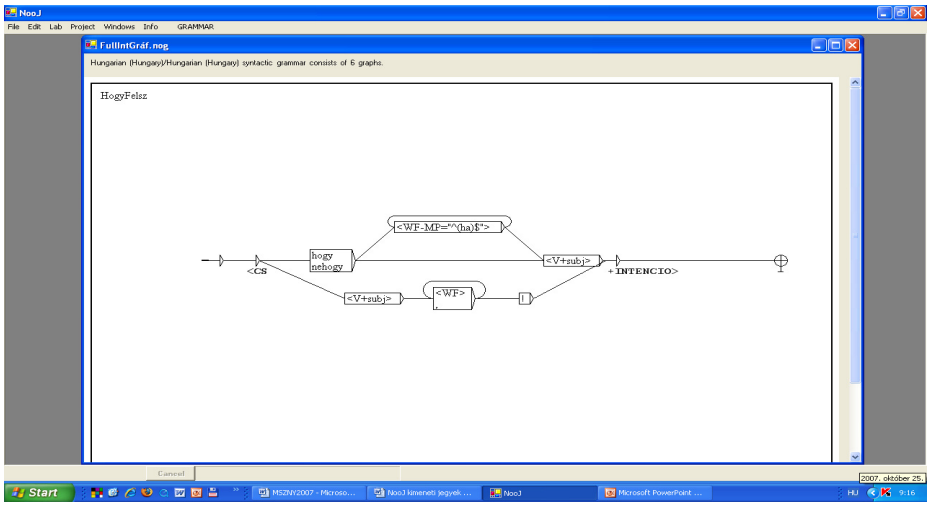


4. ábra Intencionális határozószók, melléknevek és névutó gráfja

2.6 Célhatározói alárendelő mondatszerkezet lokális grammatikája

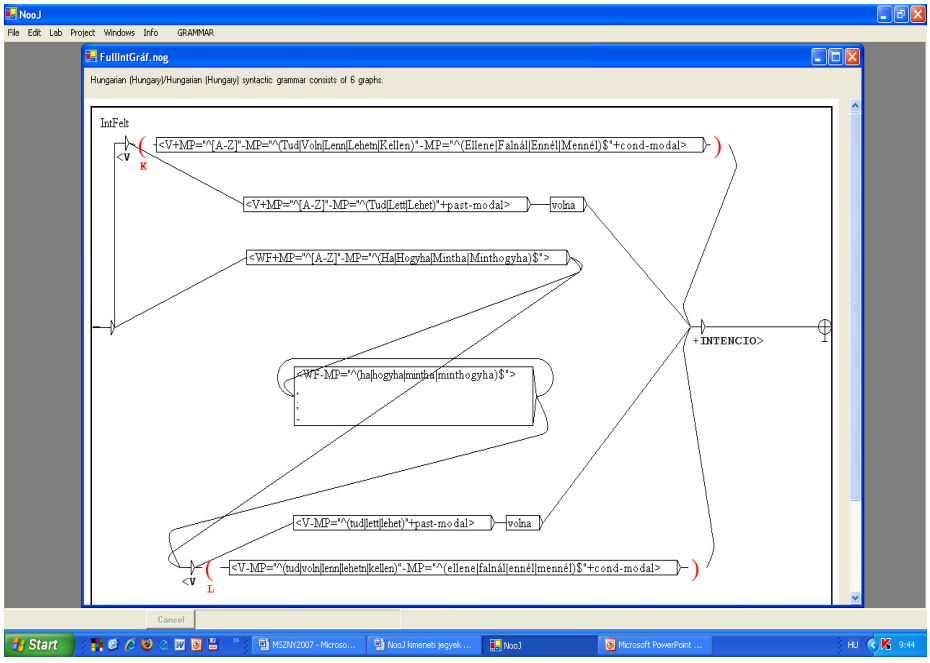
Ha megjelöljük a szövegben a konkrét cselekvés célját, pl. *Azért írom a cikket, hogy bemutassam a modulomat.*, az szintén intencionális. Ezt keressük a szövegben.

Ezt a szerkezetet csak az alábbi nagyon egyszerű szerkezettel, nagyon általános szinten tudtam megragadni. Célhatározói alárendelést jelez az *'azért'* kötőszó, de a magyar nyelv sajátossága és szépsége, hogy nem mindig használjuk. (*Azért jöttem, hogy beszéljek veled*. illetve *Eljöttem, hogy beszéljek veled.*) Ez a gráf sok téves találatot hoz, pl. a beszédaktusokhoz kanyarodva: *Megparancsolta/Megkérte, hogy maradjak.*, vagy az azonosságok miatt: *Lehetetlen, hogy ennyi kávé fogyjon a szünetben!*. Ezeket manuálisan kell kiszelektálnunk a találatok közül, mivel nyelvtanilag nem elkülöníthetőek. A téves találatok ellenére azonban nem tekinthetünk el ettől a gráftól, mert az intencionalitás tekintetében is gyakori a használata, tehát sok jó találattól is elesnénk.



5. ábra Célhatározói alárendelő szerkezet gráfja

2.7 A feltételes mód lokális grammatikája



6. ábra Feltételes mód gráfja

Ez a gráf a feltételes mód azon nem kívánatos találatait szűri ki, melyek valaminek csak a lehetőségét jelenítik meg, nem magát a szándékot. A *Ha most nem cikket írnék, kávét innék és beszélgetnék…* példa alapján még lehetne szó a vágy sejtéséről, de a *Ha most Afrikában lennék, nem innék ennyi vizet.* mondat kapcsán ez már nem merül fel. (A gráf részletesebb kifejtését ld. Ferenczhalmy és László, 2006.)

2.8 A gráf találati pontossága

11849 szóból álló szövegen lefuttatva a találatok a következők.

Kézi kódolással: 160 találat

NOOJ találatok:
Összes találat: 157
Összes jó találat: 136 – 85%
Téves találatok: 21 – 13%
NOOJ kihagyás: 24 – 15%

3 Történelemkönyvek narratív pszichológiai elemzése

3.1 A kutatás ismertetése

A kutatásban a történelemkönyveket mint szociális reprezentációkat vizsgáltuk, melyek lényegesek a nemzeti identitás alakulásában. László, Ehmann és Imre (2002) szintén tartalomelemzésen alapuló kutatása a jelentős pozitív és negatív történelmi események szociális reprezentációinak vizsgálatára irányult, melyet általános- és középiskolásokkal, illetve egyetemistákkal végeztek. Ez a vizsgálat kimutatta, hogy a pozitív események többsége a régmúltban történt, melyet azután hosszú negatív, a balsors által vezérelt időszak követett, ami a rendszerváltás óta megint kicsit pozitívba fordult. [4] Jelen kutatásban ezen események reprezentációját a történelemkönyvek elemzésével vizsgáljuk.

A kutatásban a jelenleg használatos, több kiadótól és szerzőtől származó általános- és középiskolás történelemkönyveket vizsgáltuk témakörök szerint. Így az intencionalitás tekintetében a tatárjárás és az ország újjáépítése, Hunyadi János, Mátyás király, Mohács, Trianon és a II. világháború került az eddigiekben feldolgozásra. A szándéktulajdonítást két csoportra nézve vizsgáltuk: a saját és a másik csoportra vonatkoztatva, a vizsgált eseménytől függően (ingroup-outgroup összehasonlítás).

A munka során a NOOJ program mellett az Atlas.ti-vel is dolgoztunk, lévén, hogy a NOOJ még egyelőre nem rendelkezik statisztikai programmal és az ágencia kezelhetőségének megoldása is folyamatban van, illetve a szövegek ellenőrző kézi kódolását is ezzel végeztük.

A kutatás során arra kerestük a választ, hogy a pozitív és negatív események leírásaiban, azaz a sikerek és kudarcok megjelenítésében van-e a szándéktulajdonításra nézve valami jellegzetes mintázat. Van-e eltérés, és ha van, milyen a saját és mások kudarcainak vagy sikereinek bemutatásában. Feltételezésünk szerint az identitás

szempontjából lényeges pozitív események esetében a saját siker és a szándékosság jegyei erőteljesen megjelennek a szövegben, míg negatív események esetén a saját intenció mértéke csökken, míg a másik csoporté nő.

3.2 Eredmények

Az eredményeket az alábbi táblázatokban foglalom össze az általános- és középiskolai szövegek elemzéséből kapott adatok alapján.

1. Táblázat: Eredmények 2/1

		∑ Intenció	∑ Saját	∑ Másik	Saját		Másik	
					∑ Siker	∑ Kudarc	∑ Siker	∑ Kudarc
Általános iskola	Tatár 6169 szó	90	65	25	5	4	4	5
		1,45%	46	16	∑9		∑9	
	Hunyadi 5334 szó	67	52	15	16	5	0	4
		1,25%	31	11	∑21		∑4	
	Mátyás 4098 szó	48	44	4	12	3	0	0
		1,17%	29	4	∑15		∑0	
	Trianon 3297 szó	45	22	22	1	4	0	1
		1,36%	17	21	∑5		∑1	
	II.VHáb 13869 szó	218	170	48	15	17	6	4
		1,57%	138	38	∑32		∑10	

2. Táblázat: Eredmények 2/2

		∑ Intenció	∑ Saját	∑ Másik	Saját		Másik	
					∑ Siker	∑ Kudarc	∑ Siker	∑ Kudarc
Középiskola	Tatár 8103 szó	105	73	23	7	8	0	2
		1,29%	58	21	∑15		∑2	
	Hunyadi 10442 szó	151	107	30	25	18	1	4
		1,44%	64	25	∑43		∑5	
	Mátyás 2487 szó	39	38	1	7	4	0	0
		1,56%	27	1	∑11		∑0	
	Trianon 5226 szó	70	37	35	1	7+7Neg	2	0
		1,33%	12	33	∑15		∑2	
	II.VHáb 13646 szó	174	111	63	13	10+7Neg	7	8
		1,27%	81	48	∑30		∑15	

Az eredmények csak részben támasztották alá a feltevéseinket. Azt láthatjuk, hogy László és munkatársai által leírt pozitív, a régmúltba visszanyúló eseményeknél a saját csoport intenciója jelenik meg mind a szándék, mind ezen belül a siker és kudarc tekintetében. A másik csoporté (tatárok és törökök) alulreprezentált.

Ezzel szemben a negatív eseményeknél megjelenik a másik csoport intenciója is, Trianonnál az intenció aránya a két csoport között ki is egyenlítődik. Lényeges kiemelni, hogy sikerek és kudarcok tekintetében is a saját csoport jelenik meg leginkább, míg a másik csoport, pl. mint győztesek, ezen a síkon szintén alulreprezentáltnak mondható. Szembeötlő ez Trianon esetében, amit lentebb részletesen is ismertetek.

3.3 Eredmények – Trianon szövegek elemzése

3.Táblázat: Trianon szövegek elemzésének eredménye

		∑ Intenció	∑ Saját	∑ Másik	Saját		Másik	
					∑ Siker	∑ Kudarc	∑ Siker	∑ Kudarc
Ált. isk.	Trianon 3297szó	44 1,36%	22	22	1	4	0	1
			17	21	∑5		∑1	
Közép isk.	Trianon 5226szó	70 1,33%	37	35	1	7+7Neg	2	0
			12	33	∑15		∑2	

Az eredmények alapján azt látjuk, hogy az intencionalitás tekintetében a két csoport egyenlő arányban jelenik meg. Azonban ha ezt a sikerek és kudarcok bontásában nézzük, akkor ezen a síkon, tehát a pszichológiai síkon, szinte csak a saját csoport reprezentálódik. A 'Neg' kifejezés a kényszer következtében bekövetkező cselekvést jelzi, tehát az intenció egy specifikus fajtájának tekinthetjük, az intenció hiányára utal.

A találatok számadatainak elemzésén túlmenően érdemes megvizsgálni a konkrét találatokat is az egyes csoportokat illetően.

Saját csoport: remény, aláírta volna, sikerrel (+neg), kénytelen-kelletlen, tekintette volna, célul, kívánták alakjai, sikerült, hogy hazánkban a nyugalom helyreálljon, mindhiába, eredménytelen, törekvést, nem akaró, célja, kerüljenek vissza, hiába.

Másik csoport: akarták alakjai, tervezettel, elhatározták, eldöntötték, hiába, hogy Mo elveszítse, kénye-kedvére, hogy a magyarok kisebbségbe kerüljenek, tekintették fő feladatuknak, hogy…akadályozzák meg, vette volna (figyelembe), tervezett, hajlandó, kísérletet sem tettek, kívánták, hogy azok körbevegyék, tudatosan, módszeresen.

Az aláhúzott kifejezések a kudarcra irányulnak, melyeknek eloszlása így is érzékelhetővé válik.

3.4 Eredmények megvitatása

Az eredmények kapcsolódnak ahhoz az elgondoláshoz, hogy a régmúlt eseményei, azaz a gyökerek, az identitás szempontjából döntőek, felelevenítésük annak kialakulását és fenntartását szolgálja, tehát a saját csoport intencionalitása, sikerei ebben az esetben nagyon erőteljesen megjelennek. Ez hozzájárulhat a büszkeség, kontroll, pozitív identitás érzéséhez.

A negatív események azonban, melyek esetünkben még nem a régmúlt eseményei, tehát a kommunikatív emlékezet szerves részét képezik, még feldolgozatlanok. A feldolgozást itt olyan értelemben nézzük, hogy az események megjelenése a narratívumban mennyire integrált és koherens, ami a jó történet szempontjából döntő kritérium. Azaz mennyire lehet ezeket az eseményeket objektíven, reflektíven szemlélni. Itt a hiányoknak, a „hallgatásnak" éppolyan jelentősége van, mint a megjelenő tartalmaknak. Ilyen tekintetben a negatív eseményeknél, pl. Trianon, mint nemzeti trauma megjelenítésénél azt látjuk, hogy még erőteljes hiányosságok vannak, a másik csoport szándékai, eredményei még nem tudnak megjelenni.

Bibliográfia

1. Bruner, J.: Valóságos elmék, lehetséges világok. Új Mandátum Könyvkiadó, Budapest (2005)
2. Ferenczhalmy, R., László, J.: Az intencionalitás modul kidolgozása NOOJ tartalomelemző programmal. MSZNY, Szeged (2006)
3. László, J.: A történetek tudománya. Új Mandátum Könyvkiadó, Budapest (2005)
4. László J., Ehmann, B., Imre, O.: Történelem történetek: a történelem szociális reprezentációja és a nemzeti identitás. Pszichológia 22, (2), 147-161. (2002)

Az érzelmek reprezentációja történelmi regényekben és történelemkönyvekben

Fülöp Éva[1], László János[2]

[1]PTE-BTK, Pszichológia Doktori Iskola
7624 Pécs, Ifjúság útja 6.
fulopeva1@freemail.hu

[2] MTA Pszichológiai Kutatóintézet
1132 Budapest, Victor Hugo u. 18-22
laszlo@mtapi.hu

A kutatásban célunk az, hogy történelmi szövegek elemzésével differenciált képet kapjunk a magyar történelemábrázolásban előforduló érzelmekről. Megvizsgáljuk a saját csoportnak és a külső csoportoknak tulajdonított érzelmeket, hogy azok alapján az identitás konstrukcióra következtessünk. Az érzelmek felismerésére a szövegben a NooJ nyelvtechnológiai rendszerben szótár alapú nyelvi algoritmusokat fejlesztettünk. A vizsgálatba hat történelmi regény valamint általános iskolai és középiskolai tankönyvek történelmi eseménnyel foglalkozó fejezeteit vontuk be. Az eredményeket az érzelmek és a csoport identitás közötti összefüggésekre vonatkozó modellekkel összevetve értelmeztük.

1. Elméleti háttér

1. 1 Történelmi elbeszélések, identitás, tartalomelemzés

Moscovici [3] szociális reprezentáció elmélete szerint a társas valóságot a személyek és csoportok számára a szociális reprezentációk jelentik. Ezek a reprezentációk gyakran narratívumok formájában jelennek meg és kijelölik az azonosulás formáit. [1]. A történelem narratívumokban kifejezett reprezentációja annak konkrét leírásán kívül (személyek, események, idő) tartalmazza a csoportok lehetséges érzelmi viszonyulásainak készletét is [4]. Ebben a vizsgálatban a történelmi elbeszélések érzelmi szerveződésének megjelenését várjuk.

1. 2. Szociálpszichológiai modellek

1. 2. 1 Infrahumanizáció

Leyens és mtsai. [2] a külső csoport hátrányos megkülönböztetésének érzelmi oldalát vizsgálják. 'Infrahumanizáció' elméletük szerint a másodlagos érzelmeket –melyeket

jellemzően az emberi esszencia alapelemének találtak- nagyobb mértékben tulajdonítják a saját csoport tagjainak, míg a külső csoporttagokat megfosztják azoktól.

Feltételeztük, hogy infrahumanizáció nemcsak laboratóriumi közegben, de a természetes nyelvhasználatban is tetten érhető, mint pl. történelemkönyvekben vagy csoportközi konfliktusokat tartalmazó történelmi regényekben *A magyar történelem jelentős-más nemzeteket is bevonó- eseményeit kiválogatva azt feltételeztük, hogy a történelem által meghatározott nemzeti identitás részét képező érzelmi reakciók, melyekkel a regények szereplőit felruházzák, változnak a történelem függvényében.*

Fő hipotézisünk az volt, hogy az infrahumanizáció jelenségének megléte függ a saját és a külső csoport kapcsolatától. Ezzel együtt erőteljesebb dehumanizáló tendenciát vártunk azon írások esetén, amelyek jelenlegi konfliktusokat, fennálló ellenségeskedést beszélnek el vagyis a magyar szerzők kevesebb szociális érzelmet tulajdonítanak majd ezen regények külső csoportjainak, míg ez a tendencia nem lesz észlelehető az időben távol levő eseményeknél, csoportoknál.

1. 2. 2 Az érzelmek valenciája

A magyarokat érintő traumákat lehetetlen pozitív érzelmekkel leírni, a negatív érzelmek megosztása szerves része a feldolgozási folyamatnak.

Hasonlóan az infrahumanizáció jelenségéhez az időben a jelenhez közel eső, még feldolgozás alatt álló történetek (két világháború) esetén jósolható intenzívebb saját csoporthoz tartozó érzelemkinyilvánítás, különösen a negatív tónusúaké. A külső csoportnál nem várunk kifejezett mintázatot.

1. 2. 3 Konkrétság-absztraktság

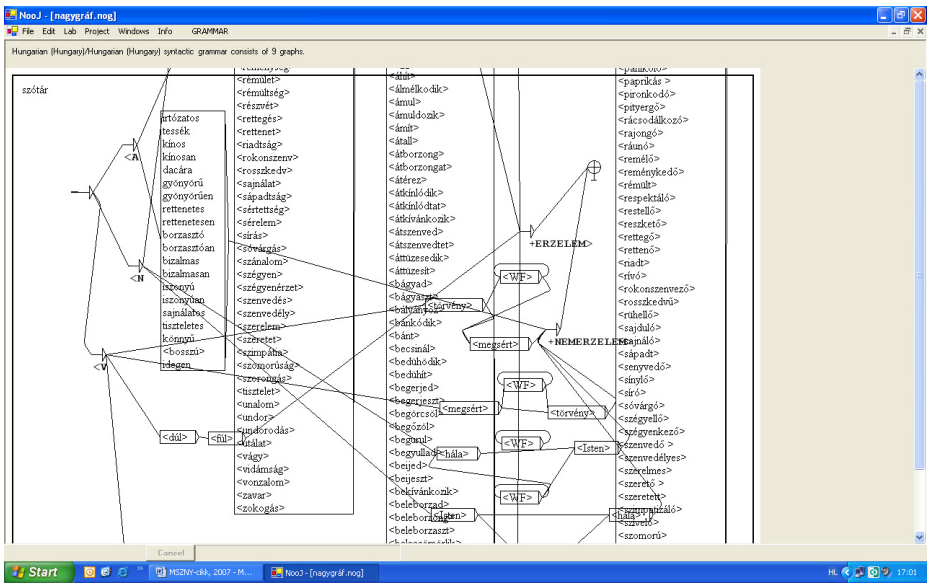
Az érzelemtulajdonításnak nemcsak tartalmi (alap-szociális), de formai különbségei is reflektálhatnak valamely szociálpszichológiai változóra. A nyelvi kategória modell alapján [5] feltételeztük, hogy a konkrét nyelvi kifejezések pszichológiai távolságot közvetítenek, míg az absztrakt alakok segítik az empátiát. A saját –külső csoport megkülönböztetést elősegítendő több absztrakt kifejezés tulajdonítását várjuk a saját csoportnál és a konkrétak többségét a külső csoportoknál. Az érzelmi közelségtávolság a nyelvi kategóriák használatának közvetett útján kívül közvetlenül is kifejeződik, hiszen egyes érzelmek önmagukban közelítést vagy távolítást implikálnak, itt is hasonló mintázatot várunk.

Végül a fenti kritériumokon kívül összegyűjtöttük minden regényben és a történelemkönyvekben a leggyakrabban előforduló érzelmeket.

2. NooJ

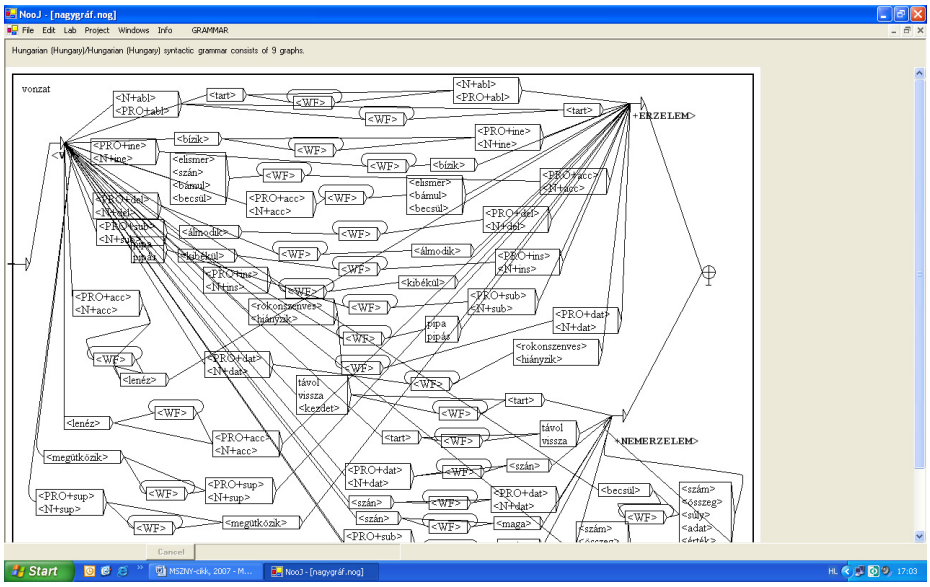
2. 1. Szótár

A nagy terjedelmű szöveg elemzésére kialakítottunk egy szótáralapú számítógépes programot. A szótár a magyar nyelv összes érzelmi jelentéssel bíró kifejezését tartalmazza, külön az igéket, főneveket, határozószókat, mellékneveket (1. ábra). Az érzelmi konnotációval bíró, ám nem közvetlenül érzelmet kifejező szavak pl. agresszió vagy- szorongásszavak külön kigyűjtésre kerülnek egy másik modulban.



1. ábra: Érzelemszótár

Ezt a szótárt beleillesztettük a NooJ nyelvelemző programba [6] és helyi nyelvtanokat alakítottunk ki a többjelentésű szavak és más nem megfelelő találatok kiszűrésére. Az egyes szavakon kívül, azokra a nyelvtani szerkezetekre, mintázatokra is nyelvtanokat építettünk, melyek szintén a helyes, érzelmet jelentő kifejezésekhez tartoznak. Például azon igék esetén, melyek csak bizonyos vonzattal hordoznak érzelmi jelentést (pl. ’ tart valakitől’) (2. ábra).



2. ábra: Vonzatkeretes érzelemszavak

Jelenleg a téves találatok kiszűrése és a saját-másik csoporthoz tartozó érzelmek elkülönítése kézi ellenőrzést igényel, melyet a NooJ és az Atlas.ti programok együttes használatával érünk el.

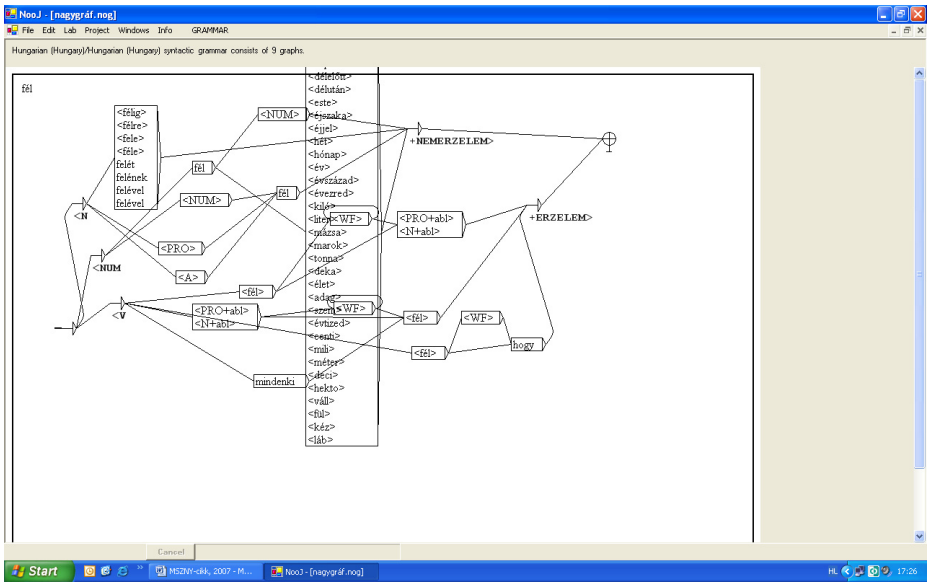
Az érzelem-gráf az eddigi munkák alapján kb. 80%-os pontossággal működik, ám ez a találati arány szövegenként változik, történelmi szövegek esetén kisebb, pszichológiai tartalomnál nagyobb is lehet. Sok esetben az ’unknown’-ok vagyis annotálatlan szavak rontják ezt az értéket.

2. 2. Problémák

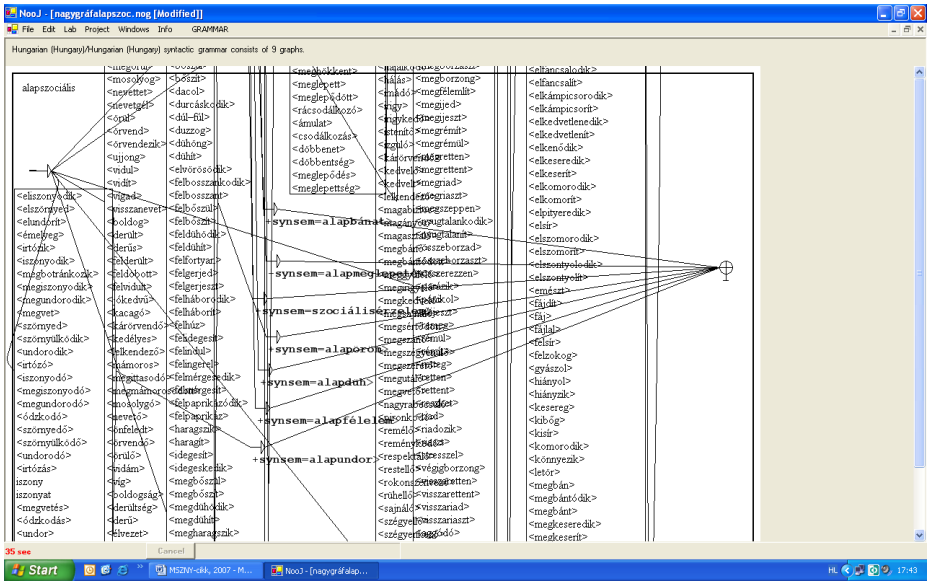
A téves találatok kizárásának elégtelensége miatt néhány esetben úgy oldható meg a szűrés, hogy pl. a többjelentésű szavaknál két kódot kap az adott szó, egyet az érzelmi jelentést közvetítő nyelvi kontextusban, egyet pedig a nem érzelmiben, így a kettős találatoknál a ’nemérzelem’ kimenetelű eredmények automatikus kizárásával nyerjük ki a valódi érzelemtalálatokat. (3. ábra)

2. 3. Kategóriák

A szótárt négy dimenzió mentén kategorizáltuk, az egyes kategóriákat szintén beépítettük a NooJ program érzelem-gráfjába. (4. ábra)



3. ábra: Példa a kizárásra (’fél’)



4. ábra: Kategóriák az érzelem-gráfban, itt például az alap-szociális érzelmek

1. Az első az érzelmek valenciája volt, ami lehet pozitív (pl. öröm), negatív (pl. félelem) vagy semleges (pl. meglepődés). 2. Az érzelmi minőséget négy alcsoport alkotja: az affektusok, érzések, érzelmek és az érzelmet implikáló cselekvések. Az affektusok az aktivációs kontúr változásával járó érzelemi folyamatok (pl. lenyugszik), az érzések alcsoport az általános hangulati állapotokat, nem specifikált érzéseket foglalja magába. (pl. jókedvű, nyugodt). Az érzelem címkével illetett csoport azokra az érzelmekre utal, melyek a történések kognitív kiértékelésével keletkeznek (pl. szégyen). Az utolsó alkategória az érzelmi minőségek közül az érzelmet implikáló cselekvéseké (pl. mosolyog, sír). 3. A harmadik csoportosító dimenzió az alap, vagyis elsődleges (pl. szomorúság) és a társas, vagyis másodlagos érzelmeket (pl. bűntudat) jelenti. 4. A negyedik dimenzió az érzelmi távolság kifejezéseiről szól, melyben három lehetőség szerepel: személyközi közelség (pl. szeretet), érzelmi távolodás (pl. elhidegül) és a pszichodimanikailag mindkettőt magába foglaló szavak (pl. harag).

3. Kísérleti anyag

Regények: Kós Károly: *'Az országépítő'* (1934) (I. István és az államalapítás történetét meséli el), Gárdonyi Géza: *'Egri csillagok'*(1899), Fekete Sándor: *'Kossuth Lajos'* (1970), Cseres Tibor: *'Vízaknai csaták'*(1988) (egy erdélyi magyar-román család története a XIX-XX-sz. fordulóból), Kálnay Adél: *'Háborús történet'* (2005) (egy I. világháborúban orosz hadifogságba eső férfiről szól) és Kertész Imre: *'Sorstalanság'*(1975). Vagyis a külső csoportok: németek, törökök, osztrákok, románok, oroszok és náci Németország.

Történelemkönyvek: Általános és középiskolás történelemkönyvek a Magyar történelem jelentős eseményeit tartalmazó részeinek Ehmann Bea által válogatott gyűjteménye jelentette a kísérleti anyagot.

4. Eredmények

4. 1. Elsődleges és másodlagos érzelmek gyakorisága

A hat regényből négyben a saját és külső csoportnak tulajdonított érzelmek megközelítőleg azonos mértékben jelentek meg. A *Háborús történetben* szignifikáns különbség volt a két csoport között: több szociális érzelmet kapott a saját csoport. ($\chi^2(1)= 3, 75$ $p =0,05$). Hasonlóképpen a *Sorstalanságban* szignifikánsan több társas érzelem jelent meg a saját csoport vonatkozásában. ($\chi^2(1) = 4, 48$ $p< 0, 05$).

Az egyes érzelmeket illetően a történelemkönyvekben az 'öröm' szignifikánsan többségben volt a külső csoportnál, mint a sajátnál ($\chi^2(1)= 3,91$ $p< 0, 05$), míg a 'félelem' sokkal jellemzőbb volt a saját csoportra ($\chi^2(1)= 4,2$ $p< 0, 05$). A leggyakoribb érzelem magyarokra vonatkoztatva, szignifikánsan eltérően a külső csoporttól a 'remény' érzése volt ($\chi^2(1) =7,4$ $p < 0,01$).

4. 2. Az érzelmi minőségek gyakorisága

A történelemkönyvek elemzése szignifikáns eredményt hozott. Magyarok esetén megbízhatóan több érzelem fordult elő ($\chi^2(1)= 4,8$ $p < 0, 05$) és kevesebb érzés ($\chi^2(1)= 4$ $p < 0,05$). Egyik regényben tendenciaszintű eltérést találtunk az érzelmeket implikáló cselekvések tekintetében a csoportok között a másik csoport javára ($\chi^2(1) =3,3$ $p = 0,07$).

4. 3. Pozitív, negatív és semleges érzelmek gyakorisága

Ebben a dimenzióban két regényben lehetett megfigyelni szignifikáns eltérést. A *Sorstalanságban* mind a pozitív, mind a negatív érzelmek tekintetében különbözetek a csoportok. A külső csoport szereplői több pozitív, a saját csoport tagjai több negatív érzelmet mutattak a könyvben ($\chi^2(1)= 7, 3$ $p< 0, 01$, $\chi^2(1) = 4, 4$ $p< 0, 05$). A *Háborús történetben* összehasonlítva a saját csoporttal, több pozitív érzelmi történést jelenítettek meg a külső csoportbeliek ($\chi^2(1)= 4,4$ $p < 0, 05$).

4. 4. Az érzelmi távolság kifejezéseinek gyakorisága

A történelemkönyvekben saját csoport különbözött a többi külső csoporttól érzelmi távolság mutatóiban: a szerzők szignifikánsan több távolító nyelvi elemet társítottak nekik ($\chi^2(1) = 3, 9$ $p < 0, 05$). Hatból kettőben, az *Egri csillagokban* és a *Háborús történetben* az érzelmi közeledést és távolodást egyszerre kifejező szavak (nagyrészt ez a ‘haragot’ jelenti) többségben voltak a külső csoportnál ($\chi^2(1) = 13, 48$ $p < 0, 01$; $\chi^2(1) = 6, 4$ $p= 0, 01$).

5. Megvitatás

A fenti kísérletben a NooJ programban kifejlesztett érzelem-gráf gyakorlati alkalmazásának egy lehetséges példája került bemutatásra.

Fő hipotézisünk az infrahumanizációra vonatkozóan teljes alátámasztást nyert az adatok által. Az infrahumanizáció jelensége sokkal komplexebb folyamatnak tűnik, melyet sok szociálpszichológiai tényező befolyásol, mint pl. a saját és külső csoport közötti jelenlegi kapcsolat, az időbeli távolság a másikcsoport- vonatkozású eseményektől, érdekellentétek megléte, stb. Jelen kutatás a jelenség szelektív jellegére hívja fel a figyelmet. Eredményeink szerint egyazon külső csoport reprezentációja- beleértve az érzelmi összetevőt- változhat időről időre függően a csoportközi és történelmi helyzettől.

A dehumanizációnak több funkcionális értéke lehet konfliktus esetén, mint megbékéléskor vagy együttműködéskor és valószínűbb, hogy kevésbé feldolgozott élmények esetén tűnik fel, melyek még kevéssé összeállt, kikristályosodott narratívummal rendelkeznek. Ebben a kutatásban négy különböző évszázad eseményeit vizsgáltuk, és az infrahumanizációs tendencia csak a XX. századi regényekben, a *Háborús történetben* és a *Sorstalanságban* jelent meg. Az I. és a II. világháború

élményei még mindig érzelmileg involválóak, a külső csoport tagjainak társas érzelmektől való megfosztása része lehet a nemzet ezen veszteségeinek feldolgozásában.

A 'remény' és a 'félelem' érzelmének dominanciája és az 'öröm' minimális jelenléte összeillesztve, egy olyan nemzet reprezentációját mutatja, melynek történelme küzdelmekkel és kudarcokkal telített.

A vizsgálatban érzelemtulajdonítás többletet találtunk a történelemkönyvekben a saját csoport részére, ami absztrakt kifejezésre utal, ami az empátia érzését támogatja a nemzetre vonatkozóan.

A negatív érzelmek megosztása, kommunikációja kulcsfontosságú lehet a feldolgozás szempontjából. A vesztes oldalon állva tehát, nem meglepő, hogy a saját csoport több negatív érzelmet és kevesebb pozitívat kapott, hasonlóan az infrahumanizációhoz a két XX. sz-i vagyis még még feldolgozás alatt álló eseményt elbeszélő regényben.

A történelemkönyvekben a magyarok hajlamosabbak érzelmi távolságot tartani, pszichésen elmozdulni a másiktól. A külső csoporttól való különbözőség hangsúlyozása a nagyobb távolsággal segítheti a szelf-kategorizációt.

Bibliográfia:

1. Breakwell, G. (1993) Social Representation and Social Identity In : Papers on Social Representation 2., 3.pp.198-217

2. Leyens, J-P., Paladino, P. M., Rodriguez- Torres, R., Vaes, J. Demoulin, S., Rodriguez-Perez, A., Gaunt, R.(2000): The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups, *Personality and Social Psychology Review*, Vol.14., No.2, 186-197

3. Moscovici, S. (1961). La psychoanalyse, son image et son public. Presses Universitaires de France, Paris.

4. Rimé, B. & Christophe, V. (1997). How individual emotional episodes feed collective memory. In J. W. Pennebaker, D. Paez & B. Rimé (Eds.). Collective memory of political events: Social and psychological perspectives (pp. 131-146). Hillsdale, NJ: Erlbaum.

5. Semin, G. R. (2007). Implicit indicators of social distance and proximity. In K. Fiedler (Ed.), Social Communication: Frontiers of Social Psychology (pp. 389-409). New York: Psychology Press.

6. Silberztein, M. : NooJ Manual: a Linguistic Annotation System for Corpus Processing. 2006

Történelemkönyvek és az idő viszonya: beszámoló a NooJ program segítségével végzett tartalomelemzéses vizsgálatokról

Garami Vera[1] és Ehmann Bea[2]

[1] PTE BTK Pszichológiai Doktori Iskola, 7624 Pécs, Ifjúság útja 6.
garamivera@yahoo.com
[2] MTA Pszichológiai Kutatóintézet, 1132 Budapest, Victor Hugó u. 18-22,
ehmannb@mtapi.hu

1 Bevezetés

Az időélmény pszichológiájának tartalomelemzéses vizsgálatairól korábbi publikációk számolnak be. [1, 2, 3, 6]. A jelenlegit közvetlenül megelőző kutatások [4] a NooJ program segítségével a szubjektív időélmény lényeges aspektusainak nyelvi markereit próbálták megragadni. Az egyes szókategóriákat a Korpusznyelvészeti Osztály munkatársaitól kapott, a tíz-tízezer leggyakoribb magyar igét, határozószót és melléknevet tartalmazó szólistákból állítottuk össze. [10]

Az eddig kidolgozott időkategóriáknak most egy lehetséges alkalmazását szeretnénk bemutatni, amelynek során általános iskolás történelemkönyvek szövegeit elemeztük narratív pszichológiai szempontok szerint, a csoportidentitás alakulását vizsgáló kutatás keretében. Úgy gondoljuk, hogy ezek a szövegek forrásai és lenyomatai lehetnek a magyar történelem eseményeiről kialakított szociális reprezentációknak.[6]

2 Narratív idő a történelemkönyvekben

A történetírás elbeszélés, de jellemzői eltérnek a hétköznapi vagy a fikciós narratíváktól. Az idő dimenzióját a történészek tudatosan is alakítják, foglalkoznak a korszakolással, kronologikus sorrendbe rendezik az eseményeket, megemlítik a pontos dátumokat.

Jean Leduc [7] "*A történészek és az idő"* című könyvében különböző szempontok mentén tekinti át a történetírás és az idő viszonyát. Foglalkozik a történelemtanítás és tanulás kérdésével, és tankönyvek szövegeinek elemzésével is.

Gerard Genette nyomán bevezeti a *„narratív tempók"* fogalmát. Ez alatt azt érti, hogy egy eseménynek a történetben való megjelenítése milyen arányban áll az esemény eredeti idejével. Ennek alapján négyféle narratív tempót különít el:

1. **ellipszis**: a szöveg átugrik a valóságban megtörtént időmennyiséget Pl.: „Két évvel később",
2. **összefoglalás**: az elbeszélés hirtelen felgyorsul,
3. **jelenet**: úgy érezzük, hogy az elbeszélés ideje arányos a cselekvés valós idejével,

4. **szünet**: az elbeszélés megáll – beilleszt egy leírást vagy kitekintést a narratív folyamatba.

Leduc történelmi szövegek elemzésénél vizsgálja az elbeszélés idejét az igeidők, és más „időjelölők" azonosításával. Megkülönbözteti az abszolút és relatív időjelölőket. Az abszolút időjelölők pontos dátumokat, évszámokat és korszakhatárokat jelentenek, a relatív időjelölők pedig a szövegben megjelenő utalások az események időbeli sorrendjére (előtte, közben, utána, stb.).

Leduc azt kutatja, miben különbözik a fikció és a történetírás ideje. Úgy találja, hogy az időjelölők a jelenetben a legritkábbak, és ez a forma nem igazán jellemző a történetírás narratív idejére. Mivel gyakoribbak a rövid és hosszabb periódusok leírásai, az időjelölők nagyobb arányban fordulnak elő, mint a fikciós szövegekben. Ezen belül a hosszabb időtartamokat átfogó elbeszélésekben az abszolút időjelölők aránya nagyobb, szemben a relatívakkal.

Korábbi kutatásunkban [4] beszámoltunk arról, hogy az idő nyelvi markereit tartalmi és funkcionális kategóriákba soroltuk. Leduc időjelölői az általunk tartalminak nevezett kategóriába tartoznak, és jól megragadhatónak bizonyultak a NooJ program segítségével.

A különböző "narratív tempók" vizsgálata is érdekes lehet számunkra. A 4. pontban felsorolt "szünet" nyelvi markere lehet például, ha a jelenre történik utalás a szövegben, ami végig múlt időben játszódik. Leduc hosszan elemzi a jelen idő megjelenését és eluralkodását a (francia) történetírásban, és ennek lehetséges okait. A mindenhol jelenben írott szöveg nálunk még nem terjedt el, valószínűleg más a jelentése, használata, mint a francia nyelvben, de az elbeszélés egységes múlt idejének megszakadása, a jelenre való utalások az általunk vizsgált szövegekben is megjelennek, és ezek fontos jelentéseket hordozhatnak számunkra. A lineáris idő megszakításának oka lehet, hogy a szerző kissé el akar távolodni az elbeszélt eseményektől, és ezért közbeszúr leíró részeket, például kulturális tények ismertetését. Ugyanakkor előfordul a jelen idő alkalmazása az események drámaiságának fokozásakor is. A múltból a jelenbe váltás ilyenkor a „felgyorsulás" érzetét keltheti az olvasóban. Természetesen más a tudatosan konstruált történelemkönyv szövege, mint az interjúban vagy másutt megjelenő énelbeszélés, de fontos itt megemlítenünk, hogy az önéletrajzi emlékezettel foglalkozó, illetve narratív pszichológiai kutatások szerint a múlt idejű elbeszélésben történő igeidő váltás érzelmi átélést [8,9], illetve a traumatikus események elbeszélésekor a feldolgozás, távolságtartás, lezártság hiányát jelezheti [2].

A történetírás egyik alapvető kérdése, hogy a hosszabb távú folyamatokat, ciklusokat, visszatéréseket és összefüggéseket, vagy az egyedi eseményeket hangsúlyozza-e. Kutatásunkban arra keressük a választ, hogy a történelemkönyvekben a hosszabb távú folyamatok, illetve a konkrét, rövid lefolyású, meghatározott, körülírható időben lejátszódó események ideje hogyan tér el egymástól. Feltételezhető, hogy a rövid időben játszódó, eseményszerű történetírás közelebb áll az önéletrajzi emlékezet személyes idejéhez, így átélhetőbbé válik az olvasó számára.

A fentiek alapján kutatásunk kérdései a következők voltak:

1. Ha a rövid időszakok leírásai kevesebb időjelölőt tartalmaznak, melyek lesznek azok az események a magyar történelemben, ahol az elbeszélés több, illetve kevesebb időjelölő kifejezést használ?

2. Mennyire jelenik meg a történelemkönyvekben a szubjektív időélményt kifejező funkcionális időkategória? (huzamos-pillanatnyi, gyors-lassú, kezdet, befejezés, ismétlődés) Milyen az aránya a funkcionális és tartalmi időkategóriának?

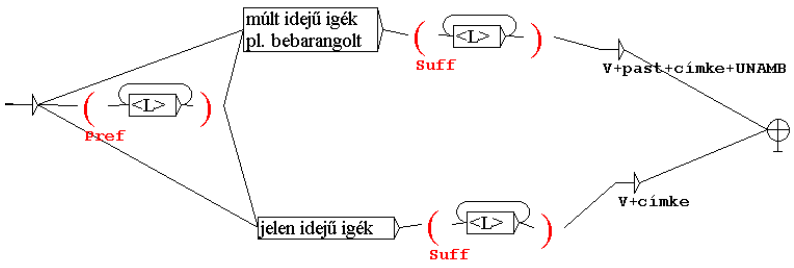
3. Milyen arányban fordulnak elő a tartalmi kategórián belül az abszolút és a relatív időjelölők? Megjelenik-e különbség az abszolút és a relatív időjelölők arányában, az elbeszélt események időtartamától függően? (Hosszabb időtartamnál több abszolút időjelölő megjelenését vártuk).

4. Megjelenik-e, és milyen mértékben a jelen idő a történelemkönyvek szövegében? Hol, milyen eseményeknél gyakoribbak a jelenre való utalások?

3 Módszer

A vizsgálatban hat különböző kiadónál megjelent általános iskolás történelemkönyv anyagából állítottuk össze a szövegkorpuszt. A négy osztály anyagának tankönyveiből 16 magyar történelmi eseményt választottunk ki, és az ezekre vonatkozó szövegrészeket tettük alkalmassá a nyelvi elemzésre. Így összesen 94 szöveges dokumentumot kaptunk (minden esemény esetében 6 szöveg, kivéve kettőt, ahol 5). A történelmi események, illetve időszakok a következők voltak: Honfoglalás, István uralkodása, a Tatárjárás, Hunyadi János háborúi, Mátyás király, Mohács, Végvári háborúk, Rákóczi-szabadságharc, 48-as forradalom, Kiegyezés, I. Világháború, Trianon, II. Világháború, Holocaust, 1956, és a Rendszerváltás.

A szövegeket a Nooj morfoszintaktikai elemző programba vittük be, és először elvégeztünk rajtuk egy lexikai elemzést, amelyhez a szótárt az MTA Nyelvtudományi Intézetével együttműködésben, Váradi Tamás és munkatársai biztosították. Ezután következne szintaktikai gráfoknak a futtatása, amelyek az általunk vizsgált időkategóriáknak megfelelő konkordancia találatokat képesek megadni. Ahhoz, hogy ezek a gráfok kevesebb téves, és több valóban használható találatot adjanak, először a lexikai elemzést bővítettük ki saját, a Magyar Nemzeti Szövegtár alapján készült szótárainkból vett, időbeliséget kifejező igékkel és határozószókkal, amikor ez szükséges volt. Ehhez morfológiai gráfokat készítettünk. Természetesen az adott korpusznak megfelelően igyekeztünk a szótárakat módosítani, alakítani. Az 1. ábrán látható egy igéket felismerő morfológiai gráf vázlata.



1. ábra. Gráf a lexikai elemzés bővítéséhez

A gráf felső ágán látható, hogy igyekeztünk egyértelműsíteni a múlt idejű igealakokat, amelyre jelen idejűség kiszűrésére készített szintaktikai gráf későbbi megfelelő működése miatt volt szükség. Ez a gráf már eleve hozzáad a megtalált szavakhoz egy címkét, pl. „gyors", és ezután a szintaktikai gráf, amely a funkcionális időkategóriákat, és ezen belül a gyors szavakat keresi, úgy is megtalálja, hogy <V+gyors>, így nem kell az egész szótárat beilleszteni a „gyors" időkategória gráfjába.

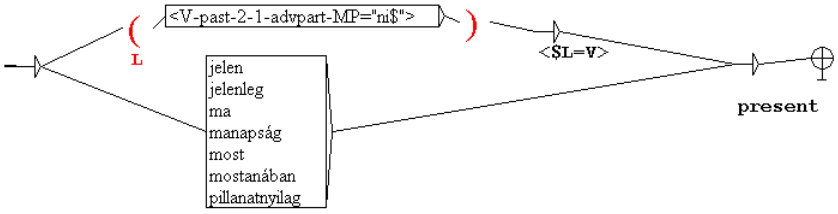
Vizsgálatunkban a következő szintaktikai elemző gráfokat készítettük el és futtattuk le a szövegeken:

1. **Funkcionális időkategória**. A gráf magában foglalja algráfként, a huzamos, pillanatnyi (vagy pontszerű), gyors, lassú, ismétlődés, kezdet, befejezés gráfokat. A konkordanciában több címkét is kap a találatunk: pl. „funkcionális"+ „huzamos".

2. **Tartalmi időkategória**. Tartalmazza az idő léptékével kapcsolatos szavakat, az „abszolút időjelölőket", vagyis a különböző dátumokat, és a relatív, előidejűségre (előtte, hajdan…), utóidejűségre (utána) és egyidejűségre (ugyanakkor, aznap) utaló kifejezéseket.

3. **Jelen idő**. A múlt idejű igék lexikai szótárral történő azonosítására alapul a jelen idő gráfja. Számos nehézségbe ütközik a jelen idejűség automatikus megtalálása, megoldandó feladat például a függő beszéd, vagy a főnévi igenevek kiszűrése. (2. ábra)

4. Külön gráfot készítettünk az „**ellipszis**" jellegű narratív tempó azonosítására is. Ebből azonban nagyon kevés találatot nyertünk, így inkább a tartalmi időkategóriába soroltuk be az „egy évvel később" típusú fordulatokat.



2. ábra. A jelen idő gráfja

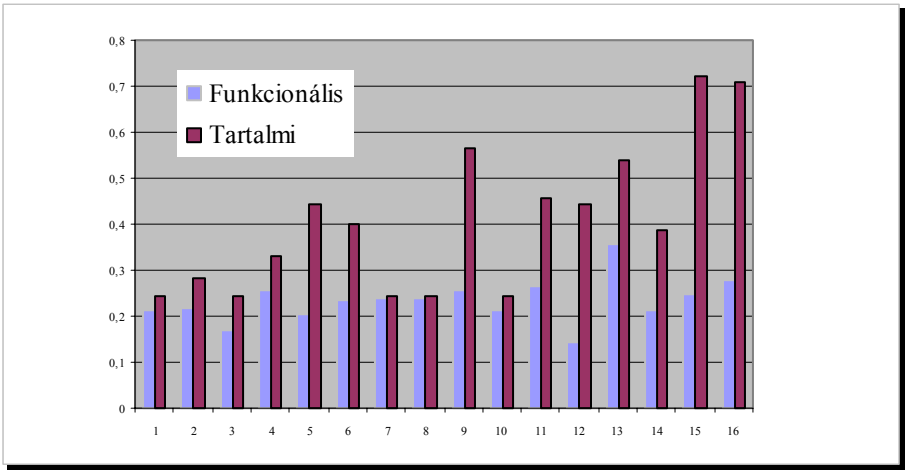
A futtatás után minden szövegnél átnézzük a konkordanciát, ebből fény derül a gráf hiányosságaira vagy hibáira, és a téves találatokat is ilyenkor szűrjük ki. A legrosszabb találati arányt a jelen idő gráfja adta, a többi esetben csak 10 % körüli volt a felesleg.

A konkordancia statisztikailag értelmezhető adatokká alakítását a NooJ-ban még nem látjuk megoldhatónak, ezért a kapott eredményeket tartalmazó output fájlokat txt. formátumban elmentettük, és az Atlas/ti program segítségével, egyszerű automatikus kódolással számszerűsítettük. Minden szövegnek van tehát egy saját konkordancia kimenete, amely az Atlas/ti-ben kapja meg a hozzárendelt számadatokat. Pl. 5 gyors, 6 huzamos, 20 funkcionális, stb. A szövegek eltérő hosszúsága miatt a végén minden kapott értéket elosztunk az adott szövegben található mondatok számával, így összehasonlíthatóvá válnak az egyes esetek.

5 Eredmények

Ha az összes időt kifejező kategória átlagát vesszük, akkor azt látjuk, hogy 4-5%-ban fordul elő ilyen kifejezés vagy szó a 48-as forradalom, az I. és II. Világháború, az 56-os forradalom és a rendszerváltás szövegeiben.3 százalék körüli ez az arány a Hunyadi, Mátyás király, Mohácsi csata, Trianon és a Holocaust történeteiben. Ennél kevesebb, 2% körül mozog a Honfoglalás, István király uralkodása a Tatárjárás, a végvárak háborúi, Rákóczi szabadságharc, és a Kiegyezés esetében.

A 3. ábra grafikonján látszik, milyen arányban fordulnak elő a funkcionális és a tartalmi elemek a szövegekben. A funkcionális kategória kiemelkedően magas a II. Világháborút (13) tárgyaló részekben, de ugyanitt a tartalmi elemek aránya is magas. Tartalmi elemek szempontjából kimagaslik a 48-as forradalom és szabadságharc, az 1956-os forradalom, és a rendszerváltás is. Megfigyelhető, hogy ha kiemelkedően magas a tartalmi elemek aránya, akkor a funkcionális elemek száma nem nő ezzel együtt. A két változó között nincs korreláció, a Trianonról szóló elbeszélésekben például átlagosan kevesebb, 2 százalék alatti a funkcionális időkategória kifejeződése, míg a tartalmi elemek aránya viszonylag magas.

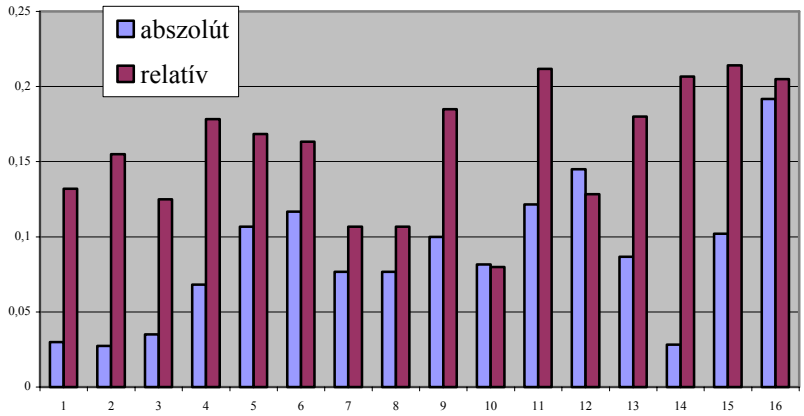


3. ábra. Funkcionális és tartalmi időkategóriák átlagainak megoszlása

Ha részletesebben megvizsgáljuk a funkcionális időkategória összetevőinek megoszlását, az derül ki, hogy a legtöbb találatot a „kezdet” algráf hozta. Második helyen a „pillanatnyi” kategória áll, a „befejeződés” és a „lassú” idő pedig elenyésző számban jelent meg a szövegekben. Itt már nagyon kis számokról van szó, az egyes kategóriák átlagának százalékos aránya a szövegekben 1% alatti. Mindenképpen érdemes azonban figyelembe venni, hogy a funkcionális gráf találatainak nagy része a „kezdet”, a „pillanatnyi”, illetve a „gyors” és a „huzamos” kategóriából kerül ki. A II. Világháború esetében, ahol kiemelkedő számban jelennek meg a funkcionális időszavak, a legtöbb ilyen a „kezdet”, a „pillanatnyi” és a „gyors” alkategóriából került

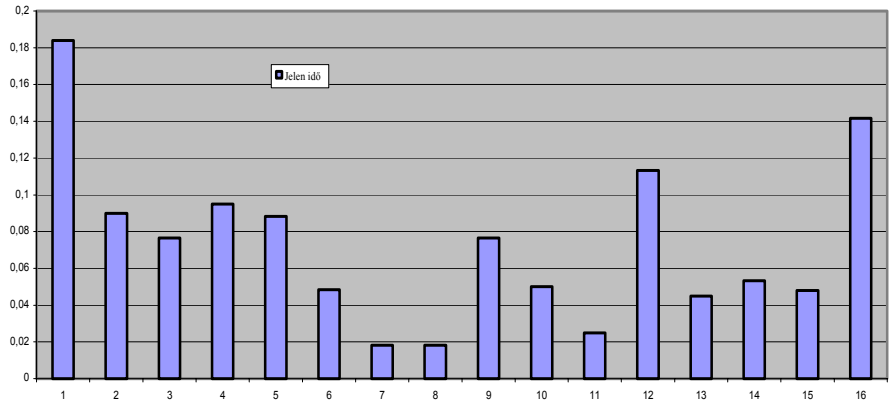
ki. Külön talán csak a „kezdet” kategória elemezhető, a legmagasabb az aránya – 1% körüli – a rendszerváltás és a II. Világháború történeteiben, a legalacsonyabb pedig a Trianon elbeszélésekben.

A tartalmi kategória algráfjainak eloszlását mutatja be a 4. ábra grafikonja. A relatív időkategória átlagosan nagyobb százalékban van jelen, mint az abszolút. Ezen belül az abszolút időmeghatározások aránya kiemelkedően magas a Rendszerváltás és Trianon történeteiben, és alacsony a Honfoglalás, a Holocaust, István király és a Tatárjárás elbeszéléseiben. A relatív kategória magas az I. Világháború, a Holocaust és a Rendszerváltás esetében, és legalacsonyabb a Kiegyezés szövegeknél. Azt is megvizsgálhatjuk, mikor vannak ezek a kategóriák közel egyenlő arányban, és mikor jelenik meg nagyobb különbség. A különbség mindig a relatív kategória javára növekszik, és a legnagyobb a Holocaust esemény elbeszélésekben. A legtöbb eseménynél nem ekkora, de azért számottevő különbséget találunk a relatív kategória javára, néhány esetben pedig kiegyenlített a két időjelölő előfordulása (Végvárak, Rákóczi, Kiegyezés, Trianon, Rendszerváltás).



4. ábra. Abszolút és relatív időjelölők átlagainak megoszlása

Az ötödik ábra grafikonja a jelen idő megjelenésének százalékos előfordulását ábrázolja. A legtöbb jelenre utalást a Honfoglalás és a Rendszerváltás elbeszélésében találtuk. Legkevésbé a Végvárak, Rákóczi és az I. Világháború elbeszélései mozdultak ki a jelen időbe.



5. ábra. A jelen idő előfordulásának átlagos megoszlása

6 Eredmények értékelése, megvitatás

A módszer működésével kapcsolatban úgy gondoljuk, hogy az időszavak 1-5%-os találati aránya a szövegekben megfelel a valós előfordulásnak. Mintánkban megfigyelhető, hogy magasabb az időjelölők használata akkor, ha több eseményt mutat be egymás után a szöveg, több az elbeszélő rész, kevesebb a leírás és a magyarázat.

Azok a korszakok vagy események, amelyek a magyar történelem kezdeti időszakára esnek, a többi eseményhez viszonyítva kevesebb időjelölőt tartalmaznak. Ennek az lehet az oka, hogy a történészek több eseményt sűrítenek be a jelenhez közeli, mint a távolabbi múlt leírásába, illetve kevesebb abszolút időjelölőt használnak – bizonytalanabb is a történetírás ezeken a területeken, és talán jobban közelít a fikcióhoz, a meséhez. (Honfoglalás, István, Tatárjárás). A hosszabb időtartamok, korszakok leírásánál az abszolút időjelölők Leduc által megfigyelt nagyobb aránya a mi mintánkban nem mutatkozott meg.

A II. Világháború, a 48-as forradalom, 1956 és a rendszerváltás, ahol a legtöbb tartalmi kifejezés fordult elő, fontos szerepet játszanak a magyar történelem szociális reprezentációjában. Ezek közül három esemény meglehetősen közeli, és sokat szerepel a mindennapi diskurzusban is, a 48-as forradalomra pedig nemzeti ünneppel emlékezünk minden évben: az eseményszerűség, „drámaiság” könnyebbé teszi az azonosulást ezzel a csoportidentitást alapjaiban meghatározó történettel.

A funkcionális időkategóriát a szubjektív időélmény vizsgálatára fejlesztjük, valószínűnek tartjuk, hogy ezek azok a kifejezések, szavak, amelyek nem csupán megérthető, hanem átélhető, érzelmi élménnyé teszik számunkra az idő múlását, tempóját, ritmusát vagy tartamát az elbeszélésekben. Az a tény, hogy ezek a jegyek megjelennek a szövegekben, mutatja, hogy a történészek nem csupán adatokat sorolnak fel könyveikben, hanem megpróbálják átélhetővé tenni azt az időt, amelynek nem lehettünk részesei.

A jelen idő a Honfoglalásnál és a Rendszerváltásnál jelent meg leginkább, amely mutatja, milyen különböző jelentése lehet az igeidő váltásoknak. A Honfoglalás történeteiben megjelenő jelen talán a történész szándékát jelzi, hogy összekapcsolja a kezdetet, a régmúltat a mával, és közelebb hozza ezt a bizonytalan, elmosódott és a legendák idejében játszódó korszakot. A Rendszerváltás viszont élénken benne van a diskurzusban, főszereplői ma is ismert politikusok, sokak számára személyes élmény. Az itt megjelenő jelen idő inkább leírásokat, ismertetést tartalmaz az alkotmányról, ami most is érvényben van, a lehetőségekről, amelyek ekkor teremtődtek, és most is megvannak.

A történelemkönyvek meghatározzák az időfogalom alakulását, olyan kategóriákat vezetnek be, mint például a történelmi korok, korszakok, amelyek segítségével gondolkodni tudunk arról az időről, ami meghaladja az emberi élet, vagy egy generáció közvetlen emlékezetének határait. A saját személyes identitás kialakításának kezdete serdülőkorban [5] a személyes múlt, jelen és jövő közötti kapcsolatteremtést jelent. Ez egybeesik a történelemtanulás kezdeteivel, a csoport közös történetét részben az iskolában elsajátított elbeszéléseknek köszönhetően ismerhetjük meg. Ezeknek a narratíváknak az idői szerveződése, mind kihatással lehet arra, hogyan alakul a csoportidentitás és a jövőperspektíva.

Bibliográfia

1. Ehmann Bea, Kiss Balázs, Naszódi Mátyás, László János: A szubjektív időélmény tartalomelemzéses vizsgálata. A LAS Vertikum időmodulja. Pszichológia, 2005/2. 133-142. (2005)
2. Ehmann Bea: Az énelbeszélés idői szerkezetei. In: Személyiséglélektantól az egészségpszichológiáig. Tanulmányok Kulcsár Zsuzsanna tiszteletére. Trefort Kiadó, Budapest. /in press/(2007)
3. Ehmann Bea: Tartalomelemzési módszerek a szubjektív időélmény vizsgálatára laikus beszélők szövegeiben. In: Szerk.: Erős Ferenc: Az elbeszélés az élmények kulturális és klinikai elemzésében. Pszichológiai Szemle Könyvtár 8. Akadémiai Kiadó, Budapest, 57-73. (2004)
4. Ehmann, Bea, Garami Vera, Szabó Júlia: NooJ fejlesztések a szubjektív időélmény tartalomelemzéses vizsgálatára. (2006)
5. Erikson, Erik H: Az életciklus: Az identitás epigenezise. In: A fiatal Luther és más írások. Gondolat, Budapest. (1991)
6. László János: A történetek tudománya. Bevezetés a narratív pszichológiába. Budapest, Új Mandátum Könyvkiadó. (2005)
7. Leduc, Jean: A történészek és az idő. Elméletek, problémák, írásmódok. Kalligram, Pozsony, (2006)
8. Pillemer, D. B., Desrochers, A: B., & Ebanks, C. M. Remembering the past in the present: Verb tense shifts in autobiographical memory narratives. In C.P. Thompson, D. J. Herrmann, D. Bruce, J.D. Read, D.G. Payne, & M.P. Toglia (Eds.). Autobiographical memory: Theoretical and applied perspectives. (pp. 145-162) Lawrence Erlbaum Associates Inc., Mahwah, NJ. (1988).
9. Pólya T. Identitás az elbeszélésben. Szociális identitás és narratív perspektíva. Budapest, Új Mandátum Kiadó. (2007).
10. Váradi T (szerk.):Magyar Nemzeti Szövegtár. MTA Nyelvtudományi Intézet, Budapest, (2004)

A pszichológiai perspektíva előfordulása történelem tankönyvi szövegekben

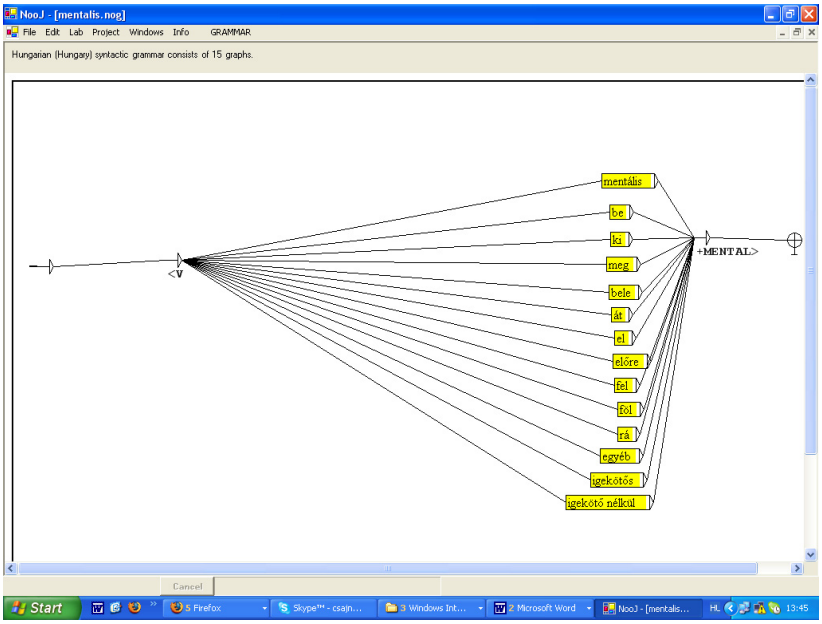
Pólya Tibor[1], Vincze Orsolya[2], Fülöp Éva[2] és Ferenczhalmy Réka[2]

[1]MTA Pszichológiai Kutatóintézet, Pf.: 398
1394 Budapest, Magyarország
[2]PTE Pszichológiai Intézet
7624 Pécs, Magyarország
polya@mtapi.hu
vinor@freemail.hu
petymeg81@freemail.hu
ferreka@freemail.hu

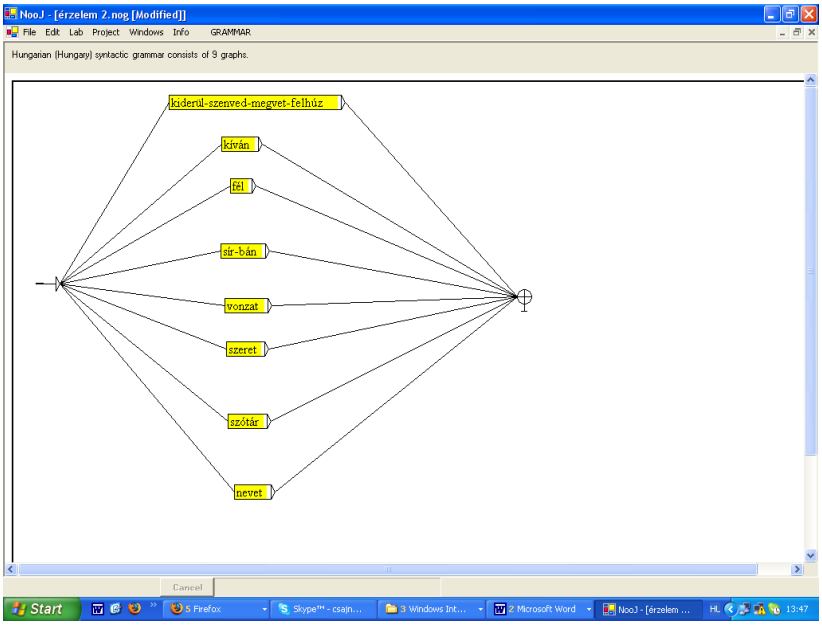
Kivonat: Az előadás a pszichológiai perspektíva modult mutatja be, amely a kognitív, érzelmi és intencionalitás gráfokat foglalja magában. A pszichológiai perspektíva két formáját különböztetjük meg. Belső perspektíva érvényesítése esetén az elbeszélő bemutatja a szereplők tudattartalmait, külső perspektíva érvényestésekor a szereplők viselkedésének bemutatására korlátozódik az elbeszélés. A pszichológiai perspektíva modul pszichológiai relevanciájának feltárásához három hipotézist fogalmazunk meg, amely összefüggést feltételez a belső perspektíva érvényesítése és a történelmi szövegek három jellemzője (valencia, időbeli távolság, és a történet megértési nehézsége között. A hipotézisek teszteléséhez 16 jelentős magyar történelmi eseményt leíró általános és középiskolai történelem tankönyv szövegét elemeztük. Az eredmények azt mutatják, hogy a belső perspektíva előfordulása könnyebben megérthetővé teszi a történelmi eseményeket bemutató történeteket.

1 Bevezetés: Pszichológiai perspektíva

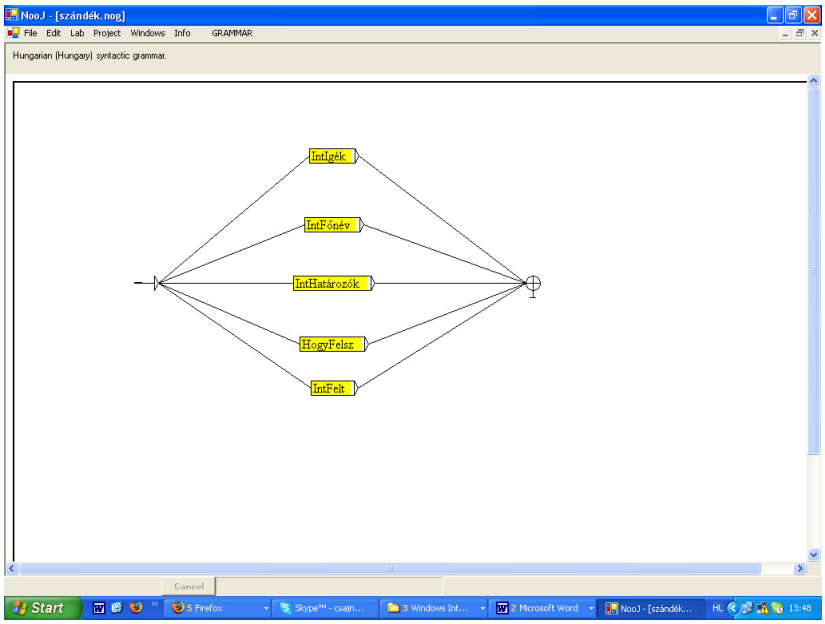
A pszichológiai perspektíva fogalmán a történetek azon jellegzetességét értjük, hogy az elbeszélő a szereplők belső tudattartalmait is bemutathatja. A pszichológiai perspektíva két alapvető formája a belső és külső perspektíva. Belső perspektíva érvényesítése esetén a szereplők mentális állapotáról is tudomást szerzünk, míg külső perspektíva érvényesítése esetén az elbeszélés a szereplők viselkedésének bemutatására korlátozódik. A pszichológiai összetevő automatikus kódolására elsőként Wiebe [1] fejlesztett ki algoritmust, amely angol nyelvű történetek elemezését végzi. A pszichológiai perspektíva automatikus elemzésére általunk kifejlesztett eszköz három önmagában is összetett gráfot foglal magába: a szereplők mentális állapotának bemutatását azonosító kognitív gráfot, amelyet Vincze Orsolya dolgozott ki [2, lásd 1. ábra], a szereplők affektív állapotait kódoló érzelmi gráfot, amelyet Fülöp Éva fejlesztett ki [3, 4 lásd 2. ábra] és a szereplők viselkedésének szándékosságát azonosító intencionalitás gráfot, amelyet Ferenczhalmy Réka dolgozott ki [5, 6 lásd 3. ábra].



1. ábra. A kognitív gráf



2. ábra. Az érzelem gráf



3. ábra. Az intencionalitás gráf

2 A pszichológiai perspektíva előfordulása történelem tankönyvi szövegekben

2.1 Hipotézisek

A pszichológiai perspektíva pszichológiai vonatkozásait elsősorban irodalmi narratívumokon vizsgálták. Több vizsgálat eredményei is azt jelzik, hogy a pszichológiai perspektíva hozzájárulhat az adott eseményben résztvevő szereplők felelősségének megítéléséhez [7]. Belső perspektíva érvényesítése növelheti a szereplők felelősségét, hiszen a tudattartalmak bemutatása alapján a cselekvés szándékosságára következtethetünk. Ezzel szemben a külső perspektíva használata semlegesnek tekinthető ebben a tekintetben. Ez alapján azt várhatjuk, hogy összefüggést találunk az események pozitivitása versus negativitása és a pszichológiai perspektíva előfordulásának mértéke között. Hipotézisünk az, hogy pozitív történelmi események leírásában többször fordul elő belső perspektíva, mint negatív történelmi események bemutatásában.

Második hipotézisünk a vizsgált szövegek azon sajátosságához kapcsolódik, hogy történelmi eseményeket bemutató történeteket vizsgálunk. Azt feltételezzük, hogy a pszichológiai perspektíva előfordulása összefügghet a történetbe foglalt esemény és a

történelem tankönyvek megírása közt eltelt idő mértékével, mivel a belső perspektíva érvényesítése kevésbé tűnik valóságosnak időben távoli eseményeknél szemben az időben közeli eseményekkel. Ez alapján azt a hipotézist fogalmazzuk meg, hogy időben közeli történelmi események leírásában gyakrabban jelenik meg a belső perspektíva, mint távoli történelemi események ismertetésében.

Végül harmadik hipotézisünk a belső perspektíva előfordulása és a történet megértési nehézsége között fogalmaz meg kapcsolatot. A történet megértését nagy mértékben befolyásolja az, hogy az olvasók milyen mértékben képesek kapcsolatokat kialakítani a történetbe foglalt események között. Az események közötti kapcsolatok kialakításának egyik leghatékonyabb eszköze a szereplők céljainak bemutatása, amelyek összefűzhetik az eseményeket. Mivel a szereplők mentális aktusainak, érzelmi állapotainak és szándékainak kifejezése gazdag hátteret biztosít a szereplők céljainak megállapításához, azt várhatjuk, hogy a belső perspektíva előfordulása könnyebbé teszi a történetek megértését. Ez alapján azt a hipotézist fogalmazzuk meg, hogy az általános iskolai történelem tankönyvekben többször fordul elő belső perspektíva, mint a középiskolai tankönyvekben.

A hipotézisek teszteléséhez az MTA Pszichológiai Kutatóintézetében Ehmann Bea vezetésével összeállított történelem tankönyvi szövegek korpuszának egy részén futtattuk a pszichológiai perspektíva modult. A korpusz a magyar történelem 16 jelentős eseményét öleli fel a honfoglalástól egészen a rendszerváltásig. A korpuszban az összes ma használatban lévő általános és középiskolai történelem tankönyv szerepel. A vizsgált szövegmintába 32 történelemtankönyv fejezetei kerültek bele. A fejezetek a következő 3 szempont alapján kerültek kiválasztásra: a történelmi esemény valenciája (pozitív versus negatív, a történelmi esemény óta eltelt idő (közeli versus távoli), és a történelemtankönyv olvasói (általános iskolás versus középiskolás diákok). A kiválasztott történelmi eseményeket az 1. Táblázat tartalmazza.

1. Táblázat: Az elemzett történelmi események jellemzői

Történelmi esemény		
Időbeli távolsága	**Valenciája**	
	Pozitív	**Negatív**
Távoli	Honfoglalás Mátyás király uralkodása	Tatárjárás Mohácsi csata
Közeli	1848-49-es forradalom és szabadságharc Rendszerváltás	II. Világháború Trianoni békekötés

2.1 Eredmények

Az elemzés első lépéseként a pszichológiai perspektíva modult alkotó három gráf találatainak kapcsolatát vizsgáltuk meg. Ha a kognitív, érzelmi és szándékosságra utaló kifejezések valóban egy konstruktumhoz kapcsolódnak, akkor magas pozitív korrelációt várhatunk ezek előfordulása között. A 2. táblázatban látható korrelációs együtthatók megerősítik ezt a várakozást. Azaz a szereplők mentális aktusainak, ér-

zelmi állapotának és a cselekvések hátterében álló szándékok bemutatása egyaránt hozzájárulnak a szereplő tudattartalmainak kifejezéséhez.

2. Táblázat: A kognitív, érzelem és szándékosság gráfok találatai közötti korrelációk

	Érzelem gráf	Szándékosság gráf
Kognitív gráf	0,44*	0,59**
Érzelem gráf		0,56**

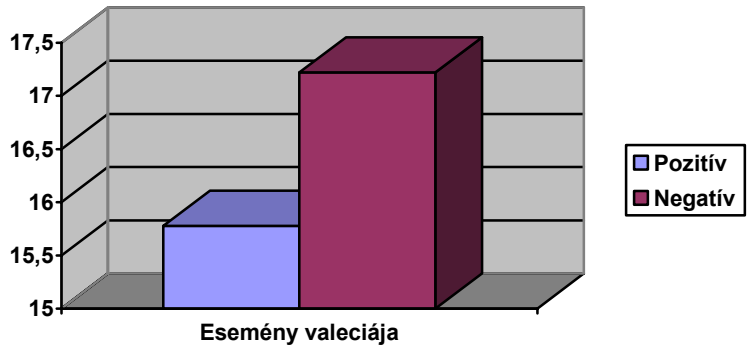
** $p < 0,01$; * $p < 0,05$

Az elemzés második lépéseként a belső perspektíva előfordulását vizsgáltam meg a pozitív versus negatív, időben közeli versus távoli történelmi események leírásában, illetve az általános és középiskolás tankönyvek szövegeiben (lásd 4-6. ábrák). Különbséget csak ez utóbbi szempont elemzésekor találtunk. Az általános iskolai tankönyvekben tendenciaerősséggel gyakrabban fordult elő a belső perspektíva forma mint a középiskolai szövegekben. Ez a különbség megerősíti harmadik hipotézisünket, és azt jelzi, hogy a szereplők tudattartalmainak bemutatása könnyebben megérthetővé teszi a történelmi eseményeket bemutató történeteket.

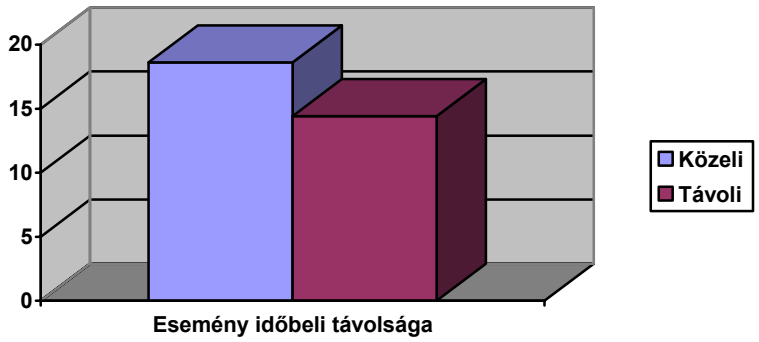
2.1 Következtetések

A pszichológiai perspektíva történelem tankönyvi előfordulásainak vizsgálata alapján két következtetést fogalmazhatunk meg. A szereplők tudattartalmainak bemutatása a valóságos eseményeket bemutató történetekben is segíti a történetmegértését. Ebből arra következtethetünk, hogy a történelmi alakok tudattartalmainak bemutatása fontos szerepet játszik a történelmi események értelmezésében.

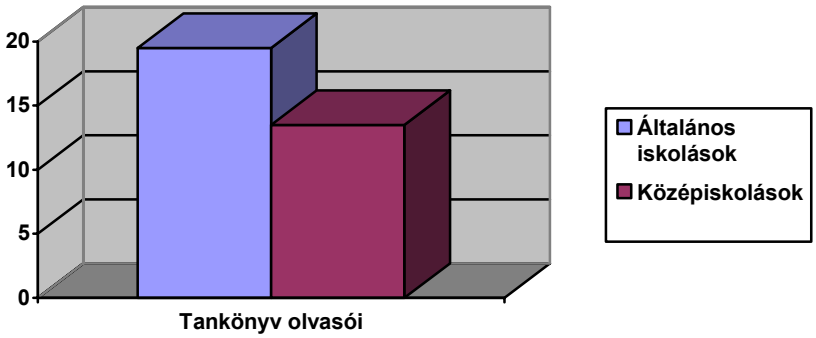
A vizsgálat eredményei alapján arra is következtethetünk, hogy a mentális, érzelmi és intencionális gráfokat összefogó pszichológiai perspektíva modul önállóan értelmezhető szintet képez a tartalomelemzési kategóriák pszichológiai vonatkozásainak vizsgálatában.



4. ábra A belső perspektíva forma előfordulása a pozitív és negatív történelmi eseményekben



5. ábra A belső perspektíva forma előfordulása az időben közeli és távoli történelmi eseményekben



6. ábra A belső perspektíva forma előfordulása az általános iskolai és középiskolai tankönyvekben

Bibliográfia

1. Wiebe, J.M. (1991). Tracking point of view in narrative. *Computational Linguistics. 20(2),* 233-287
2. Vincze O., László J. (2006). A mentális igék szótára, valamint alkalmazása az automatikus tartalomelemzésben. IV. Magyar Számítógépes Nyelvészeti Konferencia. 339-345. Szeged.
3. Fülöp É., László J. (2006). Az elbeszélések érzelmi aspektusának vizsgálata tartalomelemző program segítségével. IV. Magyar Számítógépes Nyelvészeti Konferencia. 296-304. Szeged.
4. Fülöp É., László J.: Az érzelmek reprezentációja történelmi regényekben és történelemkönyvekben V. Magyar Számítógépes Nyelvészeti Konferencia. Szeged.

5. Ferenczhalmy R., László J. (2006). Az intencionalitás modul kidolgozása NooJ tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia. 285-295. Szeged.
6. Ferenczhalmy R., László J.*:* Intencionalitás modul. V. Magyar Számítógépes Nyelvészeti Konferencia. Szeged.
7. Baumeister, R.F., Stillwell, A., Wotman, S.R. (1990). Victim and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger, *Journal of Personality and Social Psychology, 59(5),* 994-1005.
8. László, J. & Pólya, T. (2007). Level of abstraction versus subjectivity-objectivity in Linguistic Intergroup Bias. 2nd Warsaw – Jena Seminar on Prejudice. Warsaw.

Az aktív és passzív igék gyakorisága a csoportjelenségek tükrében

Történelemkönyvek szövegeinek narratív pszichológiai vizsgálata

NooJ tartalomelemző programmal

Szalai Katalin[1], László János[2]

[1] Pécsi Tudományegyetem Pszichológiai Intézet Doktori Iskola
7624 Pécs, Ifjúság útja 6.
szalai_katalin@freemail.hu

[2] MTA Pszichológiai Kutatóintézete
1132 Budapest, Victor Hugo u. 18-22.
laszlo@mtapi.hu

Kivonat: A szociális percepció és az identitás tanulmányozásában is meghatározó szerepe van az ágencia kérdésének. A személyészlelés során megjelenhet valaki saját életeseményeinek aktív résztvevőjeként, vagy ezen események passzív elszenvedőjeként. Az aktivitás a narratív szövegekben az aktív és passzív igék gyakoriságában érhető tetten. Tavaly bemutatásra került az 'aktivitás – passzivitás' szótár kialakításának folyamata NooJ tartalomelemző programmal. Ennek során igeszótárak, illetve ezt kiegészítve – a NooJ program kínálta lehetőségek által – lokális nyelvtanok, azaz ún. gráfok készültek.
Jelen kutatás során a program tesztelését az in-group - out-group asszimetria jelenségkörében történelemkönyvek szövegein végeztük. Az eredmények azt mutatják, hogy a saját csoport inkább a pozitív esmények során jelenik meg aktív ágensként, míg az idegen csoport inkább a negatív eseményeknél. Továbbá olyan negatív történelmi eseményeknél, mint Trianon, a saját csoport az igék gyakoriságát alapul véve inkább az események passzív elszenvedőjeként jelenik meg, míg az idegen csoport aktív ágensként.

1 Bevezetés

1.1 Az identitás, a történet és a történelem kapcsolata

Szociálpszichológiai nézőpontból az identitás két koncepcióját kell megemlíteni: A személyes identitás eriksoni fogalma kifejezi, hogy az egyén az életútján bekövetkező változások ellenére is megőrzi önazonosságát. Tajfel pedig bevezeti a szociális identitás fogalmát. Az identitás ezen formája egy csoporttal való azonosulást jelöl, mely során az egyén átveszi ezen csoport normáit, értékeit, ezáltal pedig biztonságot és pozitív önértékelést, önbecsülést nyer a csoporttól. [3]

Az identitás egyes elméletekben társas közegben csiszolódó, folyamatosan újra- és újraszerkesztett élettörténetként jelenik meg (Ricouer [8]; McAdams [7]; Gergen és Gergen [2]), ebből a narratívumból tehát következtethetünk a történetmondó lelki folyamataira, viselkedésére.

Moscovici nyomán szociális reprezentációnak nevezzük azt a folyamatot, ahogy egy csoport jelentéssel ruházza fel a jelenségeket, melyek ezáltal ismerős, jelentésteli reprzenetációkká, a szociális világ részeivé lesznek. Ezen reprezentációk a csoport-kommunikációban formálódnak, lehetővé teszik a csoport tagjainak szociális világban való eligazodását, valamint a csoporttudat kialakulásához is hozzájárulnak. [3]

A szociális reprezentációk sokszor narratív formában jelennek meg. Halbwachs szerint is az emberek világuk megértése érdekében történeteket alkotnak, történeteket osztanak meg egymással. Halbwachs kollektív emlékezetről beszélve kifejti, hogy az emlékezet tárgya társadalmi keretek között folytatott kommunikációban közvetítődik; csak az lesz emlékezetre méltó egy csoport számára, ami beleillik az ott jelen lévő szociális sémákba. [4]

A szociális reprezentációk, közös történetek megosztása, fenntartása hozzájárul a csoportidentitás – jelen esetben a nemzeti idenitás – fenntartásához.

Assmann az elbeszélés és az identitás témakörét az emlékezet nézőpontjából vizsgálja: az emlékezés kultúrája teszi folytonossá a csoport identitását. Megkülönböztet kommunikatív és kulturális emlékezetet. A kulturális emlékezet főként szövegekben jelenik meg az írásbeli kultúrákban. Ezen narratívumok – akár történészek akár írók által létrehotott narratívumok – kifejezik, hordozzák, továbbörökítik a csoport idenitását, közvetítik a csoport által elfogadott identitásmintákat. [1] [3] [11]

Mindezeket figyelembevéve elmondhatjuk, hogy a történelmi szövegek tartalomelemzése alkalmas a nemzeti identitás vizsgálatára, az identitást érintő szociális reprezentációk elemeinek feltárására [3]. Az iskolai tankönyvek tehát nemcsak egyszerűen információkat, történelmi tényeket rögzítenek és továbbítanak, hanem a csoport-identitás jellemzőire, elfogadott viselkedésformákra adnak mintát.

1.2 Narratív kutatások a csoportjelenségek területén

Korábban is végeztek kutatásokat az igék implicit jelentéséről, illetve ezek vizsgálatáról csoportjelenségek tükrében. Semin és Fiedler nyelvi kategória modellje (LCM) az absztrakció szintjét érintő 'nyelvi reprezentációs formákat' különböztet meg, amik a kommunikáció résztvevőinek világát strukturálják [9], [3].

A modellhez kapcsolódó vizsgálatban azt találták, hogy a saját csoport társadalmilag kívánatos viselkedését absztraktabban írják le, mint a külső csoportét, illetve a társadalmilag nem kívánatos viselkedésnél ellenkezőleg, elkerülik az absztrakt leírást a saját csoport tagjainál, hiszen ez egy általános reprezentációt hívna életre az adott eseményről [6].

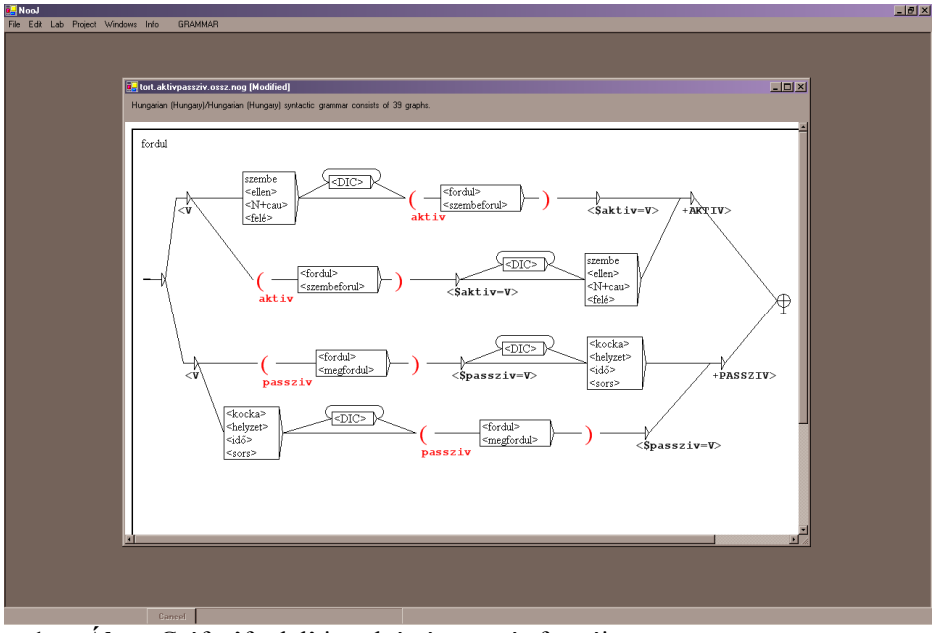
A narratív kutatás magyar történelmi szövegeket illetve történelmi eseményeket is többször választott tárgyául. Diákok legpozitívabbnak és legnegatívabbnak tartott magyar és európai eseményeiről szóló elbeszéléseket vizsgáltak [5], s többek között a Holocaust tekintetében találtak – az ágenciát érintő – eredményeket.

2 Az 'aktív – passzív' szótár

2.1 A szótár bemutatása

Munkacsoportunk a NooJ program fejlesztését különböző modulok mentén végzi. Ezen modulok egy-egy pszichológiai jelenséghez kapcsolódó nyelvi formák felismerésére és kigyűjtésére hivatottak a szövegek elemzése során. Ilyen modulok pl. a narratív perspektíva, a közelítés-távolítás, a szereplő-funkciók, az idő, a szelf-referencia, a tagadás, az értékelés, a mentális igék, az érzelem, az intenció vagy az aktivitás-passzivitás.

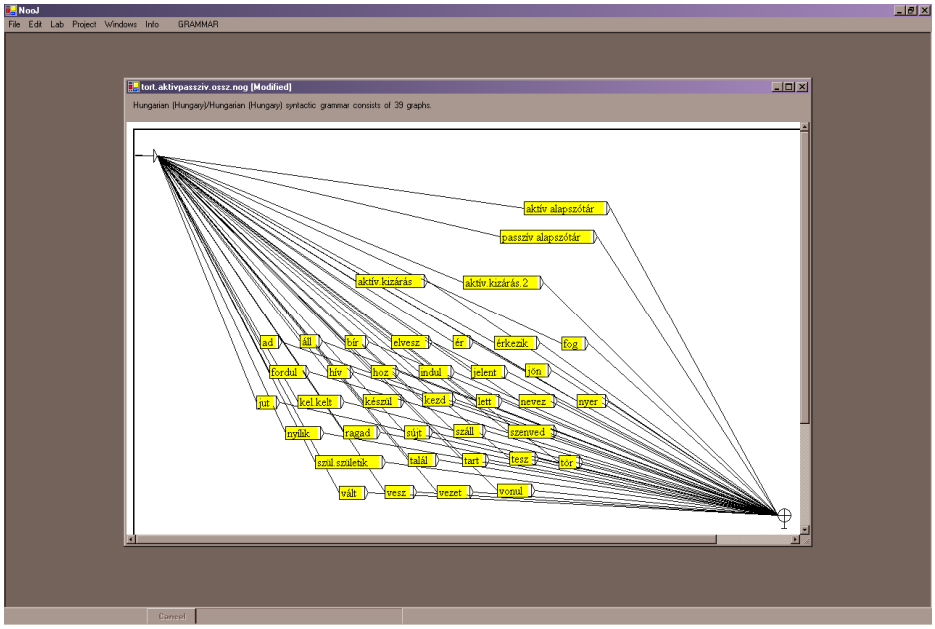
Jelen kutatás az 'aktivitás - passzivitás' NooJ – szótár segítségével, történelmi szövegeken végzett elemzést mutat be. Az igeszótár kidolgozása folyamán kialakult igekategóriák a következők: Aktivitás (pl.: harcol, menekül), passzivitás – azaz az állapotváltozás, történés igéi, illetve az állapotot, folyamatosságot kifejező igék (pl.: történik, alakul, kezdődik), az „aktivációs kontúr" igéi (pl.: fokoz, lassít, leheveredik, időzik) és a 'stop/go' igék (elkezd, abbahagy). Az igék egy bizonyos része nem szerepel a szótárban: mentális igék (beszédaktusok kivételével); érzelmeket kifejező igék (kedvel, utál); segédigék; környezetünk eseményeinek leírására vonatkozó igék, természeti állapotokra és azok változásaira, vagy fiziológiai folyamatokra vonatkozó igék (hajnalodik, remeg, kiizzad). [10]



1. Ábra: Gráf a 'fordul' ige aktív és passzív formáira

Az igéknek pusztán ezen kategóriákba való besorolása nem elegendő, hiszen vonzatkeretük, változatos igekötőik, idiómákban való megjelenésük alapján jelentésük módosulhat. Lokális nyelvtanok, azaz ún. szintaktikai gráfok különítik el ezen különböző jelentéseket, és sorolják be az igéket a megfelelő kategóriába.

Példa erre az 1. ábrán látható lokális nyelvtan, mely a „fordul” ige különböző változatait hivatott megtalálni és osztályozni. Ha a szövegben az egyik szereplő egy másik ‘ellen fordul’, vagy ‘szembefordul vele’, esetleg ‘feléfordul’ – úgy aktívként ismeri fel és nevezi el a gráf az igét. Ha viszont olyan idiómákban jelenik meg, mint ’úgy fordult a helyzet’, vagy ’fordult a kocka’, esetleg a ’sorsa rosszra fordult’ – úgy a gráf passzív kimenttel jelöli meg. Ez a lokális nyelvtan figyelembe veszi, hogy a magyar nyelvnek nincs kötött szórendje, tehát ezen kifejezések tagjai a mondatban szabadon felcserélhetők, illetve más szavak ékelődhetnek közéjük.

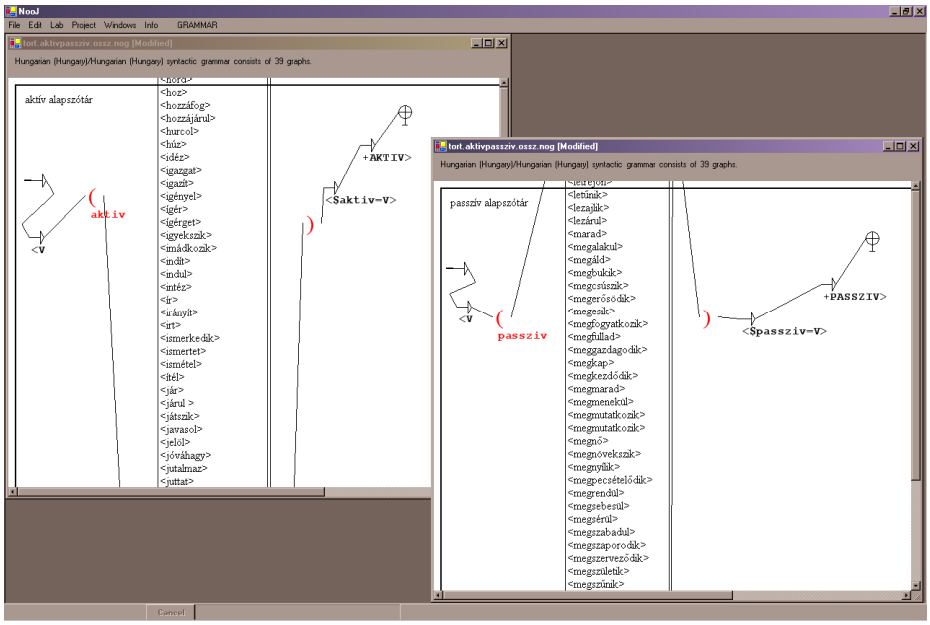


2. Ábra: Az aktív – passzív szótár gráfja: 38 további gráfot tartalmaz

2.2 A szótár alkalmazása

A 2. ábrán látható az ‘aktív-passzív’ szótár gráfja, mely 38 további gráfot foglal magában. Helyet kapnak benne az aktív és a passzív igéket tartalmazó szólisták (ld.: 3. ábra), amik az egyes szavak különböző eseteit, ragozott formáit is képesek megtalálni, és a megfelelő kimenettel ellátni. (Az aktív alapszótárban a jelen kutatásban szereplő szövegbázis 705 db igéje szerepel, a passzív alapszótárban pedig 227 db ige.) Ezen kívül helyet kapnak itt a ’kizáró’ gráfok is, melyek a helytelen találatokat külö-

nítik el: pl. a tör ige E/1 esetét (török), mely a történelemszövegekben igen nagyszámú téves találatot okozott az aktív igék rovására. De kizárandóként jelöltük meg például a különböző igeneveket is, hiszen pl. a 'menekül' igét aktívként kell számba venni, de a 'menekülve' határozói igenév már nem képezi a szótár részét. A gráf tartalmaz további 34, egyes – a szövegbázisban nagy gyakoriságú – igékre szabott gráfot (pl: hív, hoz, indul, kezd). Ezek a fent bemutatott (ld.: 1. ábra) „fordul" ige példájához hasonló módon különítik el az igék passzív illetve aktív jelentéseit szövegkörnyezettől függően.

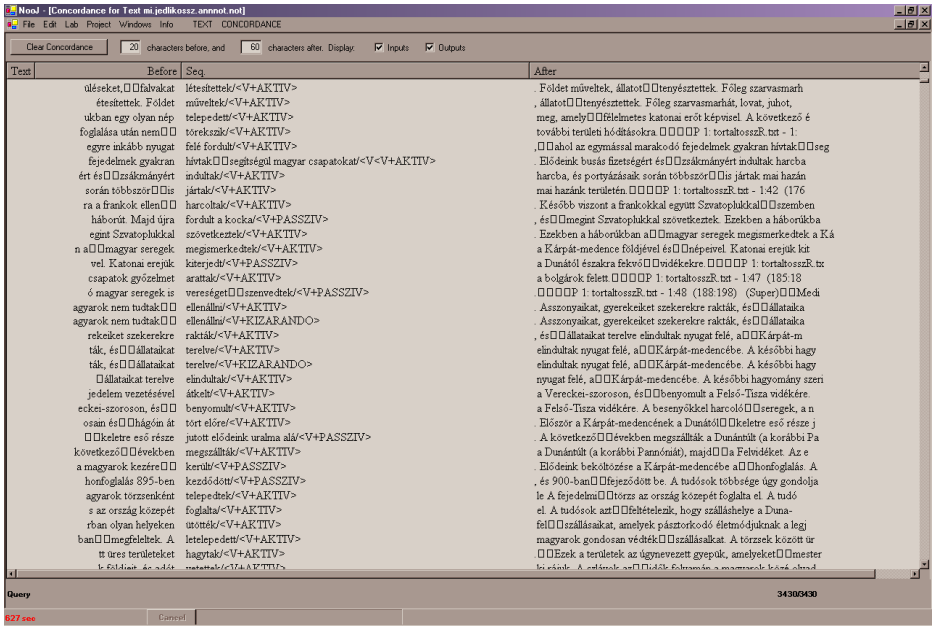


3. Ábra: Az 'aktív' és a 'paszív' alapszótár gráfjai

Ezen lokális nyelvtanok ellenére továbbra is születnek helytelen találatok, főként az ismeretelen – azaz a NooJ programban egyelőre még nem annotált – szavak miatt, illetve a magyar nyelv sajátosságai pl. homonímia jelensége miatt. (Erre példa: „szerzeteseket *küldött* a keleten maradt magyarok felkutatására", „az ellene *küldött* török csapatokat sorra megverte". Ezen egybeesések szétválasztása egyelőre nem megoldott.) .A szótár jelen állapotában a kutatásunkban szereplő szövegbázison 74 %-os egyezést mutat az Atlas.ti segítségével végzett manuális ellenőrzéssel.

A Nooj program jelenleg még nem képes ágencia mentén különbséget tenni (nem tudja "megmondani" hogy a törökök vertek-e el minket, vagy mi őket). Tehát egyelőre nem tudjuk a program segítségével meghatározni, hogy a mondat állítmánya saját vagy idegen csoportbeli alanyra vonatkozik-e, így az ingroup és outgroup szövegrészletek szétválogatását jelen kutatás során az Atlas.ti program segítségével, manuálisan tettük meg. Fejlesztés alatt áll egy program, ami a mondatbeli szerepet képes lesz megállapítani.

Az alábbi, 4. ábrán láthatók a gráf találatai, azaz az adott pszichológiai jelentést hordozó szavak, kifejezések az adott kiementtel ellátva.



4. Ábra: Az ’aktív – passzív’ gráf konkordanciája egy szövegrészleten

2 Történelemkönyvek szövegeinek vizsgálata

2.1 A vizsgálat menete

Jelen vizsgálatunkat történelemkönyvek szövegein végeztük. Az általános és középiskolás tankönyvek szövegrészletei több (általános iskolai taknönyvek 7, középisolai 8) kiadótól, tehát több szerzőtől származnak, s a magyar történelem főbb eseményeit érintik: honfoglalás, I. Szent István király, tatárjárás, Hunyadi János, Mátyás király, végvárak harcai a török támadások idejében, a kiegyezés, I. világháború, Trianon és a II. világháború (Ehamnn Bea válogatása). A teljes általunk vizsgált szöveg mintegy 150ezer szót tartalmaz. A szövegek egyes részeit – mint pl. kultúrtörténeti leírások – nem vettük figyelembe.

A vizsgálathoz a NooJ és az Atlas.ti tartalomelemző programot is használtuk. A fent megnevezett okok miatt az Atlas.ti program segítségével választottuk szét a szöveg saját csoportra és az idegen csoportra vonatkozó részeit, majd az abból képzett hipertexten (azaz porgram által egybegyűjtött azonos kóddal ellátott szöveg-

részleken) futtattuk le a a NooJ gráfjait. (Végül a vizsgálat befejező részében szintén Atlas.tivel végzetük el az ellenőrzést.)

Jelen kutatás során az fent bemutatott aktív – passzív igeszótáron kívül a munkacsoportunk által kidolgozott értékelő igeszótárt is alkalmaztuk (Bigazzi, Csertő, Nencini; 2007), mely szótár az igéket pozitív, semleges és negatív dimenzió mentén osztályozza. Néhány példa a gráf találataira: pozitív – megment, támogat, segít, létrehoz, sikerül; negatív – támad, fenyeget, akadályoz, üt, kihasznál, megsebesül; semleges – kér, folytat, áll, jön, megy.

E két kategróia (aktív-passzív és értékelő) mentén besorolt igék számát mutatja az alábbi táblázat, történelmi események illetve saját csoport/idegen csoport szerinti felosztásban. (1. táblázat)

1. Táblázat: Eredmények

Történelmi események	**In-group**						**Out-group**					
	Akt Neg	**Akt Poz**	**Akt Seml**	**Pas Neg**	**Pas Poz**	**Pas Seml**	**Akt Neg**	**Akt Poz**	**Akt Seml**	**Pas Neg**	**Pas Poz**	**Pas Seml**
Honfoglalás	20	16	71	19	4	33	29	1	38	11	6	13
Szent István	18	15	39	12	3	22	1	1	3	0	0	3
Tatárjárás	23	34	91	44	8	44	54	11	72	5	6	20
HunyadiJános	38	47	106	46	8	42	23	15	79	9	11	6
Mátyás király	34	25	64	12	14	24	6	1	4	0	0	1
Mohács	27	12	41	33	5	33	27	6	50	1	10	9
Végvárak	16	35	40	24	4	9	22	5	57	7	10	12
Kiegyezés	2	3	11	8	4	5	4	1	5	7	0	5
I. világháború	5	1	12	10	3	10	35	18	98	23	7	34
Trianon	3	3	39	35	2	38	17	4	43	0	5	5
II.világháború	51	35	203	83	6	128	97	13	179	19	8	38
Össz	**237**	**226**	**717**	**326**	**61**	**388**	**315**	**76**	**628**	**82**	**63**	**146**

2.1 Az eredmények értékelése

Az általunk vizsgált szövegekben az összes találatot tekintve a saját csoport képviselői gyakrabban jelentek meg pozitív igék ágenseként, mint az idegen csoport tagjai, és fordítva, a negatív igék ágenseként többször álltak idegen csoport képviselői, mint saját csoport tagjai. Az eredmények feldolgozásánál chi-négyzet próbát alkalmaztunk.

Egy korábbi kutatás [5] megállapította, hogy a magyarok az első és második világháborút illetve a trianoni békét tartják a magyar történelem legnegatívabb eseményeinek. (Ezek a „csak vesztettünk” narratív sémába illeszkedő történetek.) Jelen kutatás eredményei szerint a II. világháború és Trianon eseményénél többször jelentek meg az in-gorup tagok mint az események elszenvedői, az out-group tagok pedig inkább aktív ágensként fordultak elő.

A II. világháború az aktív és negatív, aktív és pozitív illetve passzív és negatív igék tekintetében mutatja a fenti állításhoz illeszkedő statisztikailag szignifikáns eredményeket. A passzív és pozitív igék ezt a különbséget nem mutatják.

Érdemes megemlíteni, hogy a Trianonról szóló szövegrészeknél a negatív igék (mind aktív negatív, mind passzív negatív) tekintetében megtalálható a fenti – statisztikailag szignifikáns – különbség a csoportok között, viszont a pozitív igék rendkívül kicsi száma miatt ezen különbség nem mutatható ki. A szövegrészlet összes igéihez mérten nagyon kicsi a pozitív igék aránya – ami az esemény traumatikus voltáról vall.

3 Összegzés

Mint látható, a történelem szövegek narratív pszichológiai elemzése csoportközi viszonyok szempontjából hozhat értékelhető eredményeket. Mind a szövegbázisunkban, mind a kapott eredményekben vannak további, eddig kihasználatlan lehetőségek. Munkacsoportunk elérkezett ahhoz, hogy az egyes modulok külön-külön tesztelésén továbblépve, együttes vizsgálatokat végezzünk, azaz a különböző modulokat egyszerre használva a szövegek többszempontú, árnyaltabb narratív elemzését adhatjuk a NooJ tartalomelemző program további fejlesztésével együtt.

Bibliográfia

1. Assmann, J.: A kulturális emlékezet. Atlantisz, Budapest (1999)
2. Gergen, K.J. – Gergen, M. M.: A narratívumok és az én mint viszonyrendszer. In.: .: László J. – Thomka B. (szerk.) Narratívák 5. Narratív pszichológia. Budapest, Kijárat Kiadó (2001) 77-120.
3. László J.: A történetek tudománya. Bevezetés a narratív pszichológiába. Új Mandátum Kiadó. Budapest (2005)
4. László J.: Szociális emlékezet: A tröténelem szociálpszichológiája. Magyar Tudomány 2003/1.
5. László J. – Ehmann B. – Imre O.: Történelem történetek: a történelem szociális reprezentációja és a nemzeti identitás. Pszichológia, 22, (2002) 147-162.
6. Maass, A., Salvi, D., Arcuri, L., Semin, G.R.: Language use in intergroup contexts: The linguistic intergroup bias . Journal of Personality and Social Psychology , 57, 981—993 (1989)
7. McAdams, D.P.: A történet jelentése az irodalomban és az életben. In.: László J. – Thomka B. (szerk.): Narratívák 5. Narratív pszichológia. Budapest, Kijárat Kiadó. (2001) 157-175.
8. Ricoeur, P.: A narratív azonosság. In.: László j. – Thomka B. (szerk.) Narratívák 5. Narratív pszichológia. Budapest, Kijárat Kiadó (2001) 15-27.
9. Semin, G. R. – Fiedler, K.: The Linguistic Category Model, Its Bases, Applications and Range. European Review of Social Psychology, 2, (1991) 1-30
10. Szalai K. – László J.: Az aktivitáss – passzivitás modul kidolgozása NooJ tartalomelemző programmal. In.: IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
11. Vincze O. – Kőváriné Somogyvári I.: A nemzeti identitás reprezentációja a sikeres történelmi regényekben. In.: Magyar Tudomány 1. (2003)

Mentális kifejezések jelentősége a perspektíva-felvételben a csoportidentitás tükrében

Vincze Orsolya[1]

[1] PTE BTK Pszichológia Intézet
7624 Pécs, Ifjúság út 6.

orsolyavincze@hotmail.com

Kivonat: A mentális kifejezések szerepet játszanak az empátia és az azonosulás folyamataiban. A mentális állapotok figyelembevételének képessége elősegíti a cselekvő személy perspektívájának felvételét. A folyamat során a megfigyelő kiterjeszti saját tudatát a cselekvő személyre, amely megkönnyíti a cselekvő személy belső történéseinek elképzelését, cselekvésének megértését. Vizsgálatomban a Nooj nyelvi program keretében kidolgozott mentális gráfot alkalmaztam 1900 és 2007 között megjelent középiskolai történelemtankönyvek Osztrák-Magyar Monarchiával kapcsolatos szövegrészeinek elemzésére. Előadásomban a mentális akciók gyakorisági eloszlása és a csoport identitás összefüggéseire szeretnék rámutatni.

1 Bevezetés

A társas viselkedés alapvető feltétele a perspektíva-felvétel képessége. Lényeges szerepet játszik a segítségnyújtásban [1], csökkenti a szociális agressziót [9], valamint a társas ítéletalkotás hibáit [10].

A perspektíva-felvétel szituatív kezelése a szociálpszichológiában nem új keletű dolog, azonban közvetlen vizsgálata csupán az utóbbi évtizedekben kezdődött. Tudománytörténeti gyökerei az attribúció tárgyköréhez vezetnek vissza, amely a cselekvések oki magyarázatával foglakozik. A cselekvések mögött rejlő okokat tekintve általános jelenség, hogy mások viselkedésének magyarázatában nem vesszük figyelembe a nyilvánvaló külső okokat [14]. Másszóval indokolatlanul feltételezzük, hogy mások cselekvéseit azok személyes, belső tulajdonságai jobban meghatározzák, mint a környezeti tényezők, amelyekben a cselekvés megjelenik. Ugyanakkor önmagunk viselkedésének magyarázatában éppen ellenkezőleg járunk el, vagyis inkább a helyzeti körülményeket hangsúlyozzuk. Erre a diszkrepanciára, amelyet *cselekvő-megfigyelő torzításnak* hívunk, Jones és Nisbett [11] hívta fel a figyelmet, akik rámutattak arra, hogy a nézőpont eltérő tartalmakat eredményez a cselekvések értelmezésében. Ez a torzítás azonban nemcsak önmagunk és mások relációjában jelenik meg. Fiske és Taylor [12] vizsgálatában a megfigyelő egy interperszonális helyzet két résztvevőjének viselkedését látta különböző nézőpontból. Annak a személynek a viselkedését, akit a téri elrendezésből fakadóan jobban látott, szignifikánsan több

diszpozícionális okkal indokolta. A cselekvések oki magyarázatával foglalkozó kutatások tehát a perspektívát elsősorban a helyzet perceptuális körülményeinek függvényében vizsgálták. A legelső vizsgálatok egyike, amely kifejezetten a perspektíva-felvétellel foglalkozik Regan és Totten [13] nevéhez köthető. Kísérletükben nyíltan arra kérték a résztvevőket, hogy a cselekvő perspektíváját felvéve magyarázzák annak viselkedését. Nyilvánvalóvá vált, hogy a cselekvő-megfigyelő torzítás hiánya ebben az esetben egy kognitív erőfeszítés, a perspektíva-felvétel következménye, amelyben a megfigyelő nem csupán a célszemély viselkedéséhez könnyen illeszthető diszpozíciókat veszi figyelembe, hanem a cselekvő mentális állapotára is kiterjeszti a figyelmét. A célszemély gondolatainak, érzelmeinek, intencióinak figyelembevétele a cselekvés finomabb, részletesebb értelmezését eredményezi, amely szükségszerűen implikálja a szituatív tényezők tekintetbe vételét.

A kutatók között konszenzus mutatkozik arra vonatkozóan, hogy a perspektíva-felvétel szoros kapcsolatban áll az *észlelt hasonlósággal*. Davis és munkatársai [3] egy vizsgálatukban arra kérték a kísérleti személyt, hogy vegye fel a cselekvő perspektíváját, majd később jellemezze saját magát, valamint a cselekvő személyt egy előre megadott tulajdonságlistán. A kutatók azt találták, hogy a perspektíva-felvétel szignifikánsan több hasonló tulajdonságot eredményezett a megfigyelő és a cselekvő között, mint azoknak a személyeknek az esetében, ahol nem volt perspektíva-felvétel instrukció.

1.1 A perspektíva-felvétel szerepe a csoportközi folyamatokban

A csoportidentitás folyamatában lényeges szerepet játszik az észlelt hasonlóság. Önmagunk meghatározásában gyakran használunk olyan tulajdonságokat, amelyekben saját csoportunk más tagjaival osztozunk. A szociálpszichológiában ez az elgondolás a társadalmi kategorizáció elméletében jelenik meg [id. 15]. Egyik központi fogalma a *deperszonalizáció*, amely egy olyan folyamatot jelöl, amelyben a személy úgy fogja fel önmagát, mint egy szociális kategória kicserélhető tagját, s nem mint egy önálló személyiséggel rendelkező entitást. A deperszonalizáción keresztül a személyes tulajdonságok és a csoport tulajdonságai kibogozhatatlanul összekapcsolódnak [15]. Ennek alapján számos kutató követlen összefüggésbe hozza a csoportazonosulást az észlelt hasonlóság - vagyis a személyes és a csoport tulajdonságok átfedésének – mértékével [15] [18] [16]. Coats és munkatársai [2] újabban azt is bizonyították, hogy az észlelt hasonlóság a tulajdonságokon túl, az attitűdökre is kiterjed.

A saját csoporttal szemben észlelt hasonlóság funkcionális jellegű, amennyiben elősegíti a külső csoporttól való megkülönböztetést. A megkülönböztetés, vagy másság igénye az információfeldolgozás és megjelenítés eltérő módozatait implikálja csoportközi helyzetben. Például a saját csoport sikereit és a külső csoport kudarcait általában belső tényezőknek tulajdonítjuk, míg a saját csoport kudarcait illetve a külső csoport sikereit helyzeti tényezőkkel magyarázzuk. A pozitív megkülönböztetés igénye nyelvi szinten is megjelenik. A saját csoport pozitív, illetve a külső csoport negatív viselkedésének absztrakt nyelvi síkon történő leírása (melléknevek használata) időben tartós és konstans értéket tulajdonít a csoportnak. Míg a saját csoport nega-

tív és a külső csoport pozitív magatartásának konkrét, kontextusfüggő nyelvi megjelenítése (leíró igék használata) a viselkedés egyediségére utal, annak helyzetspecifikusságát implikálja, amely ezáltal megakadályozza az általánosítást [8].

Újabban Leyens és munkatársai [7] az érzelmekkel kapcsolatban, Kozak pedig általában a mentális állapotok tulajdonítására vonatkozóan is kimutatta a csoportközi torzítást. Az infrahumanizációs elmélet [7] szerint léteznek olyan humán sajátosságok, amelyek kizárólag az ember sajátjai, s amelyek birtoklása révén megkülönböztethető más alacsonyabbrendű élőlényektől. Ilyenek például a másodlagos, vagy szociális érzelmek (csodálat, büszkeség, vágyódás), amelyek társadalmilag meghatározottak, szemben az elsődleges, biológiai alapokkal rendelkező érzelmekkel (szomorúság, öröm, harag). Leyens és munkatársai kísérletileg igazolták a külső csoport infrahumanizációját a másodlagos érzelmek tekintetében. Ez az eredményekben úgy jelent meg, hogy a kísérleti személyek szignifikánsan több szociális érzelmet tulajdonítottak a saját csoportnak, mint a külső csoportnak. Kozak [6] hasonló elgondolásból kiindulva igazolta, hogy az infrahumanizáció a kognitív kapacitás tágabb tartományára is kiterjed, így például a gondolatokra, az intenciókra, és a célokra. Kísérleti eredményeiben beszámol arról, hogy az előítéletességet mérő skálán magas pontszámot elért személyek egy rákövetkező vizsgálatban szignifikánsan kevesebb mentális állapotot tulajdonítottak a színesbőrű amerikai célszemélynek, mint saját szociális kategóriájuk tagjának.

Az elmúlt évtizedben egyre hangsúlyosabbá váltak az elgondolások, amelyek a perspektíva-felvételt a csoportközi konfliktus csökkentésének egyik lehetséges módjaként kezelik. Galinsky és Moskowitz [5] abból a feltételezésből indultak ki, hogy a perspektíva-felvétel képes módosítani a külső csoport reprezentációját, amely ezáltal a másik csoport pozitívabb megítélését eredményezi. Az eredmények azt mutatták, hogy azok a személyek, akiket egy külső csoporttag perspektívájának felvételre instruáltak, nagyobb hasonlóságot észleltek a célszeméllyel, valamint jóval pozitívabban is értékelték a kontroll csoport tagjaihoz képest. Azonban az észlelt hasonlóság és a pozitív attitűd nem csupán a célszemély esetében volt kimutatható, hanem azon kategóriával szemben is, amelybe a célszemély tartozott (idős emberek csoportja). A vizsgálati személyek hasonló tulajdonságokkal ruházták fel önmagukat és a célszemélyt, mi több, ezek a tulajdonságok jóval kisebb számban tartalmaztak sztereotip kifejezéseket a célszemély kategóriáját illetően a kontroll csoporthoz képest. A vizsgálat alátámasztja azt a feltevést, hogy csoportközi kontextusban a külső csoporttag perspektívájának felvétele észlelt hasonlóságot eredményez, amely meggátolja a sztereotip konstrukciók használatát a célszemély értékelésében, valamint kategórizációjában.

2 Perspektíva-felvétel: a mentális igék szerepe a szövegben

2.1 Csoporttörténetek és a csoport perspektívája

A csoporttörténetek a csoportidentitás hordozói, és közvetítői egyszerre. Ezek az elbeszélések a történések tényszerű leírásán túlmenően, az eseményeket sajátos narratív perspektívában ábrázolják. Az elbeszélésben megjelenő szereplők tudat-

tartalmainak megjelenítése (mit gondol és érez), cselekvőképességük (passzív elszenvedő vagy aktív cselekvő), valamint cselekvéseik valenciája a csoportidentitás közvetítésének finom narratív eszközei. Ezeknek az eszközöknek a nyelvi kifejezései az események olyasfajta ábrázolását teszik lehetővé, amely a csoportidentitást elfogadható módon jeleníti meg.

Az előző fejezetben számos kísérlet bemutatása mentén megpróbáltunk rámutatni arra, hogy a perspektíva-felvétel lényeges szerepet játszik a cselekvő értékelésében, valamint viselkedésének értelmezésében; növeli az észlelt hasonlóságot, elősegíti a belső állapotok figyelembevételét, amely egy csoportközi helyzetben csökkenti a sztereotip konstrukciók használatát és a célszemély – valamint a csoport, amelybe beletartozik – értékelését pozitívan módosítja.

A következő elemzés bemutatása során, ezeket az eredményeket alapul véve abból feltételezésből indultunk ki, hogy a csoporttörténetekben a saját csoport szereplőinek mentális aktusai elősegítik a csoport perspektívájának felvételét, ezáltal hangsúlyossá teszik a csoportidentitást. A külső csoport mentális akciói csökkentik a sztereotip ismereti konstrukciók használatát és csökkentik a csoportközi konfliktust.

2.2 Vizsgálat

A bemutatásra kerülő vizsgálatban 1900 és 2007 között megjelent magyar történelem tankönyvek Osztrák-Magyar Monarchiával foglalkozó fejezeteit elemeztük a kognitív és érzelmi igék eloszlását tekintve. A tankönyveket évtizedekre lebontva vizsgáltuk, és minden évtizedből különböző szerzőktől származó tankönyv idevonatkozó részeit elemeztük.

A szövegeket a Nooj szoftver keretében vizsgálatuk és a korábban kidolgozott kognitív (Fig. 1.) [17] , valamint érzelmi (Fig. 2.) [4] gráfokat futattuk rajtuk.

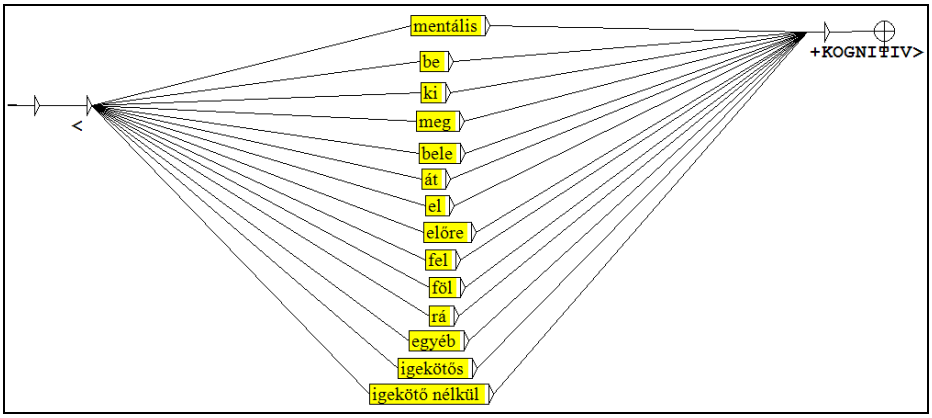


Fig. 1. Kognitív gráf

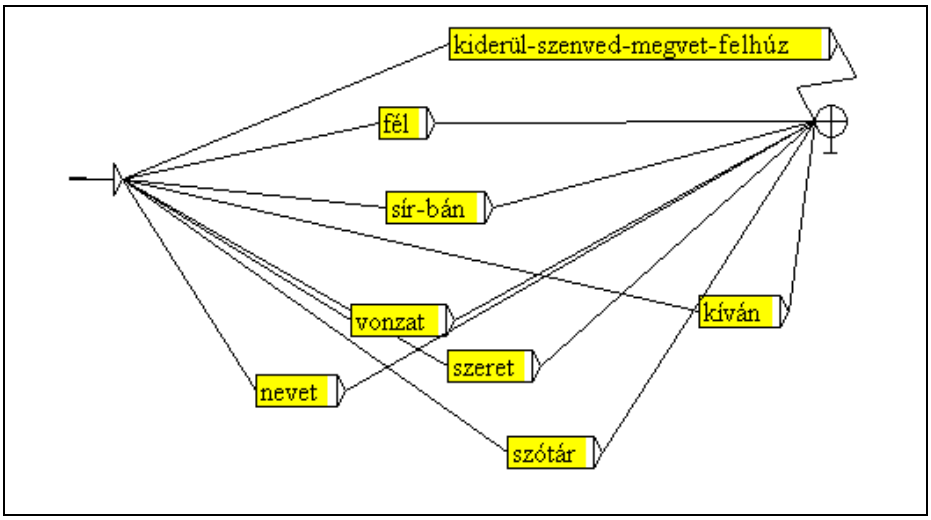


Fig. 2. Érzelem gráf

A vizsgált történelmi szövegeken nyelvi elemezést végeztünk a Nooj szoftver segítségével, amelynek eredményeképpen a kognitív és az érzelmi gráfok találatai megjelentek a szövegben (Fig. 3.).

Erzsébet királynét, nemzetünk Patrónáját is, aki mindig velünk <V ERZELEM>**könnyezett**</>, ha szenvedtünk, és velünk <V ERZELEM>**örvendett**</>, ha <V ERZELEM>**örvendtünk**</>, aki velünk <V KOGNITIV><>**érzett**</></V>jóban rosszban, s a legelső magyar Asszony volt mióta elfoglalta örök emlékű ősének, Gizellának a trónját, s nemzetünk jóságos,<V ERZELEM>**szerető**</></> szívű Anyja mióta Magyarország királynéjává avattatott. E nagy nap óta új nemzedék nőtt fel, mely tovább fejlesztette és megerősítette a nemzeti fellendülés örökét a nagy Uralkodónak védőszárnyai alatt, a ki <V KOGNITIV><>**látta** szabadságharcunkat</></V>, a <V ERZELEM>**szomorú**</> bukás után bekövetkezett sötét időket, <V KOGNITIV><>**látta**</></V> szívós kitartásunkat, hűségünket, és érezte meleg <ERZELEM>szeretetünket</> még a legzordabb években is, a kinek nemes és lovagias szíve, mert <V KOGNITIV><>**megismerte** nemzetünket</></V>,lelke pedig <V KOGNITIV><>**megértette** az idők jeleit</></V>, szintén <V ERZELEM>**szeretettel**</> és <ERZELEM>**bizalommal**</> fordult hozzánk.

Fig. 3. A kognitív és érzelem találatok megjelenítése a szövegben

Az így kapott szövegeket az Atlas-ti programmal manuális kódolás alá vetettük. Erre azért volt szükség, mivel a Nooj nyelvi fejlesztő környezetben a személyi referenciákat egyenlőre nem tudjuk kezelni. A manuális kódolás során tehát azonosítottuk a kognitív és érzelmi cselekvések alanyait. Ennek megfelelően a következő kódokat állapítottuk meg a szereplők azonosítására:

- ***magyar személy*:** meghatározott magyar személy (Kossuth, Deák, Széchenyi)
- ***magyar csoport*:** általánosan megnevezett magyar személyek nem intézményesült csoportja (hazafi, költők, írók, vezetők, politikusok, ellenállás, katonák, segédmunkások)
- ***magyar intézmény*:** megnevezett intézményesült csoport (társadalmi osztály, tisztek, rendőrség, törvény, kormány, miniszter, sajtó)
- ***magyar nemzet:*** (nemzet, magyarok)
- ***osztrák személy*:** meghatározott osztrák személy (Ferenc József, Erzsébet királyné, Haynau, Schmerling, Bach, uralkodó)
- ***osztrák csoport:*** általánosan megnevezett osztrák személyek nem intézményesült csoportja
- ***osztrák intézmény*: megnevezett intézményesült csoport cselekvése** (Bécs, hivatalnokok, rendszer, törvény)

2.3 Eredmények

A mentális állapotok, vagyis a kognitív és érzelmi kifejezések eloszlását tekintve látható, hogy összességében a magyar csoport szignifikánsan több mentális akciót hajt végre, mint az osztrák csoport (χ^2=82,474, p< 0, 05). Az eredmények idői lefutása alapján azonban észrevehető, hogy a két csoport között nincs szignifikáns különbség a század elején publikált történelemkönyvek esetében. A ’40-es évektől kezdődően ez a kiegyenlített eloszlás felborul, és a mentális állapotok szignifikánsan gyakrabban jelennek a magyar szereplők esetében[2]. A fentebb vázolt felvetésünk tükrében értelmes feltételezésnek tűnik, hogy az Osztrák-Magyar Monarchia idejében íródott történelemtankönyvekben az események narratív ábrázolása a társadalomban reálisan élő identitásszükségletnek megfelelően történik. Azaz a mentális állapotok kiegyensúlyozott megjelenítése a perspektíva-felvételt mind a két nemzet esetében elősegíti (1. Táblázat).

[2] 1940-ben χ^2=5,000 p<0,05; 1950-ben χ^2=43,085 p<0,05; 2000-ben χ^2=14,727 p<0,05

1. Táblázat: mentális állapotok nyelvi markereinek eloszlása a magyar és az osztrák szereplők között

Évtized	Szószám	Mentális állapotok	
		Magyar szereplők	Osztrák szereplők
1900	5301	49	40
1910	7070	65	45
1920	4000	32	19
1930	7091	50	30
1940	10970	67*	44
1950	21413	139*	49
2000	10352	62*	26
összesen		**439***	**280**

Az eredmények részletesebb elemzéséből az is kitűnik, hogy a mentális állapotok kognitív (gondol, remél, emlékszik, hisz) valamint érzelmi (szeret, gyűlöl, bízik, hű) nyelvi márkerei tekintetében is szignifikáns eltérés mutatkozik a két csoport között. A magyar szereplők belső, kognitív történései (χ^2=29,097 p<0,05) valamint érzelmi megnyilvánulásai (χ^2=32,814 p<0,05) is szignifikánsan gyakoribbak, mint az osztrák szereplőké. Az egyes korszakokat tekintve észrevehető, hogy a század elején a szereplők belső, kognitív állapotai kiegyensúlyozott eloszlása mellett, az osztrákokhoz való viszony megjelenítésének igénye is tükröződik az érzelmi igék előfordulásának szignifikáns eltérésében (1900-ban χ^2=9,757 p<0,05, 1910-ben χ^2=6,095 p<0,05). Az eseménytől való idői távolodás a kognitív igék gyakoriságának eltérésén keresztül hangsúlyosabbá teszi a magyar csoport perspektíváját (1930-ban χ^2=4,898 p<0,05, 1950-ben χ^2=22,915 p<0,05, 2000-ben χ^2=15,211 p<0,05) (2. Táblázat).

2. Táblázat: kognitív és érzelmi nyelvi markerek eloszlása a magyar és az osztrák szereplők között

évtized	szószám	Kognitív igék		Érzelmi igék	
		Magyar szereplők	Osztrák szereplők	Magyar szereplők	Osztrák szereplők
1900	5301	21	31	28*	9
1910	7070	36	32	29*	13
1920	4000	17	10	15	9
1930	7091	38*	21	12	9
1940	10970	51	36	16*	8
1950	21413	85*	33	44	16
2000	10352	55*	21	7	5
összesen		**331***	**206**	**162***	**74**

A belső, mentális állapotok nyelvi kifejezései olyan narratív eszközök, amelyek az empátián keresztül elősegítik a szereplők és az általuk képviselt csoport perspektívájának felvételét [19], valamint megjelenítik az érzelmi viszonyulások minőségét. Identitás-közvető szerepük abban nyilvánul meg, hogy a csoporttörténet tartalmait - a szereplők és viselkedéseik értékelését, valamint cselekvőképességüket – hangsúlyosabbá teszik. Az Osztrák-Magyar Monarchia időszakát az eseménnyel egy időben történő, valamint attól időben távolodó ábrázolása kongruens módon jeleníti meg, az aktuális és a visszatekintő identitásszükségletnek megfelelően.

Bibliográfia

1. Batson, C. D.: Prosocial motivation: Why do we help others? In A. Tesser (Ed.), Advanced social psychology. McGraw-Hill, Boston (1994) 333-381
2. Coats, S., Smith, E. R., Claypool, H., & Banner, M.: Overlapping mental representations of self and in-group: Response time evidence and its relationship with explicit measures of group identification. Journal of Experimental Social Psychology, 36, (2000) 302-315
3. Davis, M. H., Conklin, L., Smith, A., & Luce, C.: Effect of perspective taking on the cognitive representation of persons: A merging of self and other. Journal of Personality & Social Psychology, 70, (1996) 713–726
4. Fülöp, É., László, J.(2006): Az elbeszélések érzelmi aspektusának vizsgálata tartalomelemző program segítségével. *IV. Magyar Számítógépes Nyelvészeti Konferencia konferenciakötete*, Juhász Nyomda, Szeged, (2006), 296-304
5. Galinsky, A. D., & Moskowitz, G. B.: Perspective taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. Journal of Personality and Social Psychology, 78, (2000) 708–724
6. Kozak, M. N., Correll, J., & Doan, T. Mind Attribution and Prejudice (megjelenés alatt)
7. Leyens, J. Ph., Paladino,P.M., Rodriguez, R. T.,Vaes, J.,Demoulin, S., Rodriguez, A. P., et al.: The emotional side of prejudice: The role of secondary emotions. Personality and Social Psychology Review, 4, (2000) 186-197
8. Maas A., Salvi, D., Arcuri, L., & Semin, G. R.: Language use in intergroup contexts: The linguistic intergroup bias. Journal of Personality and Social Psychology, 57, (1998) 981-993
9. Richardson, D. R., Hammock, G. S., Smith, S. M., Gardner, W., & et al.: Empathy as a cognitive inhibitor of interpersonal aggression. Aggressive Behavior, 20 (4), (1994) 275-289.
10. Savitsky, K., Van Boven, L., Epley, N., & Wight, W.: The unpacking effect in responsibility allocations for group tasks. Journal of Experimental Social Psychology, 41, (2005) 447–457
11. Jones, E. E., & Nisbett, R. E.: The actor and the observer: Divergent perceptions of the causes of behavior. In: E. E. Jones (Ed.), Attribution: Perceiving the Causes of Behavior (pp. 79–94). Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc. (1971)
12. Fiske, S. T., & Taylor, S. E.: Social Cognition. (2nd ed.). McGraw-Hill. New York (1991)
13. Regan, D. T., & Totten, J.: Empathy and attribution: Turning observers into actors. Journal of Personality & Social Psychology, 32 (5), (1975) 850-856

14. Ross, L.: The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology*, 10, Academic Press. New York (1977)
15. Smith, E. R., & Henry, S.: An in-group becomes part of the self: Response time evidence. Personality and Social Psychology Bulletin, 22, (1996) 635-642
16. Tropp, L. R. & Wright, S. C.: Ingroup identification as inclusion of ingroup in the self. Personality and Social Psychology Bulletin, 27, (2001) 585-600
17. Vincze O., László J.: A mentális igék szótára, valamint alkalmazása az automatikus tartalomelemzésben. *IV. Magyar Számítógépes Nyelvészeti Konferencia konferenciakötete*, Juhász Nyomda, Szeged, (2006) 339-349
18. White C. S., Aron, A., Brook, S., Tropp, R. L.: Hogyan teszünk más embereket és csoportokat énünk részévé? Énkiterjesztés és csoportközi viszonyok. In: Forgas P. J. és Kipling D. W. (szerk.) A társas én. Az önmegismerés szociálpszichológiája. Kairosz Kiadó. Budapest (2001) 43-56
19. Liu, J.H., László, J.:A narrative theory of history and identity: Social identity, social representations, society and the individual. In: Moloney G., and Walker, I.: Social Representation and Identity; Content, Process and Power. (megjelenés alatt)

VIII. Poszter– és laptopos bemutatók

Az első magyar nyilvános, internetes beszédadatbázis bemutatása

Abari Kálmán[1], Olaszy Gábor[2]
[1] Debreceni Egyetem, Pszichológia Intézet és Matematikai és Számítástudományi Doktori Iskola
abarik@delfin.unideb.hu
[2] BME Távközlési és Médiainformatikai Tanszék
olaszy@tmit.bme.hu

1 Számítógépes bemutató

Az adatbázis címe: http://fonetika.nytud.hu/cvvc

A szógyűjtemény az általános beszédkutatás és az oktatás segítésére készült [1]. Célja, hogy bemutassa, és közzé tegye magyar a beszéd szegmentális akusztikai szerkezetét, a koartikulációs hatásokat a hangok egymáshoz való kapcsolódásánál. Használhatja bárki helytől és időtől függetlenül. Az adatbázis egy korábbi fejlesztés továbbfejlesztett változata [2].

Az adatbázisa szavakat tartalmaz, minden szót férfi és női ejtésben. Egy-egy szó a hangsorából adódó hangoknak és azok összekapcsolódásainak akusztikai bemutatására szolgál. A hangkapcsolatok 9 magánhangzó és 25 mássalhangzó kombinációit jelentik. Nem tettünk különbséget a hosszú és a rövid hangok között. Az adatbázis tartalmazza a lehetséges hangkapcsolódási formák mindegyikét, a CV,VC, a VV, a VVV, és a CC, CCC, CCCC kapcsolatokat. A CV és VC kapcsolatokból az elméletileg lehetséges 2x225-ből összesen 224, illetve 216-féle kapcsolatra ad példát az adatbázis. A VV kapcsolatok mindegyikére adunk példát (81 db), VVV kapcsolatokból 15-féle hangkapcsolódásra, a négy- és ötelemű magánhangzó kapcsolatokra 1-1 mintaszót tartalmaz az adatbázis. A mássalhangzó-kapcsolatok tekintetében pedig a CC kapcsolatokból minden kapcsolatra (373 db), a több elemű mássalhangzó-kapcsolatokból pedig a leggyakoribbakra (525 db) adnak példát a mintaszavak. Az adatbázisban összesen 1124 mintaszó található. A mintaszavakból számított artikulációs sebesség átlaga 10,5 hang/s.

Az adatbázis minden mintaszóra a következő adatokat tartalmazza: a szó szöveges (karakteres) alakja, a szót alkotó hangsor hangjainak szimbólumai (saját hangjelekkel), a szó hullámformája, a hanghatárok jelzői, valamint a szó akusztikus diagramjai. Ez utóbbiak előállításában a Praat 4.0 fonetikai szoftvert használtuk. A saját hangjelek a számítógépes feldolgozás megkönnyítését szolgálják, a legtöbb esetben megegyeznek a hang betűjelével, kivéve néhány hangot. A kivételek a következők:

á= A: , ü=U, é=E:, ö=O, gy=G, ty=T, ny=N, sz=s, s=S, zs=Z, cs= C.
A kettőspont a hosszú hangot jelöli. Példa: *gyáros* = GA:roS

Az adatbázis szolgáltatásai

Az adatbázis használatát egy erre a célra fejlesztett, felhasználó barát kereső és megjelenítő program könnyíti. A keresővel kikereshetők az kívánt hangkapcsolatokat tartalmazó mintaszavak, a megjelenítővel pedig a szó hullámformája tehető láthatóvá, továbbá a szó akusztikai képei (spektrogram, inteniztásmenet, hanghatárok, a szót felépítő beszédhangok szimbólumai). Az akusztikai diagramokon méréseket is végezhetünk (formánfrekvenciák, intenzitás értékek, hangidőtartamok stb.). Két szó akusztikai diagramjai is összehasonlíthatók (egymás alatt jelennek meg a képernyőn).

Bibliográfia

1. Abari, K., Olaszy, G,. A magyar beszéd hangkapcsolódásainak bemutatása az interneten. *Beszédkutatás 2007*.
2. Abari Kálmán – Olaszy Gábor: *Beszédadatbázis a magyar mássalhangzó kapcsolódások akusztikai szerkezetének bemutatására*. In: IV. Magyar Számítógépes Nyelvészeti Konferencia. Szerk.: Alexin Zoltán és Csendes Dóra. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, 2006.

A frázisstrukturált Szeged Treebank átalakítása függőségi fa formátumra

Alexin Zoltán

Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék,
H-6720 Szeged, Árpád tér 2.
e-mail: alexin@inf.u-szeged.hu

Abstract. A CoNLL (Conference on Computational Natural Language Learning) nemzetközi konferencia szervezői évről évre különböző versenyeket írnak ki a résztvevők számára. Az elmúlt években került sor, például a tagmondatokra bontás (2001), tulajdonnév felismerés (2003), szemantikus szerep annotáció (2005) témakörében feladatok kiírására. A 2007. nyarán Prágában megrendezésre került konferencia versenyfeladványa a függőségi struktúrák gépi tanulása volt. A kitűzött feladatok között a magyar Szeged Treebank 2.0-ból kialakított tréning adatbázis is szerepelt. A versenyben egymástól nyelvészetileg rendkívül eltérő nyelvekre készített adatbázisok vettek részt (arab, baszk, katalán, kínai, cseh, angol, görög, magyar, olasz, török), amelyek 9 nyelvcsaládból származtak (sémi, elszigetelt, újlatin, kínai-tibeti, szláv, germán, hellén, finn-ugor, török). A szerző a függőségi fa formára történő automatikus gépi átalakítást mutatja be, valamint a verseny eredményeként kapott néhány megállapítást a nyelvcsaládokra vonatkozóan.

1. Bevezetés

A CoNLL 2007 konferencia szervezői megkeresték a világ különböző országaiban működő kutatókat, nyelvi korpuszok fejlesztőit, hogy nyújtsanak segítséget egy a függőségi struktúrák gépi tanulása témában kiírandó verseny feladatainak kialakításában. Nyújtsanak segítséget egy tréning és egy teszt adatbázis kialakításában. A kívánt tréning adatbázis mérete 50-100 ezer token, a teszt adatbázis mérete 5-10 ezer token volt. A magyar Szeged Treebank 2.0 [1] készítői is kaptak egy ilyen felkérést. A szervezők megkülönböztetett érdeklődést mutattak egy új, struktúrájában merőben más, eddig számukra ismeretlen nyelv iránt. A Szeged Treebank azonban a mondatok elemzését frázisstrukturált formában tartalmazza, amelyet egy szűk időkeretben függőségi fa formára kellett átalakítani.

A frázisstrukturált korpuszban a mondatok tagmondatokból felépülő hierarchikus struktúrát alkotnak. Maguk a tagmondatok pedig igékre, az igék vonzataira (ezek névszói szerkezetek) és egyéb alkotóelemekre bonthatók, amelyek az egyes szinteken belül azonban nem alkotnak hierarchiát. A függőségi fa formátum ettől abban tér el, hogy minden egyes szó a mondatban szigorúan egy másik szó alárendeltségében van. A mondatfa csúcsán egy mesterséges gyökér elem (ROOT) található, amelynek alá-

rendeltjei lesznek a mondatban előforduló szavak. A függőségi fában szereplő kapcsolatokat címkékkel is ellátják, amelyek a kapcsolat jellegére utalnak.

A verseny számára a Népszabadság és a HVG folyóiratok egy-egy számából kialakított korpuszrészletet választották ki, amely méretben megfelelt az elvárásoknak és a nyelvezete is kellően stabil volt. A verseny céljaira kialakított függőségi fa korpuszt a szervezők által biztosított segédprogramokkal ellenőrizték.

A szervezők a verseny eredményeinek összefoglalását egy hosszú tanulmányban tették közzé [2]. A résztvevők különböző gépi tanuló algoritmusokat használtak. A kapott eredmények azonban azt mutatták, hogy nem annyira az alkalmazott módszer, hanem az egyes nyelvcsaládok jellegzetességei, és az adott korpusz határozza meg a tanulás sikerességét. Így a nyelveket a tanulhatóság szempontjából három csoportba lehetett sorolni. A magyar nyelv a középső osztályba került, ami fontos visszajelzés arról, hogy a treebankben az igei vonzatok annotációja jó minőségű és releváns információt hordoz, illetve, hogy a konverzió algoritmusa kellően stabil. Ez lehetőséget adhat arra, hogy e munka eredményeként a teljes Szeged Treebank 2.0-át automatikus módszerekkel függőségi fa alakúra konvertáljuk.

2. A konverzió

A Szeged Treebank tartalmazza a mondatokban szereplő igék vonzatainak megjelölését, valamint egy a kapcsolat jellegére utaló címkét is. Az átalakítás fő feladata az volt, a vonzatokban kódolt függőségeket a konvertáló program vonja ki az adatbázisból, a nem jelölt függőségeket pedig automatikusan határozza meg.

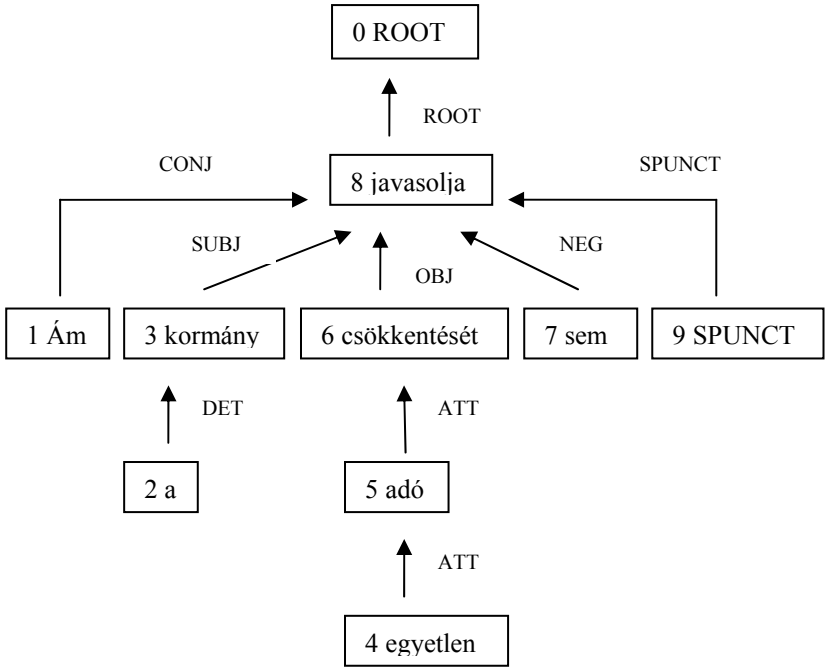
```
1       Ám      ám      C       Cc      ctype=coordinating      8       CONJ
2       a       a       T       Tf      def=yes 3       DET
3       kormány kormány N       Nc
                        n=singular|case=nominative|proper=no 8       SUBJ
4       egyetlen        egyetlen        A       Af
        deg=positive|n=singular|case=nominative     5       ATT
5       adó     adó     N       Nc
        n=singular|case=nominative|proper=no 6       ATT
6       csökkentését    csökkentés      N       Nc
        n=singular|case=accusative|proper=no|pperson=3rd|pnumber=singula
r       8       OBJ
7       sem     sem     R       Rm      _       8       NEG
8       javasolja       javasol V       Vm
        mood=indicative|t=present|p=3rd|n=singular|def=yes  0       ROOT
9       .       _       SPUNCT  SPUNCT  _       8       PUNCT
```

1. ábra. A tréning adatfájl egy részlete (HVG.2.6.4 mondat)

A treebankben az egyes névszói szerkezetek belső felépítése részben hierarchikus (az egymással birtokviszonyban álló névszók esetén), az egyes névszókhoz tartozó névelő, számnevek és jelzők viszont a hierarchia jelölése nélkül szerepelnek a frázis fejében lévő kulcs névszóval, alapvetően egy főnévvel közös szerkezetben. A főnevek bővítményeit a konvertáló program automatikusan a főnév alárendeltségébe tette. Névutós szerkezet esetén a konvertáló program a névszót tette a névutó alárendeltségébe – ez volt a legegyszerűbb és leginkább kézenfekvő megoldás.

Az algoritmus számára a legnehezebb feladat az volt, amikor egy tagmondaton belül több igei szerkezetet is talált, és a vonzatok felváltva tartoztak az egyes igékhez (igenevekhez) valamint az igék (igenevek) között is függőségi kapcsolat állt fent. Az elkészített program három igei szerkezetet tud kezelni egy tagmondatban, ami a tesztek alapján elegendő volt.

Az 1. ábrán látható a konvertált állomány egy kis részlete. A mondat szavai 1-től kezdve egy-egy sorszámot kapnak. A ROOT elem a 0-s sorszámot kapja. A táblázatos (tabulátor karakterekkel tagolt fájl) egyes oszlopai a következők: sorszám, ortográfia, szótári alak, a morfo-szintaktikai kód (MSD) első betűje, részletes morfo-szintaktikai kód (a verseny szervezői kérték, hogy az MSD kódokból itt csak az első két karakter szerepeljen), további fontos lexikális tulajdonságok | (bar) karakterrel elválasztva, a fölérendelt szó sorszáma, a függőségi kapcsolat jellege.



2. ábra. A konvertáló program által előállított függőségi fa

A konverziót megvalósító program C# programozási nyelven készült és alapvetően heurisztikus eszközökkel oldotta meg a feladatot. Az idő rövidsége miatt került sor ennek az alkalmazására az XSLT transzformációs eszköz helyett. A program tagmondatonként rekurzívan végezte el a konvertálást. Első lépésben megszámolta a tagmondatban szereplő igék (igenevek) számát. Ha volt ige, akkor megkereste a hierarchiában legfelsőt, és elhelyezte a többi igét (igenevet) ez alá, majd pedig végighaladt a bővítményeken és ezeket is a megfelelő ige alá helyezte el. A gazdátlan bővítmények a legfelső szintű ige alárendeltjei lettek. Főnévi szerkezetek esetén a jobb oldali

első főnevet (fejet) tette az algoritmus a legfelső pozícióba és ez alá helyezte el a frázis további elemeit. A névelő függőségi címkéje DET, a jelzőké ATT lett.

A rendszerben természetesen előfordulhattak hibák. A rövid 2-3 hetes határidő miatt nem volt mód minden hibajelenség okát felkutatni, ezért a rendszer a szerkezeti hibás mondatokat egy belső ellenőrző eljárással észlelte és az output állományból egyszerűen törölte, ezek száma azonban nem volt túl sok: HVG: 9 mondat (2179-ből), Népszabadság 38 mondat (3905-ből).

3. Eredmények

A konvertált tréning és teszt állományt a szervezők a verseny résztvevőinek kiadták és 21 csapattól érkezett eredmény a magyar nyelvre. Nagyon sok különböző módszert alkalmaztak: pld. SVM, véges Newton SVM, maximum entrópia, átlagolt perceptron, maximum likelihood, HMM alapú módszereket stb. a függőségi relációk predikciójára. A kapott eredmények azt mutatták, hogy nem annyira az alkalmazott módszer, hanem az egyes nyelvcsaládok jellegzetességei és az adott korpuszok határozzák meg a tanulás sikerességét. Így a nyelveket a tanulhatóság szempontjából három csoportba lehetett sorolni az elért átlagos pontosság alapján. Nehezen tanulható (76,31-76,94%): arab, baszk, görög. Közepesen jól tanulható (79,19-80,21%): cseh, magyar, török. Jól tanulható (84,40-89,61%): katalán, kínai, angol, olasz.

Az eredmények elemzése azt mutatta, hogy ez az eredmény sok esetben a korpusz nagyságával és minőségével van kapcsolatban, hiszen az arab és az újgörög nem annyira nehéz nyelvek. Mégis megelőzte őket a magyar és a cseh, amelyek nagymértékben ragozók és meglehetősen szabad szórendet engednek meg. A szervezők megvizsgálták az ismeretlen szavak arányát a teszt állományokban, amely érték a magyar és a török nyelvek esetén volt a legmagasabb. Ez nem rontotta le a magyar eredményt – a sok ismeretlen, új szót ellensúlyozni tudta a korpusz mérete.

A verseny eredményeit az [2] irodalom foglalta össze, amelyből kitűnik, hogy a heurisztika ellenére a magyar függőségi treebankkel kapott eredmények a középmezőnyben foglaltak helyet. Az elkészített program nemcsak konvertálására alkalmas, hanem arra is, hogy a Szeged Treebankben meglevő annotálási hibák után nyomozzon, segítsen ezek felkutatásában és kijavításában. A jövőben szeretnénk ezt a lehetőséget is felhasználni a treebank minőségének javítására.

4. Irodalom

[1] Csendes D., Csirik J., Gyimóthy T., Kocsor A.: The Szeged Treebank, in Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005), Karlovy Vary, Czech Republic 12-16 September, and LNAI series Vol. 3658, pp. 123-131 (2005)

[2] Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, Deniz Yuret: The CoNLL 2007 Shared Task on Dependency Parsing, in Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, pp. 915–932, Prague, (2007)

Egy egyszerű módszer modális beszéd glottalizálttá alakítására

Bőhm Tamás, Németh Géza

BME Távközlési és Médiainformatikai Tanszék, 1117 Budapest, Magyar Tudósok krt. 2.

Kivonat: A beszédtechnológia számos területén rendkívül hasznos lenne az irreguláris hangszalagrezgés (glottalizáció) megfelelő kezelése. Cikkünk ilyen irányú munkánk első eredményeit ismerteti: egy félautomatikus eljárást mutatunk be, amely képes modális (nem glottalizált) beszédet glottalizálttá alakítani. A korábbi algoritmusok a jitter növelésével próbálták mesterségesen érdes hangzásúvá tenni a beszédet. Módszerünk ezzel szemben alapperiódusok kitörlésével éri el ezt a hatást, amit a megmaradó periódusok amplitúdójának perturbálásával tovább erősít. Formális meghallgatásos tesztben kimutattuk, hogy az így előállított beszédjel természetes hangzású és hasonlóan érdes, mint a természetes glottalizált felvételek.

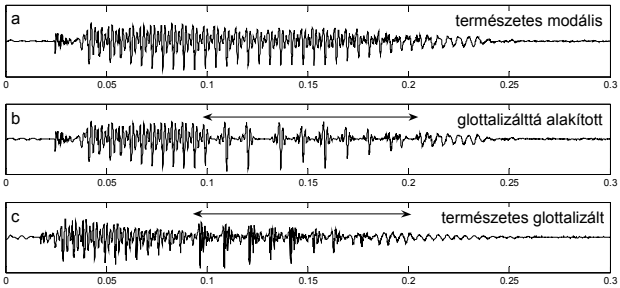
1 Bevezetés

A zöngeképzés, a fonáció során a hangszalagok általában közelperiodikusan (kváziperiodikusan) rezegnek. Ilyenkor a hangszalagok nagyjából állandó időközönként összecsapódnak – a rezgés reguláris, vagy más néven modális. Rövidebb-hosszabb ideig azonban ez a rezgés irregulárissá válhat, azaz hirtelen változások jelentkezhetnek a rezgés pillanatnyi frekvenciájában, amplitúdójában vagy mindkettőben. Ezt a jelenséget nevezzük *glottalizációnak* (1.*c)* ábra), ami gyakori jelenség mind egészséges, mind sérült gégefunkcióval rendelkező beszélők esetén. Gyakran rendkívül alacsony alapfrekvencia és a glottális impulzusok gyors lecsengése kíséri. A glottalizációt érdes, rekedt hangként érzékeljük. A jelenség produkciós és percepciós hátterének áttekintése [1]-ben olvasható.

A glottalizáció előfordulása függ a prozódiai szerkezettől (például gyakran egybeesik prozódiai egységhatárokkal és hangsúlyos szótagokkal [2]) és információt hordoz a beszélő személyéről, nyelvjárásáról [4], hangulatáról és érzelmi állapotáról [3]. Így a glottalizáció megfelelő manipulációja hozzájárulhat természetes hangzású, érzelmi töltettel rendelkező és személyre szabott beszédszintézis rendszerek építéséhez.

Azonban még nincs általánosan alkalmazható módszer modális beszéd glottalizálttá alakítására és fordítva. Számos kísérlet történt glottalizáció előállítására formánsszintézissel (pl. [9]), de ezek több tucat szintézisparaméter kézi beállítását igénylik. A másik módszer egy természetes beszédfelvételen a glottális impulzusok időzítésének perturbálása, azaz a jitter növelése [7,10]. Általánosan elfogadott tény, hogy a jitter összefügg a beszéd érdességével, de sokszor az összefüggés csak áttételes [6] és a különböző típusú perturbációk máshogy befolyásolják az érzeti zöngemi-

nőséget [5]. Ezért ebben a tanulmányban más megközelítést választottunk: úgy próbáljuk a glottalizációra jellemző hullámformát előállítani, hogy egyes kézzel kiválasztott alapperiódusokat kitörlünk a jelből, más periódusokat pedig felerősítünk vagy csillapítunk. Bár módszerünk jelentősen növeli a jelben a jittert, a glottalizáció számos más akusztikai jellegzetességét is reprodukálja.



1. ábra. Egy női bemondótól származó természetes modális (*a*) és glottalizált (*c*) felvétel hullámformája, valamint a modálisból mesterségesen glottalizálttá alakított változat (*b*). A nyilak a glottalizált részeket jelölik.

2 A módszer leírása

Módszerünk hasonlít a PSOLA algoritmusra [8]. Az alkalmazott analízis és szintézis megegyezik azzal, a különbség a manipulációban van: míg a PSOLA az alapfrekvencia módosítása érdekében időben elcsúsztatja az alapperiódusokat, esetünkben ehelyett az egyes periódusok amplitúdóját változtatjuk.

Analízis. A módszer bemenete a beszédjel és a glottális impulzusok időpontjai, azaz a *pitchmark*-ok. Az analízis célja, hogy a jelet szétbontsa alapperiódusokra. Ezt a megfelelő pitchmark környezetének kiablakozásával éri el. Egy olyan Hanning ablakkal szorozza be a beszédjelet, aminek a csúcsa az aktuális pitchmark-on van és az előzőtől a következő pitchmark-ig tart (tehát két alapperiódust fed le).

Manipuláció. Minden egyes kiablakozott alapperiódust beszorzunk egy kézzel beállított s faktorral. Így a periódusokat egyenként felerősíthetjük ($s > 1$), csillapíthatjuk ($s < 1$), kitörölhetjük ($s = 0$) vagy akár módosítás nélkül meghagyhatjuk ($s = 1$). Egyegy periódus kitörlésével érdes hangzás érhető el. A hatás több egymás utáni periódus törlésével és az amplitúdók perturbálásával fokozható. A cél a természetes glottalizációra hasonlító hosszú és irreguláris alapperiódusok létrehozása.

Szintézis. A módosított beszédjelet a faktorokkal megszorzott periódusok átfedve összeadásával (overlap-and-add) kapjuk meg. Ha nem végeztünk semmilyen manipulációt, akkor a kerekítési hibától eltekintve visszakapjuk az eredeti jelet.

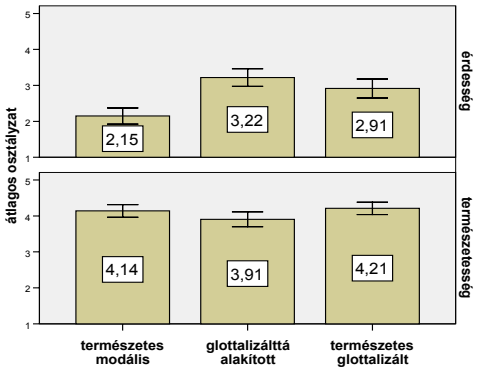
Az 1. ábra *b)* részén az *a)* részen látható felvétel glottalizálttá alakított változata látható. Összehasonlításképpen az ábra *c)* része egy természetes glottalizált felvételt ábrázol, ugyanannak a bemondónak az ejtésében.

3 Értékelés

Egy meghallgatásos kísérlettel értékeltük az eljárást. Négy rövid szót vettünk fel, amit két bemondó modálisan és glottalizált befejezéssel is felolvasott. Módszerünkkel a modális szavak végét glottalizálttá alakítottuk. Megpróbáltuk a bemondó természetes glottalizált kiejtésének glottális impulzus-időzítéseit és -amplitúdóit reprodukálni. A manipulált felvételt meghallgatva iteratívan finomítottuk a beállított *s* faktorokat.

A 12 kísérleti személy két külön tesztet végzett el: az egyikben az érdesség, a másikban a természetesség szempontjából kellett értékelniük a felvételeket. Egyesével, véletlenszerű sorrendben hallották a természetes modális, a természetes glottalizált és a mesterségesen glottalizálttá alakított hanganyagokat[23]. Ezeket egy ötpontos skálán kellett osztályozniuk (1: nagyon természetellenes/egyáltalán nem érdes; 5: nagyon természetes/nagyon érdes). A kísérlet előtt végighallgatták az összes hanganyagot, valamint hallottak pár nagyon érdes és egyáltalán nem érdes példát.

Négy kísérleti személy a természetes glottalizált felvételeket *kevésbé* érdesnek ítélte meg, mint a természetes modálisokat, így osztályzataikat a továbbiakban nem elemeztük. A maradék nyolc személy esetén egyutas varianciaanalízist (ANOVA) végeztünk a hanganyag típusának függvényében, külön a természetesség és külön az érdesség osztályzatokra. Az egyes felvételtípusokhoz tartozó átlagos osztályzatok a 2. ábrán láthatóak. Az alábbiakban részletezett különbségek Tukey-féle post hoc tesztek alapján 5%-os szinten szignifikánsak.



2. ábra. A meghallgatásos kiértékelés során kapott átlagos osztályzatok. A függőleges szakaszok az átlagokhoz tartozó 95%-os konfidencia-intervallumokat jelölik.

Ahogy várható volt, a természetes glottalizált felvételek érdesebb hangzásúak, mint a természetes modálisok. A természetes modális felvételeket glottalizálttá alakítva 1,07 osztályzattal nőtt az érdesség, ami így már nem tér el szignifikánsan a célértéktől (a természetes glottalizált hanganyagok érdességétől). Az átalakítás azonban

[23] A kísérleti személyek más típusú (pl. formánsszintetikus) hanganyagokat is értékeltek, amelyek egy másik tanulmány részét képezték, itt irrelevánsak.

csak elhanyagolható, nem szignifikáns romlást okozott a természetesség megítélésében.

4 Összefoglalás

Egy olyan egyszerű, félautomatikus, pitch-szinkron beszédfeldolgozási eljárást mutattunk be, amely képes modális beszédet glottalizálttá alakítani egyes alapperiódusok kinullázásával, valamint más periódusok amplitúdójának megváltoztatásával. A meghallgatásos értékelés azt mutatta, hogy a transzformált felvételeket a hallgatók hasonlóan érdesnek és természetesnek ítélték meg, mint a természetesen glottalizáltakat. Mivel az alkalmazott analízis és a szintézis módszerek megegyeznek a PSOLA algoritmus megfelelő feldolgozási szakaszaival, módszerünk könnyen integrálható PSOLÁ-t alkalmazó beszédszintetizátorokba. Ehhez azonban az eljárás automatizálása, azaz az *s* faktorok algoritmikus beállítása, valamint az ellentétes irányú transzformáció megvalósítása is szükséges. A szerzők jelenleg ebben a két irányban folytatják a munkát.

Bibliográfia

1. Blomgren, M., Chen, Y., Ng, M.L., Gilbert, H.R.: Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. JASA 103 (1998) 2649–2658
2. Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M.: Glottalization of word-initial vowels as a function of prosodic structure. J. Phonetics 24 (1996) 423–444
3. Gobl, C., Ní Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude. Sp. Comm. 40 (2003) 189–212
4. Henton, C.G., Bladon, A.: Creak as a sociophonetic marker. In: Hyman, L.M., Li, C.N. (eds.): Language, speech and mind. Routledge, London (1987) 3–29
5. Hillenbrand, J.: Perception of aperiodicities in synthetically generated voices. JASA 83 (1988) 2361–2371
6. Kreiman, J., Gerratt, B.R.: Perception of aperiodicity in pathological voice. JASA 117 (2005) 2201–2211
7. McCree, A.V., Barnwell, T.P.: A mixed excitation LPC vocoder model for low bit rate speech coding. IEEE Trans. Speech and Audio Proc. 3 (1995) 242–249
8. Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Sp. Comm. 9 (1990) 453–467
9. Pierrehumbert, J.B., Frisch, S.: Synthesizing Allophonic Glottalization. In: van Santen, J.P.H. et al. (eds.): Progress in Speech Synthesis. Springer, New York (1997) 9–26
10. Verma, A., Kumar, A.: Introducing rougness in individuality transformation through jitter modeling and modification. Proc. ICASSP (2005) 5–8

A szerzők szeretnék köszönetüket kifejezni Stefanie Shattuck-Hufnagelnek szakmai útmutatásáért és a Fulbright Bizottságnak a kísérleti személyek toborzásáért. A kutatást részben az NKFP 2. keretprogramja (szerződésszám: 2/034/2004) támogatta.

Magyar nyelvű beszédfelismerő rendszer diszkriminatív tanítása

Gyepesi György és Serény András

Alkalmazott Logikai Laboratórium
Budapest, Hankóczy J. utca 7. 1022
e-mail:{ggyepesi,sandris}@all.hu

A legtöbb modern beszédfelismerő rendszer az akusztikus jelből kivont, az adott hangrészletet jellemző feature-vektorok osztályozására rejtett Markov-modelleket (HMM) alkalmaz. A nagyszótáras rendszerekben általában az egyes fonémákat (esetleg környezettől függően) modellezik egy–egy HMM-mel, és alapvető feladat ezen HMM-ek paramétereinek beállítása, a tanítás. Ehhez tanító minta szükséges, ami egy (hosszú) beszéddarabból kivont feature-vektor sorozatból (ezt nevezzük megfigyelésnek) és helyesen átírt változatából áll.

Jelölje a megfigyelést o, a helyes átírást s_r, a tanítás során keresendő paraméterek összességét λ. A paraméterek beállítására a szokásos eljárás a maximum likelihood becslés (MLE), vagyis adott o megfigyelés és s_r helyes megoldás mellett keresendő λ azon értéke, melyre a modell által definiált $P_\lambda(o \mid s_r)$ valószínűség maximális; ennek megvalósítására egy hatékony eljárás a Baum–Welch algoritmus.

A diszkriminatív tanító eljárások az MLE módszer alternatíváit adják a paraméter-beállítás feladatának megoldására. Az az alapvető elképzelés, hogy a helyes megoldás mellett figyelembe vesszük a beszédfelismerő kimenetén megjelenő rossz megoldásokat is, és keressük λ egy olyan értékét, ami nem csupán azt biztosítja, hogy a helyes megoldás valószínűsége nagy legyen, de a helytelen megoldások valószínűségét is alacsonyan tartja. Innen adódik az eljárás neve: a helyes megoldást igyekszik megkülönböztetni a helytelenektől. Bár a diszkriminatív módszer kezdeti változatait HMM-ek tanítására már a nyolvanas évek végén kidolgozták, a magyar nyelvű beszédfelismerésben való alkalmazására – tudomásunk szerint – ez az első kísérlet.

A különféle diszkriminatív tanító eljárások közül igen ígéretes a Povey [1] által bevezetett minimum phone error (MPE) tanítás, ennek egy változatát valósítottuk meg és az alábbiakban ezt vázoljuk. Legyen S az összes lehetséges megoldás (mondat) halmaza, és valamely $s \in S$ esetén legyen $A(s, s_r)$ az s mondatnak a helyes megfejtéshez képest fonémákban mért pontossága, $A(s, s_r) = |s_r| - d(s, s_r)$, ahol $|\cdot|$ a fonémák száma és $d(\cdot, \cdot)$ az edit distance. Az optimalizálandó függvényünket úgy választjuk, hogy a mondatok valószínűségét pontosságukkal súlyozzuk, így a függvény maximalizálásakor a pontosabb mondatok jobban számítanak. Formálisan, keresendő a $\lambda \mapsto \sum_{s \in S} P_\lambda(s \mid o) A(s, s_r)$, az MPE célfüggvény, maximuma. Miután az összes lehetséges mondat felsorolása

és valószínűségeik kiszámítása nem megvalósítható, a következő eljáráshoz folyamodunk. Adott (o, s_r) tanító mintán először MLE tanítást hajtunk végre és az így nyert fonéma HMM-ket használva az o bemeneten felismerést futtatunk. A felismerés kimenete az első néhány legvalószínűbb mondat, ezek összességét kompakt formában, szóhálóként ábrázoljuk. A szóhálóban minden út egy mondatot reprezentál, az itt elő nem forduló mondatokat nem vesszük számításba.

Az MLE tanításhoz használt iteratív Baum–Welch eljárás adaptálható az MPE célfüggvény maximumának keresésére. Egy iterációs lépés elején a szóhálón futó forward–backward algoritmus hatékonyan számolja egyszerre a fonémákhoz tartozó statisztikákat és az $A(s, s_r)$ pontosság egy közelítését. A diszkriminatív tanítási elv szerint minden fonémához kétféle statisztikát gyűjtünk: az egyik azt tükrözi, hogy a fonéma milyen gyakran szerepel az átlagosnál pontosabb utakon („helyes megoldások"), a másik azt, hogy milyen gyakran szerepel az átlagosnál kevésbé pontos utakon („rossz megoldások"). Az újrabecslő egyenletekben mindkétféle statisztika által hordozott információ megjelenik, a fonémaparaméterek újrabecslésével egy iteráció véget ér. Az MLE esethez hasonlóan négy–öt iteráció elegendő a konvergenciához.

A fentiekben összefoglalt eljárást implementáltuk, annak futtatása, az eredmények értékelése és a baseline-nak tekintett MLE módszer eredményével való összevetése még folyik. Povey és Woodland [2] angol nyelvű korpuszokon végzett vizsgálatai a hibásan felismert szavak arányának három–öt százalékpontos csökkenését mutatják, magyar nyelvre is hasonló eredményt várunk.

Hivatkozások

1. D. Povey., Discriminative Training for Large Vocabulary Speech Recogniton. PhD thesis, University of Cambridge, 2003.
2. D. Povey and P. Woodland. Minimum phone error and I-smoothing for improved discriminative training. In Proceedings of the ICASSP, 2002.

Végesállapotú transzducerek mindenkinek

Gyepesi György[1], Gábor Bálint[2], Halácsy Péter[3], Kertész Zsuzsa[1]

[1] ALL Kut. és Fejl. Szöv. {ggyepesi,kzsuzsa}@all.hu
[2] BME Kognitív Tudományi Tanszék, bgabor@cogsci.bme.hu
[3] BME Média Oktató és Kutató Központ, hp@mokk.bme.hu

Kivonat Cikkünkben bemutatunk két véges állapotú fordítóval működő nyílt és szabad szövegfeldolgozó komponenst, a `huntokent` flexibilitásban meghaladó tokenizálót és a `hunmorph` csomag jelenlegi elemző programjánál az `ocamorph-nál` nagyságrenddel gyorsabban futó morfológiai elemzőt. Mindkét szoftverre jellemző, hogy nagy korpuszok gyors feldolgozására készült és nem csak parancssorból lehet őket használni, hanem fejlesztői könyvtárként bármilyen alkalmazásba könnyen beilleszthetőek.

1. Tokenizáló és mondatrabontó

A neten már több tucat tokenizáló program érhető el [6], azonban ezek egyike sem felel meg néhány számunkra fontos követelménynek. Egyrészt fontos a bővíthetőség és konfigurálhatóság, mert más-más alkalmazásoknak másmilyen szegmentálásra van szükségük. Másrészt nagyon fontos szempont a sebesség, hiszen gigabájt méretű korpuszok esetén egy okos, de lassú tokenizáló használhatatlanná válik.

1.1. Standoff annotáció

A hagyományos szövegfeldolgozó csővezeték az eredeti szöveget lépésről lépésre átírja, transzformálja. A végeredmény manapság általában egy XML fájl különböző elemekkel teletűzdelve. Ezzel több probléma is van: egyrészt a formátum, az ún. inline annotáció nem teszi lehetővé, hogy többféle annotációt használjunk ugyanarra a szövegre. Ez azt is jelenti, hogy későbbiekben nem tudjuk fejleszteni vagy éppen lecserélni a feldolgozó sor valamelyik komponensét.

Kevésbé triviális, de ugyanolyan fontos probléma, hogy a monolitikus inline annotáció használata esetén az eredeti szöveget, a jelenségek eredeti kontextusát elvesztjük.

Reméljük, a fenti problémákat megoldja az egyébként már több helyen is alkalmazott ún. *standoff* annotáció. Ennek a lényege, hogy az eredeti szöveget érintetlenül hagyjuk és a szöveg mellett csak felsoroljuk, hogy a szöveg mely része milyen címkét, annotációt kapott (de nem tesszük bele a szövegbe a címkét).

1.2. Sebesség

Tokenizáláshoz a leggyakrabban alkalmazott megközelítés az egymás utáni cserélgetés. Ez általában csővezetékbe kötött `lex/flex` programok (lásd huntoken vagy a Multext segment), `sed` szkriptek vagy több egymásutáni sztring replace parancs végrehajtásával végzik, mint Greffenstette [1] mára klasszikussá vált tokenizálója. Ezekben az a közös, hogy a szöveget többször olvassák végig, tulajdonképpen több reguláris kifejezés illesztés történik egymás után.

Például a hunglish projektben használt egyszerű tokenizálóban csővezetékbe kötött `sed` szkriptek egymás után több cserét hajtanak végre. Első lépésben a mondatvége jelek után szóközt, majd – hogy a kötőjeleket leválassza a szavakról – a kötőjelek elé majd után újabb szóközt szúr be, majd a több szóközt egymás mellett egyre cseréli.

Ezzel a megközelítéssel az a baj, hogy a kontextusérzékeny beszúrás – főleg, ha regexpes sztring csere műveletként implementáljuk – eléggé lassú tud lenni.

1.3. FST alapú tokenizálás

A regexp illesztés utáni cseremműveletek egymásutánja kifejezhető *véges állapotú transzducerekkel* (FST). Ennek az a nagy előnye, hogy az egymás után fűzött FST-k az ún. kompozíció művelettel egy nagy FST-vé alakíthatóak, aminek köszönhetően egy sztring feldolgozásához a sztringet csak egyszer kell végigolvasni. Így a kompozicióval előállított transzducer kevesebb input-output műveletet igényel, és a számítási igénye is alacsonyabb. Bár az állapotok száma általában nő, ez csak az automata memóriaigényét befolyásolja.

Minden karakterhez meghatározzuk azt az FST csere-szabályt, amelyik leellenőrzi, hogy ennek a karakternek az egyes előfordulásai – a kontextusuk alapján – token- vagy mondatzáró pozicióban vannak-e. A tokenizáló javarészt ilyen csereszabályok által meghatározott transzducerek kompozíciójából áll. Vannak olyan bonyolultabb esetek, amikor távoli karakterek hatását kell figyelembe vennünk egy esetleges tokenzáró karakter vizsgálatakor. Ilyen eset például amikor zárójelek között szereplő mondatok esetén csak a zárójelen kívüli – ezeket magában foglaló – mondat végét szeretnénk mondathatárnak nyilvánítani. Ilyenkor szükség lehet olyan transzducerek használatára, amelyek ideiglenes, a későbbi automaták számára üzenetet hordozó szimbólumokat helyeznek el a szövegben. Megjegyezzük, hogy ez esetben is transzducerek kompozíciójával dolgozunk, és nem egymás után futtatjuk őket.

1.4. Implementáció

Az SFST[4] nyílt forráskódú FST fordítót használjuk a hálózat építésére, de futásidőben saját algoritmussal dolgozunk. A tokenizálás azért gyors, mert tudjuk,

[4] http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html

hogy (1) determinisztikus automatával van dolgunk, (2) a felszíni szimbólumok csak karakterek lehetnek, (3) a végállapotoknak nincs jelentősége, (4) epszilon sztringgel minden állapotból csak egyfelé lehet menni és (5) ott vannak a tokenhatárok, ahol csak egy ilyen élen lehet továbbmenni.

Az SFST-vel előállított automatát egy saját elemzőprogrammal használjuk, ami folyamatosan olvassa a szöveget, és ha tokenhatárhoz ért, akkor annak pozicióját visszaadja.

2. FST alapú szóelemző

Az affix-stripping alapú morfológiai elemzés a hunspell nyelvi erőforrásait, az un. `affix` és `dictionary` állományait használja. A morphdb kiterjeszti ezt a szabályleírási formát morfológiai információkkal, és keretet biztosít a morfológiai szabályok nyelvészeti fogalmakkal történő leírására. Az FST morfológiai elemzők modellje a kétszintű morfológia [3]. Az XFST, SFST és más FST alapú morfológiai elemzők saját, gyakran körülményes, az FST-építés technikáját is szabályozó nyelven követelik a morfológiai szabályok leírását. Emiatt nyelvészek számára nehézkes az ilyen elemzők használata és a morfológiai szabályok ilyen nyelven való megfogalmazása.

Magának az FST alapú elemzésnek számos előnye van: az FST erőforrást (a transzducer állapotainak és input-output-címkézett éleinek listája) beolvasó és a beolvasott transzduceren morfológiai elemzést végző program bármilyen programozási nyelven nagyon egyszerű és rövid. Az affix-stripping alapú elemzők ezzel szemben meglehetősen bonyolult programok. Ráadásul az FST alapú elemzés sebessége legalább egy nagyságrenddel nagyobb az affix-stripping alapúakénál. Az FST erőforráson működnek az általános transzducer algoritmusok, és más eljárások is ki tudják használni a reprezentáció egyszerűségét.

Összefoglalva: az affix-stripping alapú elemzők számára egyszerű az erőforrás elkészítése, az FST alapú elemzők morfológiai szabályreprezentációja viszont sokkal egyszerűbb és az elemző algoritmus lényegesen gyorsabb. Ezért van értelme affix-stripping erőforrásból FST építésének.

2.1. FST építés folyamata

Az affix-stripping affix és dictionary fájljaiból először egy címkézett élű és csúcsú irányított gráfot építünk. A gráf csúcsai a szótárban szereplő szavak és az affix file-ban szereplő szabályok. A csúcsok címkéje az a morfológiai annotáció, ami a szóhoz illetve toldalékhoz tartozik. Egy szócsúcsból azokba a szabálycsúcsokba vezet él, amely szabályokkal a szó toldalékolható, egy szabálycsúcsból pedig azokba a szabálycsúcsokba vezet él, amely szabályokkal a toldalékolás folytatható. Két szócsúcs között akkor van él, ha a két szó összetételben szerepelhet. Az élek címkéje írja le azt a változást, amit a morfológiai művelet kivált.

Ebből a gráfból építjük a transzducert úgy, hogy végrehajtjuk az éleken szereplő műveleteket és a műveletek eredményét írjuk a transzducer éleire mint inputot. A gráf-élek végcsúcsaiban szereplő címkék lesznek a megfelelő transzducer éleken megjelenő outputok. Problémát csak az okoz, amikor az affix fájl egy suffix szabályából megy él egy prefix szabályba (pl. igeképzőkön megjelenik az igekötő, vagyis az igeképzőnek megfelelő csúcsból él megy egy igekötő csúcsába). Ilyenkor megismételjük a gráfnak azt a részgráfját, ami a suffix szabály csúcsán végződik, a másolat minden a suffix szabályétól különböző csúcsát nem elfogadóvá tesszük a transzduceren és a másolat minden kezdőállapotába élet húzunk a prefix csúcsából.

Az így elkészített transzducert betűsítjük, azaz minden élet felbontunk olyan élekre, melyeken egy-egy karakter az input, végül „minimalizáljuk": determinizáljuk és minimalizáljuk azt az automatát, amit úgy kapunk, hogy egy speciális jellel minden élen összekötjük az inputot és outputot majd újra szétválasztjuk őket.

3. Összefoglalás

A két bemutatott FST alapú eszköz, a szó- és mondathatár bejelölő és a morfológiai elemző (tövező) új NLP funkcionalitást nem hozott a BME MOKK által fejlesztett szószablya családba, viszont mérnöki szemmel nézve nagy előrelépést jelentenek.

Egyrészt sokkal gyorsabban végzik el feladatukat, mint az eddigi eszközök és támogatják a standoff annotációt. Ennél talán fontosabb az az új lehetőség, ami a bonyolult online elemzők esetében egyáltalán nem volt adott: egy BA-t végzett informatikus képes bármilyen programozási nyelvre tokenizálót, mondatrabontót és morfológiai elemzőt írni, mert a végesállapotú transzducer online rétege max. 100 sorban megírható. Cserébe az erőforrások előállítása nehezebb, de ezt egyszer kell futtatási plattformtól függetlenül előállítani. Reméljük, e két új feljesztéssel a `huntoken` és `hunmorph` programok használhatóvá válnak nagy ipari rendszerekben is, ahol a feladat sok szöveg gyors és hatékony feldolgozása.

Hivatkozások

1. Gregory Grefenstette, Pasi Tapanainen, 'What is a word, what is a sentence? problems of tokenization'. In: The 3rd International Conference on Computational Lexicography, 79–87, Budapest, (1994).
2. Halácsy, P., Kornai, A., Németh, L, Sas, B., Varga, D., Váradi, T., and Vonyó, A: A Hunglish korpusz és szótár, In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem, (2005).
3. Koskenniemi, K..: Two-level morphology: A general computational model of word-form recognition and production. Tech. rep. Publication No. 11, Department of General Linguistics, University of Helsinki, (1983).

4. Ronald M. Kaplan: A Method for Tokenizing Text, In: *Inquiries into Words, Constraints and Contextus*, (2005).
5. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., and Simon, E.: Morphdb.hu: Hungarian lexical database and morphological grammar, In: Proceedings of 5th International Conference on Language Resources and Evaluation. ELRA, 1670–1673, (2006).
6. Ying He, Ph.D. és Mehmet Kayaalp, M.D., Ph.D.: A Comparison of 13 Tokenizers on MEDLINE , Technical report.
`http://lhncbc.nlm.nih.gov/lhc/docs/reports/2006/tr2006003.pdf`, (2006).

Magyar Webkorpusz II.

Halácsy Péter[1], Kornai András[1], Németh Péter[2], Varga Dániel[1]

[1] Budapesti Műszaki Egyetem, Média Oktató és Kutató Központ,
{hp, kornai, daniel}@mokk.bme.hu
[2] Kitchen Budapest, spacecadet@kitchenbudapest.hu

1. Bevezetés

Az alábbiakban bemutatjuk a *Magyar Webkorpusz* készülő új, második kiadását. A első Magyar Webkorpusz [2] elkészülte óta feldolgozási láncunk minden fontos elemét továbbfejlesztettük, és a láncba integráltuk morfológiai egyértelműsítő rendszerünket. A második kiadás létjogosultságát technológiai fejlesztéseinken kívül természetesen az is indokolja, hogy 2003 ősze, az első Webkorpusz anyagának begyűjtése óta a magyar Web mérete lényeges mértékben nőtt, és a zsánerek arányai is jelentősen megváltoztak. Rövid absztraktunkban csak a legfontosabb új fejlesztéseket emeljük ki.

2. Crawling, Tokenizálás, Nyelvazonosítás

A nyers weboldalak begyűjtéséhez ezúttal a WIRE crawlert [1] alkalmaztuk. A WIRE nagy teljesítményű, erősen párhuzamosítható működésű webcrawler. Elemi duplikátum-szűréssel rendelkezik, amit saját, nyelvfeldolgozásra hangolt duplikátumszűrőnkkel egészítettünk ki.

A Szószablya projekthez kifejlesztett tokenizáló és mondatra szegmentáló rendszerünket véges állapotú technológiára alapozva újraimplementáltuk, lényegesen felgyorsítva ezzel.

A .hu domainből kinyert dokumentumok nem elhanyagolható százaléka nem magyar nyelvű. A dokumentumok nyelvének azonosításához szintén egy véges állapotú technológián alapuló rendszert építettünk. Ennek teljesítménye is lényegesen nagyobb, mint az első korpusz építésekor alkalmazott megoldásé.

3. Morfológiai egyértelműsítés

Morfológiai egyértelműsítőnk [3] a morphdb.hu magyar morfológiai erőforrást [4] felhasználva dolgozik. Rendszerünk sebessége megfelelő ahhoz, hogy a teljes Webkorpuszt feldolgozásnak vethessük alá. Hangsúlyozzuk, hogy az egyértelműsítés nem csupán szófaji azonosítást jelent: minden tokenhez részletes morfológiai információt rendelünk, képzést és produktív szóösszetételek felbontását

is beleértve. A rendszer elfogadható pontossággal oldja meg a sem lexikonja, sem tanítókorpusza által nem ismert szavak elemzésének (guessing) nehéz feladatát is.

4. A morfológiai erőforrás

morphdb.hu morfológiai erőforrásunk fejlesztésében az eredeti Webkorpusz felbecsülhetetlen segítséget nyújtott. Elsősorban természetesen oly módon, hogy lehetőséget adott gyakori, de a morfológiai erőforrás által mégsem ismert szóalakok megtalálására. Megjegyezzük, hogy a Webkorpusz leggyakoribb szóalakjai között is előfordul idegen (főként angol) nyelvű szó, elgépelés, helyesírási hiba és ékezethiány. Az angol nyelvű szavak automatikus kiszűrésére lehetőséget adott az angol nyelvű morphdb.en erőforrásunk használata. Az ismeretlen, de nem angol nyelvű szavak leggyakoribbjainak átvizsgálását, és az indokolt esetekben erőforrásba felvételét elvégeztük. Ezen a ponton elmondható, hogy az első Webkorpusz 60,000 leggyakoribb szóalakját a morphdb.hu helyesen elemzi. A morphdb.hu erőforrás fedésének növelése visszahat az új Webkorpusz minőségére, amennyiben javítja a morfológiai egyértelműsítés és a nyelvfelismerés pontosságát.

5. Webes keresőfelület korpusznyelvészeknek

A korpuszból épített, szógyakoriság-listát nyers és morfológiailag egyértelműsített változatában publikussá tesszük. De ezen túl építettünk a gyakoriságlistához egy olyan web-alapú kereső-felületet, amely minden a korpusznyelvészek által hagyományosan alkalmazott keresési feltételt és rendezési elvet támogat.

6. Tervbe vett fejlesztések

Technológiáink jelentős része nyelvfüggetlen. Ezért természetesen adódik az a célkitűzés, hogy a webkorpusz-építést a környező országok miénkhez hasonló méretű webes jelenléttel bíró nyelveire is elvégezzük. Ez a munka az absztrakt írásának pillanatában cseh nyelvre zajlik, de reményeink szerint több más nyelvre is elvégezzük majd.

Terveink között szerepel továbbá a szógyakoriság-lista interaktív keresőfelületének kiterjesztése olyan módon, hogy szövegkörnyezet-információt is indexeljen és kereshetővé tegyen.

Hivatkozások

1. Carlos Castillo and Ricardo Baeza-Yates. Wire: an open-source web information retrieval environment. In *Workshop on Open Source Web Information Retrieval (OSWIR)*, 2005.

2. Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In *Proceedings of Language Resources and Evaluation Conference (LREC04)*. European Language Resources Association, 2004.
3. Péter Halácsy, András Kornai, Csaba Oravecz, Viktor Trón, and Dániel Varga. Using a morphological analyzer in high precision POS tagging of Hungarian. In *Proceedings of LREC 2006*, pages 2245–2248, 2006.
4. Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of LREC 2006*, pages 1670–1673, 2006.

Magyar mondatok SVM alapú szintaxiselemzése

Iván Szilárd[1], Ormándi Róbert[2], Kocsor András[2]

[1] Szegedi Tudományegyetem, Informatikai tanszékcsoport
szilivan@inf.u-szeged.hu
[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
{ormandi,kocsor}@inf.u-szeged.hu

Kivonat: A nyelvtechnológiai alkalmazások egyik fontos elemzése a szintaxis-elemzés. Bemutatásra kerül egy gépi tanuláson alapuló szintaxis elemző, mely az SVM alapú megközelítést alkalmazza. A használt algoritmusok elméleti és implementációs részleteinek bemutatásán túl, átfogó teszteléssel igazoljuk a módszer alkalmazhatóságát. A módszer további érdekessége, hogy a strukturált kimenetű tanulás paradigmáját követi.

1 Bevezetés

A szintaxis elemzés a természetes nyelvi feldolgozás elemzéseinek azon csoportja, melyeknek célja a mondatok nyelvtani struktúrájának felderítése. Ez a struktúra leggyakrabban egy hierarchikus szerkezet, ahol a legnagyobb egység a mondat, legkisebb egységei pedig az alapszimbólumok (például a szavak szófajai, vagy azok POS kódjai). Az ilyen típusú szerkezetet általában egy fával, az úgynevezett szintaxis fával szokás reprezentálni, melynek gyökerében az egész mondatnak megfelelő csúcs áll, levelei az alapszimbólumokat, a belső csúcsok pedig az egyéb nyelvtani egységeket reprezentálják.

Az ilyen típusú elemzésnek rendkívül fontos szerepe van a természetes nyelvi feldolgozás számos területén, hiszen egy mondat szintaxisfájának helyes meghatározása alapvető fontosságú a magasabb szintű szövegfeldolgozáshoz (például szemantikai elemzés, vagy gépi fordítás).

Szintaxis elemzésre alapvetően kétféle megközelítés létezik, az egyik a szakértők által megadott összefüggéseken alapuló, a másik a gépi tanulást előtérbe helyező eljárások. Manapság a figyelem az utóbbi módszerekre összpontosul, angol nyelvre igen hatékony algoritmusok kerültek kidolgozásra, de a magyar nyelv sajátosságai (nyelvi variabilitás) miatt ezek változatlan formában történő alkalmazása jelentős hatékonyságveszteséggel jár [1], [2].

2 SVM alapú megközelítés

Jelen publikációban bemutatunk egy gépi tanuláson alapuló szintaxis-elemző eljárást, amely a manapság intenzíven kutatott SVM alapú megközelítést követi [5]. A kidol-

gozott eljárás a szintaxisfákat mint, adott valószínűségi környezet független nyelvtan feletti derivációs fákat értelmezi. Ezeket a fákat jól jellemzi, hogy a deriváció során az egyes szabályok hányszor lettek alkalmazva. A megközelítés lényege, hogy a szabályok alkalmazásának eloszlásának becslését végzi [3], [4]. Az itt előálló feladat, átalakítható olyan formára, mely a manapság intenzíven kutatott margó maximalizáló eljárások segítségével oldható meg. A módszer algoritmikus részleteinek bemutatásán túl, egy releváns gyakorlati feladaton keresztül igazoljuk a bevezetett eljárás létjogosultságát.

3 Korpusz

Az algoritmus teszteléséhez szükséges mondatok, és hozzájuk tartozó szintaxis-fák a Szeged Korpusz adattárából származnak. A korpusz több témakörben (iskolai, jogi, számítógépes, szépirodalmi, üzleti) tartalmaz szövegeket, amelyeken nyelvészek a különféle elemzéseket, mint morfológiai elemzés, szófaji egyértelműsítés, szintaxis elemzés.
A tanításhoz és teszteléshez használt mondatok az üzleti témakörben található mondatokból kerültek ki. A tanulás-tesztelés során használt szintaxisfák teljesen általános struktúrával rendelkeznek.

4 Eredményeink

Annak mérése, hogy az elemzés eredményeképpen előálló fa, mennyire jó, azaz mennyire hasonlít az elvárt szintaxisfához nem könnyű feladat. A szakirodalomban, erre három elterjedt mértéket szoktak használni:

- *Pontosság (precision):* a helyesen felismert szócsoportok számának és az összes felismert szócsoport számának hányadosa.
- *Fedés (recall):* a helyesen felismert szócsoportok számának és a mintában ténylegesen szereplő szócsoportok számának hányadosa.
- *F1-mérték:* 2*Pontosság*Fedés / (Pontosság+Fedés), azaz a Pontosság és a Fedés harmonikus átlaga.

Látható, hogy mindhárom mérték 0 és 1 között mozog, és minél nagyobb értéket vesz fel, annál ,,jobbnak" mondható az eredmény. Az 1. táblázat összefoglalja a mérési eredményeink. A táblázatban szereplő értékek F1-mértékben értendőek.

1. Táblázat: Eredmények a használt mondatok hosszának függvényében.

Mondat hossza	Tanító adatbázis	Teszt adatbázis
15-20	87.7%	89.2%
21-25	87.2%	89.0%
26-30	86.1%	86.5%
31-35	85.1%	83.3%
36-40	84.3%	79.6%
41-45	84.0%	78.7%
3-45	86.8%	87.0%

Bibliográfia

1. Brill, E.: Transformation-Based Learning. PhD thesis, University of Pennsylvania, (1993)
2. Hócza, A.: Teljes mondatszintaxis tanulása és felismerése. MSZNY (2004) 127-135
3. Joachims, T.: A support vector method for multivariate performance measures. Twenty-Second International Conference on Machine Learning (2005)
4. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large Margin Methods for Structured and Interdependent Output Variables. Journal of Machine Learning Research (JMLR) (2005) 1453-1484
5. Vapnik, V.: Statistical learning theory. Wiley and Sons Inc (1998)

A lexikalista szintaxis rangja(i)

Szilágyi Éva, Kleiber Judit, Alberti Gábor[1]

PTE BTK Nyelvtudományi Tanszék;
Pécs 7624 Ifjúság útja 6.
realis@mail.btk.pte.hu

Bevezetés

Az MSZNY konferencián többször szerepelt GeLexi [5] és Lile [6] projektek örökén szerveződött a ReALIS projekt a PTE BTK Nyelvtudományi Tanszékén 2006-ban. Alapvető célunk továbbra is az intelligens nyelvészeti alapokon történő fordítás, totálisan lexikalista keretben [1]. A lexikont egy relációs adatbázis képviseli, amit MS SQL2005-ben valósítunk meg.

Egy relációs adatbázis lényege a „relációk" definíciójában áll, amely a reprezentálandó tények leírása mellett, magát az adatbázist alkotja. A leírás az egyed-kapcsolat modellen alapul. Minden egyes entitás egy rendezett *n*-es: elemei valamilyen relációt alkotnak — ez egy rekord. A reláció maga a tábla, ahol minden egyes sor (rekord) egy rendezett *n*-es, és minden oszlop egy attribútum. A relációk egyedei közötti kapcsolatok állhatnak fel [7]. A számtalan adatbázis-kezelő rendszerből és SQL-megvalósítás közül választásunk az Microsoft SQL Server 2005 rendszerre esett: ezáltal egy komplett relációsadatbázis-kezelő keret rendelkezésünkre áll, és számos olyan szolgáltatással is rendelkezik, amelyet céljaink eléréséhez ki tudunk használni: elsődlegesen a kliens-szerver architektúrát, amelynek köszönhetően az adatbázis egyidejűleg több felhasználó és programmodul által is elérhető, vagyis a felhasználók a rá épített webes felületen keresztül, ingyenesen használhatják a rendszert.

A morfofonológiai komponenst alapvetően a Lile adatbázisából vettük át. Jelenleg a szintaxis „szabályainak" a lexikonba építésén dolgozunk, erről szól a poszterünk is. A fő komponenst az ez után fejlesztendő szemantika jelenti, amit a ReALIS [4] dinamikus szemantikai rendszer implementációjaként képzelünk el.

Predikátumok és argumentumok, régensek és vonzatok

A totális lexikalizmus eszméjével összhangban a fent említett „szabályok" nem szabályok, hanem az egyes lexikai elemekre jellemző tulajdonságok, amelyek végül unifikálód(hat)nak. Egy lexikai egységről tárolnunk kell azt, hogy milyen argumentumokat követel, illetve azokat milyen vonzatként kéri megvalósítani (akár több lehetséges változatban). Ezek mellett fontos tényező még a szórend kialakulása, amely-

[1] Ezúton szeretnénk köszönetet mondani az OTKA-nak (OTKA K60595) a támogatásért.

ről az általunk használt nyelvtan-modellben rangparaméterek adnak számot. A rangparaméterek egy „speciális” megjelenése számot tud adni az olyan esetekről is, amikor a mondatban a semleges szórendet megvariáló (az írott magyar mondatokban másképpen nem észrevehető) fókusz vagy egyéb operátor van jelen.

A lexikai egységek által követelhető (követelt) argumentumstruktúrákat[2] egy egyednek tekintjük, ennek elemeit egy olyan készlet adja, mely megmondja, milyen típusú argumentumok fordulhatnak elő. Ezek jellemzését egy olyan számparaméter adja, amely az ágentív-pacientív skálán helyezi el az argumentumot.[3] Az argumentumok így egy argumentumstruktúra, valamint az argumentumtípus kapcsolatából képződnek.

A másik oldalt a hasonlóan felépített vonzatkeretek jelenik. Egy egyedet képeznek a lexikai egységek által követelhető (követelt) „vonzatkeretek”: ezt kétirányúan tároljuk, a régens és a vonzat felől is[4], elemei pedig az ezek típusait definiáló készletből származnak. Ez a készlet nemcsak esetragokat tartalmaz, hanem az olyan alakzatokat, mint a névutók vagy az infinitívusz, illetve az olyan állandósult kifejezésekként használt szerkezetek, amelyekre egy-egy esetragos alak cserélhető. Például: *Péter elárul pár dolgot Mariról / Marival kapcsolatban.* Az egyes vonzatokat tehát a vonzatkeret és egy vonzattípus kapcsolatai jelentik.

Előfordulnak olyan esetek is, amikor a lexikai egység nem ad számot az argumentum esetéről. Például a *lakik* ige esetén két argumentum szerepel: egy *lakó* és egy *hely,* ahol ő lakik. A vonzatok között szintén kettőt találunk egy NOM esetet, amit a lakó számára tartogat a lexikai egység (amit a vonzat és az argumentum összekötése ad meg; „bármire illeszthető” típusokkal is dolgozva), és egy másik vonzatot, amelynek az esete nem specifikált – ennek ellenére hiánya agrammatikus mondatot eredményez. Erre az esetre az argumentumtípusok és a vonzattípusok is összeköthetőek: a példában az argumentum HELY típusához így szelektálhatók az olyan vonzattípusok, mint a -bAn *(egy szép házban),* az -On *(Szentesen)* vagy a mellett *(az iskola mellett).*

A szintaxisnak számot kell még adnia a szabad határozókról. Szabad határozónak azt tekintjük, ahol a rag kompozicionálisan szerepel, csak úgy mint [8], ám vonzatnak tekintjük az olyan kompozicionális elemeket, amelyeknek a jelenlétét valamelyik másik elem előírja. Ekkor a rag (vagy a lexikai egység, pl. *ott*) árulja el magáról, hogy ő az általa keresett főnévvel együtt szabad határozóként jelenhet meg. A szemantikai struktúra fog a lexikai egységeknél olyan tulajdonságokat megadni, ami alapján például a különböző típusú időhatározók közül a grammatikusakat ki lehet választani. Modellünkben ezek az elemek a mondat eventuális argumentumát keresik.

---

[2] Több argumentumkeretnek tekintjük azt a névszókra jellemző jelenséget, hogy lehetnek ők maguk vonzatok, illetve lehetnek névszói állítmányként predikátumok is. Ezek mellé kell még számolnunk a rövid birtokos formát is, ami keres egy birtokos személyjelet. (Pl. Péter *Budapesten lakik. Annak a fiúnak a neve* Péter. Péter *kalapja)*

[3] Az ilyen szempontból semleges argumentumok semleges számparamétert kapnak, mivel a centrális keretben – amely az alanyok és tárgyak viszonyainak alakulását írja le – nincsenek benne.

[4] A szabad bővítmények csak egyirányúak

Rangparaméterek

A lexikai egységek a követelményeik mellett azt is megmondják, mennyire szükséges azt kielégíteni, és a szomszédosságra vonatkozó követelmény esetén azt is, hogy az azt kielégítő elemet milyen irányban, illetve hogy a szón belül vagy egy másik szóban keresik. Ennek okát az adja, hogy ami szemantikai kapcsolatban áll a régenssel, az annak szomszédja szeretne lenni. Mivel azonban a mondatok lineárisak, a régensnek mindössze két szomszédja lehet: ám sokkal inkább csak egy, hiszen a régensek a nyelvre jellemzően egy irányból veszik fel a vonzataikat[5]. A rangparaméterek a nyelv leírásából derülnek ki, tapasztalat útján; tárolásuk az egyes vonzatoknál történik. Minden nyelv régensei ugyanazt a követelményt támasztják (ezért is egy egyed – egy tábla), ám különféle erősségű rangokkal (amelyek így a nyelvek közötti különbségekről is számot adnak, lásd pl. *scrambling* jelenségek).

A rangparaméterek kétféle kielégíthetőségi követelménnyel léphetnek fel. A *domináns* rangparaméterek esetében elegendő, ha egy ellentmondó, erősebb paraméter ki van elégítve. A szomszédossági viszonyokat megadó *recesszív* paraméterek esetén a követelményt mindenképpen ki kell elégíteni – akár részlegesen is [3].

Az igekötők esetében például kétféle rangparaméter is él, mindkettő recesszív: Az egyik esetben az *el* igekötő az *indul* elé kell, hogy kerüljön, ezt egy erős (-2) rangparaméter mondja meg. Fonológiai szempontból a hangsúly is az igekötőre esik, míg az igéről lekerül, így egy fonológiai szót alkotnak; ez az aspektualizáló argumentum[6] [2]. A második esetben az igekötő kerülhet az ige mögé is egy gyengébb (+3) rangparaméter szerint, ekkor mindkét elem külön fonológiai szóként realizálódik.

Semleges mondatban az erősebb (-2) rang elégül ki, míg egy fókusz megjelenése esetén, ami erős domináns típusú rangparamétert jelent[7], az igekötő igét megelőző helyére vonatkozó követelményt nem lehet, és nem is kell kielégíteni. A gyengébb (+3) rangparaméter követelménye viszont továbbra is jelen van, így azt ki lehet és kell elégíteni.

Nemcsak az igekötő lehet aspektualizáló: hasonló a helyzet a már említett *lakik* igénél: a helyre vonatkozó vonzatát – pl. *Budapesten* – az erős (-2) ranggal szeretné maga előtt tudni, vagy egy gyengébb (+3) ranggal maga mögött. Ez speciális eset, ahol az aspektualizáló argumentum a lokatívusz. Bizonyos esetekben igekötős ige esetén is ez az aspektualizáló vonzat: a *megszáll* igével semleges esetben a *Péter Budapesten szállt meg* a grammatikus mondat, míg semmiképpen nem semleges a *Péter megszállt Budapesten*. Az aspektualizáló argumentumot mindig két ranggal vesszük fel.

A fókusz, ami ugyan (a magyarban) hangalakot nem ölt, lexikai egységként jelenik

[5] Az irányultságot a szám 0-hoz viszonyított iránya fejezi ki, vagyis negatív számok jelentik a megelőző elemként való keresést, és pozitív számok a következő elemre vonatkozó keresést. Az erősséget a paramétert jelentő szám abszolút értéke adja.

[6] Mindig van egy olyan vonzat, amit aspektualizálásra használ az ige (Rendszerint az igekötőt, alkalmanként saját magát, pl. *Péter csalódik Mariban*).

[7] Amit először ki kell elégíteni, hogy ez után kiírtódhassanak a dominánssal ellentétes követelmények; és ez után következnek a recesszívek.

meg az adatbázisban. A mondatban két elemet keres: a fókuszált elemet és az igét. Az igének olyan rangparamétert oszt, hogy annak mögötte kell lennie, a legerősebb (+1) domináns rangban. Hasonlóan kezelhető a telikus szituációkra vonatkozó progresszív alak működése is.

Megfigyelhetünk olyan elemeket is, amelyek jelenlétében mindig található fókusz is, ilyen például a *csak,* illetve a mondatrésztagadást jelző *nem.* Ezen elemek erős domináns ranggal vonzzák a fókuszt maguk mögé, ami a szórendi változásról gondoskodik.

Összegzés

Rendszerünk előnye és újszerűsége abban rejlik, hogy a lexikonban ugyanazon keretek között kezelhető a szintaxis körébe (is) tartozó számos tényező (predikátum-argumentum, illetve régens-vonzat viszonyok, szabad határozók, szórend).

A rangparaméterek elegánsan számot adnak az egy nyelven belüli szórendi variációkról (a szón belül a morfémák sorrendjéről), valamint a nyelvek közötti különbségekről is. A domináns rangparaméterek használatával a szórendet megvariáló, sokszor láthatatlan elemek (fókusz, progresszív) is kezelhetőek.

Projektünk jelenleg a fent vázolt rendszer megvalósításán dolgozik. A következő lépés a fő komponensként elképzelt szemantika implementálása lesz, mert hisszük, hogy intelligens alkalmazások csak valós nyelvészeti alapokon készíthetők, ahhoz pedig alaposan kidolgozott szemantikai rendszerre van szükség.

Hivatkozások

[1] Alberti G.: GASG: The Grammar of Total Lexicalism; IN: Working Papers in the Theory of Grammar 6/1. Theoretical Linguistics Programme, Budapest University and Research Institute for Linguistics, Hungarian Academy of Sciences, 1999

[2] Alberti G.: Az aspektus szintaxisa a magyarban; IN: Újabb tanulmányok a strukturális magyar nyelvtan és a nyelvtörténet köréből, Osiris,Budapest, 2001; pp.145-164.

[3]Alberti G., Balogh K., Kleiber J., Viszket A.: A totális lexikalizmus elve és a GASG nyelvtan-modell; IN: Maleczki Márta (szerk.): A mai magyar nyelv leírásának újabb módszerei V. SZTE , Szeged, 2002; pp. 193-218

[4] Alberti G., Dóla M., Farkas J., Kántor Gy., Kleiber J., Ohnmacht M. (2007): ReALIS: a "reális" interpretációs rendszer; IN: Alberti G. és Fóris Á. (szerk.) A mai magyar nyelvtudomány formális műhelyei, Nemzeti Tankönyvkiadó, Budapest, pp. 139-156.

[5]Alberti G., Kleiber J., Viszket A.: GeLexi projekt: Gépi fordítás totálisan lexikalista alapokon; : Dr. Alexin Z., Csendes D. (szerk.): MSZNY2004 - II. Magyar Számítógépes Nyelvészeti Konferencia; Informatikai Tanszékcsoport, Szeged, 2004; pp.73-80

[6] Bódis Z., Kleiber J., Szilágyi É., Viszket A.: Lile projekt: Adatbázis mint "dinamikus korpusz"; IN: Dr. Alexin Z., Csendes D. (szerk.): MSZNY2004 - II. Magyar Számítógépes Nyelvészeti Konferencia; SZTE Informatikai Tanszékcsoport, Szeged; pp.11-18

[7] Halassy B.: Az adatbázis-tervezés alapjai és titkai. Avagy az út az adattól az adatbázison át az információig; IDG, Budapest 1994

[8] Gábor K. - Héja E.: Predikátumok és szabad határozók, IN: Kálmán László (szerk.) KB 120 A titkos kötet, MTA Nyelvtudományi Intézet Tinta Könyvkiadó, Bp., 2006; pp.134-152.

Névmutató