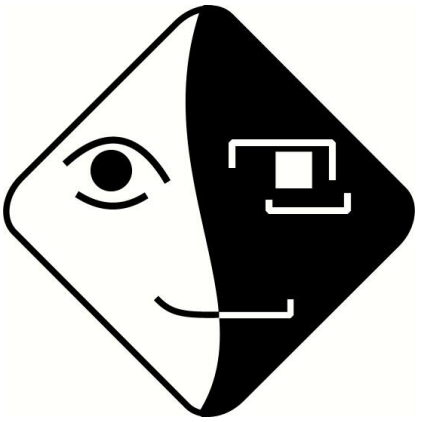
IV. Magyar Számítógépes Nyelvészeti Konferencia



MSZNY 2006

Szeged, 2006. december 7-8.
http://www.inf.u-szeged.hu/mszny2006

ISBN-10: 963-482-800-0
ISBN-13: 978-963-482-800-6

Szerkesztette: Alexin Zoltán és Csendes Dóra {alexin, 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, 2006. november

Előszó

2006. december 7-8-án negyedik alkalommal kerül megrendezésre a Magyar Számítógépes Nyelvészeti Konferencia. Nagy örömömre szolgál, hogy a rendezvény évről évre nagyszámú érdeklődőt vonz az ország különböző tájairól. A konferencia fő célja a nyelvtechnológia (elsősorban a szöveg- és a beszédfeldolgozás) területén elvégzett vagy folyamatban lévő kutatások és fejlesztések legaktuálisabb eredményeinek bemutatása. Lehetőség nyílik kapcsolódó hallgatói projektek, ill. a nyelvtechnológia ipari alkalmazásainak ismertetésére is.

Az idei felhívásra beérkezett tudományos értekezések közül a programbizottság 34-et fogadott el előadás megtartására, és további 12-t poszterprezentáció, ill. 4-et laptopos bemutató megtartására. Külön öröm, hogy idén két plenáris előadó is elfogadta meghívásunkat, így a szakmai program Kornai András és Pléh Csaba egy-egy előadásával is gazdagabb lesz. Ugyancsak az idei évben tervezzük először 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 Gordos Géza, László János, Prószéky Gábor és Váradi Tamás programbizottsági tagoknak. Szeretném továbbá megköszönni a rendezőbizottság: Csendes Dóra és Alexin Zoltán munkáját.

Csirik János, a rendezőbizottság elnöke
Szeged, 2006. november

Tartalomjegyzék

I. Korpusznyelvészet

II. Morfológia

III. Ontológia

IV. Szemantika

V. Gépi fordítás

VI. Beszédtechnológia, kommunikáció

VII. Pszichológiai szempontú szövegfeldolgozás

VIII. Poszterbemutatók

IX. Laptopos bemutatók

I. Korpusznyelvészet

Magyar internetes gazdasági tematikájú tartalmak keresése

Tikk Domonkos[1], Biró György[2], Szidarovszky P. Ferenc[1,3], Kardkovács Zsolt T.[1], Héder Mihály[1], Lemák Gábor[4]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
H-1117 Budapest, Magyar Tudósok krt. 2.
{tikk, szidarovszky, kardkovacs}@tmit.bme.hu, merlin@sch.bme.hu
[2] TextMiner Bt.
H-1029 Budapest, Gyulai P. u. 37.
george.biro@gmail.com
[3] Szidarovszky Kft.
H-1392 Budapest, Pf. 283.
ferenc.szidarovszky@szidarovszky.com
[4] GKI Gazdaságkutató Zrt.
H-1092 Ráday u. 42–44.
lemakg@gki.hu

Kivonat: A projektünk célja egy olyan keresőszolgáltatás kiépítése, amely az Interneten magyar nyelven elérhető gazdasági tematikájú tartalmak lehető legteljesebb körét egy helyen kereshetővé, és – amennyiben a tartalomszolgáltató, ill. jogtulajdonos részéről ennek nincs akadálya – elérhetővé is teszi a felhasználók számára. Jelen munkánk ismerteti a szolgáltatás funkcióit, felépítését, és megvalósítását. A komponensek közül részletesen foglalkozunk a nyelvtechnológiái módszereket alkalmazó szövegfeldolgozó és webszüretelő modulokkal.

1 Bevezetés

A projektünk célja egy tematikus, szemantikus elveket és újfajta vizualizációt alkalmazó keresőszolgáltatás létrehozása, amelyen keresztül a felhasználók az Interneten magyar nyelven elérhető gazdasági tematikájú tartalmak lehető legteljesebb körében kereshetnek, és amely – amennyiben a tartalomszolgáltató, ill. jogtulajdonos részéről ennek nincs akadálya – tartalmakhoz való hozzáférést is biztosítja a felhasználók számára. A keresőszolgáltatók egyre bővülő piacán egy olyan szegmenst célzunk meg, amely jól körülhatárolható, de korántsem elhanyagolható jelentőségű felhasználói kör. A tágabb értelemben vett gazdasági tartalmak érdekelhetik mind az átlagfelhasználót (pl. kisbefektetői kör, laikus érdeklődők), mind a vállalatvezetői, tanácsadói, döntéshozói pozícióban lévőket, mind pedig a szakmai felhasználókat – oktatók, kutatók, hallgatók.

Projektünk eredményétől azt várjuk, hogy a keresési kérésnek megfelelő dokumentumok pontosabban kielégítik a felhasználói igényeket, mint a jelenlegi alkalmazások, illetve az újfaja vizualizáció lerövidíti az információfeldolgozás idejét. Emellett a projektmegvalósítás során egy olyan know-how is létre jött, amelynek segítségével újabb tematikákkal, tudományterületekkel bővíthetjük a szemantikus keresést biztosító keresőszolgáltatásunkat, ezáltal hozzájárulva a magyar nyelvű világháló jelentés-orientálttá válásához.

Cikkünk felépítése a következő. Először a 2. szakaszban ismertetjük a kitűzött funkcionalitásokat, keresési formákat, majd a 3. szakaszban bemutatjuk a rendszer felépítését és az egyes komponenseket. A 4. szakaszban a működés során jellemző folyamatok vizsgálata következik hangsúlyozottan kiemelve a nyelvtechnológiai eljárásokat alkalmazó komponensek vonatkozó részleteit, míg az 5. szakasz a hasonló, elsősorban hazai vonatkozású kezdeményezéseket veszi számba. Végül a 6. szakaszban röviden összegzést adunk.

2 Támogatott keresési formák

A keresőszolgáltatás keresési funkcióinak meghatározása során célunk az volt, hogy a szokásos keresési lehetőségeknél fejlettebb szolgáltatásokat nyújtsunk, és támogassuk a felhasználóknak a keresési eredmények böngészése, azokon való navigálás során felmerülő továbbkeresési igényeit.

A keresőmotorok hatékonyságának növelésére egyik lehetőség, ha a felhasználó meghatározhatja a keresett tartalom tematikáját. Ez segíti a keresőmotort a keresési igény pontos meghatározásában, pl. több értelmű keresőkifejezések esetén, és leszűkíti a találat lista méretét csökkentve ezáltal az irreleváns találatok számát. A felhasználók tematikus navigációjának, illetve keresésük orientálásának támogatása a tartalmak tematikus rendszerezésével érhető el, ennek előfeltétele, hogy rendelkezésre álljon a megcélzott tematikát lefedő megfelelő részletezettségű hierarchikus kategóriarendszer (*taxonómia*). Az általános, nagy nemzetközi keresőszolgáltatások is rendelkeznek hasonló keresési lehetőséggel, (ld. pl. Google Directory, Yahoo Directory, Zeal kereső[1] stb.), de egy ilyen opció jelentősége egy tematikájában és nyelvében eleve korlátozott tartalomgyűjteményt összefogni kívánó szolgáltatás esetén sokkal számottevőbb. Az általános keresők esetén ugyanis sokkal nehezebb egy mindenre kiterjedő, kellően részletes taxonómia megalkotása és karbantartása, valamint szintén nagy kihívást jelent a taxonómia megfelelő minőségű tartalommal való feltöltése. A projektünk által megcélzott szűkebb tematika és rögzített nyelv viszont a tartalmak sokféleségéből és a témák dinamikus változásából eredő taxonómia-karbantartási feladatok bonyolultságát jelentősen csökkenti.

Esetenként a felhasználó számára rendelkezésre áll a kereséséhez egy teljes kiindulási – akár saját készítésű – dokumentum, amelyhez hasonlókat kíván megtekinteni. Az általános keresők nem támogatják a bizonyos szószámot meghaladó[2], hosszabb keresőkifejezéseket, ezért – amennyiben a dokumentum nincs indexelve – nem képesek a feladatot végrehajtani.

[1] http://www.google.com/dirhp, http://search.yahoo.com/dir, http://www.zeal.com

[2] A Google legfeljebb 32 szavas keresőkifejezéseket értelmez.

Egy keresés találati listája és annak elemei gyakran szintén fontos kiindulási pontot jelenthetnek további keresések kezdeményezésére, a keresett tartalom pontosítására, a keresés finomítására. A felhasználó számára azonban korántsem egyértelmű – még a találatok rövid átfutása után sem –, hogy milyen módon tudja leghatékonyabban bővíteni, vagy módosítani a keresését. Ezt a tevékenységet a találati listában lévő dokumentumok kulcsszavainak felkínálásával eredményesen lehet támogatni.

A felsorolt megfontolások alapján a következő keresési funkciókat határoztuk meg:

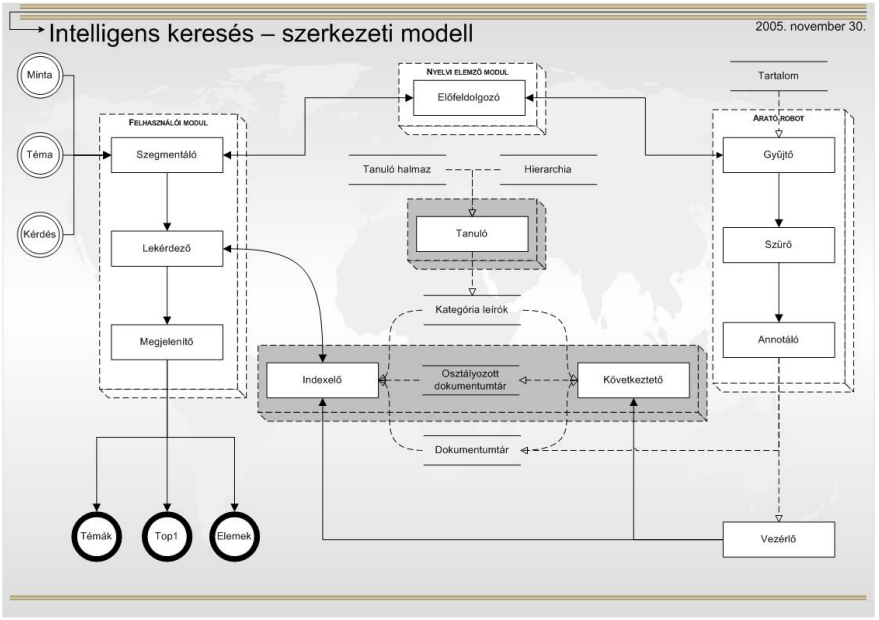
– szabadszavas kérdés-vezérelt keresést,
– mintadokumentum alapú keresést,
– tematikus böngészési lehetőséget rögzített témastruktúrában,
– keresés finomítási lehetőség a találatok kulcsszavai alapján.

3 A rendszer felépítése

A keresőszolgáltatás mögött egy négy fő komponensből álló, összetett rendszer hivatott kiszolgálni a felhasználók igényeit (ld. 1. ábra). A rendszer fő komponensei az alábbiak:

– aratórobot,
– nyelvi feldolgozó modul,
– indexelő és osztályozó motor,
– felhasználói felület.

Az alábbiakban bemutatjuk az egyes komponensek feladatát és vázlatos működését.



1. ábra. A rendszer szerkezeti modellje

Aratórobot

Az aratórobot feladata a kiválasztott, gazdasági témájú híreket (is) közlő magyar oldalakról való tartalomgyűjtés és a rendszer által használt XML formátumra való átalakítás. A cikk írásakor a tesztrendszer mintegy 50 forrásból gyűjti a dokumentumokat, melyek közt túlnyomó részben országos gazdasági tartalomszolgáltatók anyagai szerepelnek, de vannak köztük általános portálok témába vágó cikkei, és regionális tartalomszolgáltatók is.

Nyelvi feldolgozó modul

A nyelvi feldolgozó modul feladata, hogy a különböző forrásokból beérkezett dokumentumokat egységes formátumra hozza. A feldolgozás folyamatát úgy határoztuk meg, hogy akár különböző alapú szövegreprezentációs technikák (pl. szó-alapú vagy karakter n-gram alapú) is megvalósíthatóak legyenek, a feldolgozás során, pedig tetszőleges nyelvtechnológiai eszközök alkalmazásának eredményeit is integrálni lehessen.

Indexelő és osztályozó motor

Az indexelő motor feladata a keresés végrehajtásához szükséges indexállomány létrehozása, karbantartása és a keresések kiszolgálása. Az osztályozó motor a taxonómiával bővített kulcsszó alapú kereséshez szükséges kategóriainformációk nyilvántartását, ill. meghatározását támogatja. Az osztályozó modul felügyelt gépi tanulást végez, azaz tanítódokumentumok alapján megtanulja a taxonómia kategóriáinak jellemző szavait, ill. kifejezéseit. Ennek megvalósítására a HITEC osztályozóalgoritmusát integráltuk a rendszerbe [1, 2]. Az osztályozó motor segítségével tehát egyrészt lehetőség van az egyes kategóriák jellemző szavainak, ill. kifejezéseinek meghatározására, azaz ún. kategóriaprofilok kiépítésére, másrészt a rendszerbe kategóriacímke nélkül bekerülő dokumentumok kategóriáinak automatikus becslésére. Ezek a motorok korábbi fejlesztések eredményeiként álltak elő, ezt az 1. ábrán szürke alapszínnel jelöltük.

Felhasználói felület

A felhasználói modul biztosítja a keresési felületet a felhasználók felé, a lekérdezések továbbítását a keresőmotorhoz, illetve a keresőmotortól kapott eredmények megjelenítését és feldolgozását.

Taxonómia kiépítése és feltöltése

A keresőszolgáltatás hatékonysága és a keresés minőségének biztosítása szempontjából kiemelt fontosságú, hogy a taxonómia jól reprezentálja a tématerületet, kellően részletes finomítását adja a legfontosabb fő témaköröknek, ugyanakkor a kapcsolódó, ill. peremterületeket érintő tematikát is lefedje. Ezért a gazdasági témájú szövegeket tematikus osztályozásának alapját jelentő taxonómiát egy szakkönyvtár, a Budapesti Corvinus Egyetem Központi Könyvtárának tárgyszórendszere alapján alakítottuk a könyvtár szakértőinek segítségével. A kiindulási tárgyszórendszer különböző kapcsolattípusokat tartott nyilván (szűkebb/bővebb terminus, használt/nem használt terminus, kapcsolódó fogalom), valamint köröket is tartalmazott, ezért közvetlenül nem volt alkalmas egy hierarchikus, csak generatív/partitív relációkat tartalmazó taxonómia megalkotására. Az átdolgozást a könyvtár munkatársai végezték el az informatikus szakértők útmutatásai alapján. Ennek során elsődleges szempont a megfelelő struktúra kialakítása volt, úgy hogy a tárgyszórendszer élő elemei a taxo-

nómiába is átkerüljenek. A megfelelő struktúra kialakítására néhány új, korábbi tárgyszavakat egy csomópontba összekapcsoló kategóriát is létrehoztunk. Az így kialakított fastruktúrájú taxonómia 16 legfelső szintű kategóriából kiindulva összesen 2397 kategóriát tartalmaz, legnagyobb mélységében hat szintes. Amennyiben a rendszer tesztelése során kapott felhasználói visszajelzések szükségessé teszik, a taxonómia még módosulhat.

A taxonómia osztályozáshoz való felhasználására feltétlenül szükség van tanulódokumentumokra, azaz olyan mintákra, amelyek jól reprezentálják az egyes kategóriákat. Ehhez természetesen a BCE Könyvtár tárgyszórendszerét használtuk fel, mivel így számos, a könyvtár eredeti tárgyszórendszere segítségével annotált elektronikus dokumentum azonnal a rendelkezésünkre állt. A tanulókörnyezet teljessé tételét, azaz hogy minden lényeges csomóponthoz megfelelő számú tanulóadat legyen, úgy valósítottuk meg, hogy lehetőség szerint beszereztük könyvtári katalógusrendszerben elektronikus formában nem szereplő dokumentumok elektronikus verzióját, illetve újonnan annotált elektronikus dokumentumokkal bővítettük a rendszert.

4 A rendszer működése

4.1 Dokumentumok feldolgozása és tárolása

A rendszerbe való bekerülés módjától függően két dokumentumtípust különböztetünk meg: tanuló- és szüretelt dokumentumokat. Az egyedüli különbség, hogy a tanulódokumentumok rendelkeznek kategóriainformációval, míg a szüreteltek nem. (A felhasználó által megadott keresőkifejezéseket a feldolgozás szempontjából a szüretelt dokumentumokkal analóg módon kezeljük, csak ezeket nem tároljuk el.) A rendszerbe kerülő dokumentumok eredeti formátuma több féle lehet, HTML, PDF, DOC, RTF, illetve szöveges (TXT), ezeket a rendszer által használt XML alapú reprezentációra kell megfelelő konverziós eljárások alkalmazásával átalakítani. A dokumentumok tárolására egy olyan egyszerű, de a dokumentumfeldolgozás bármely lépését tárolni képes struktúradefiníciót (DTD) hoztunk létre[3], amely elsősorban a szövegbányászati feladatok elvégzésére optimalizált, de egyszersmind könnyen átalakítható bármely más szabványos XML formátumra (pl. NewsML, TEI[4], stb.).

A DTD létrehozásakor fontos szempontot volt, hogy

1. a meghatározott információk kódolására képes legyen az XML struktúra,
2. az XML formátumú szöveg tárigényének minimalizálása. Ennek kiemelt jelentősége van egyrészt a dokumentumgyűjtemény mérete, másrészt a feldolgozó algoritmusok működési sebessége és memóriaigénye miatt.

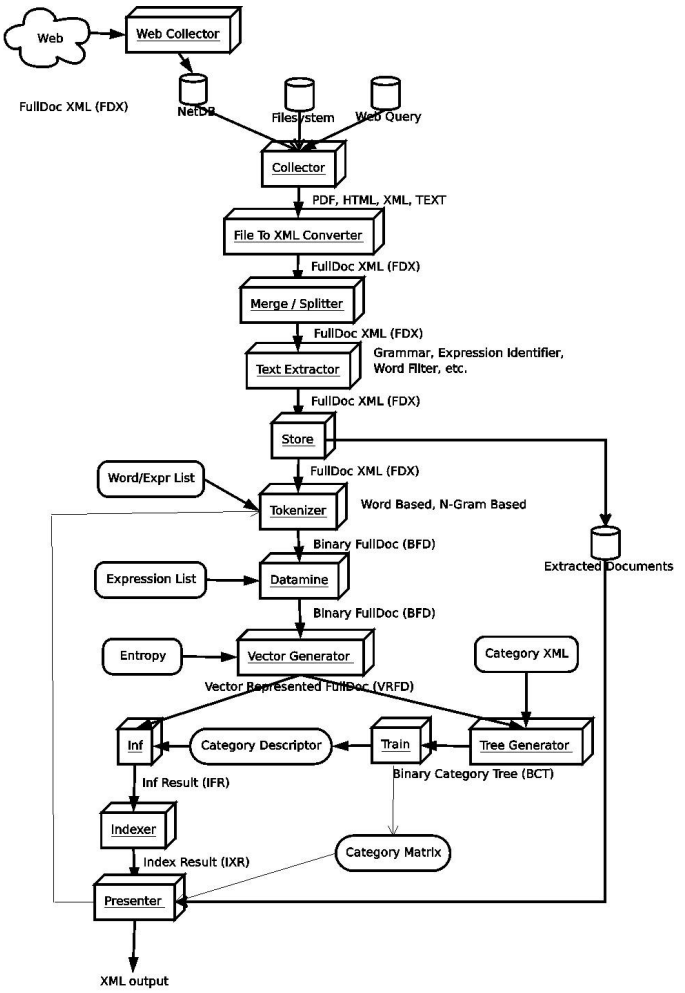
Az első követelményt a viszonylag rugalmas szerkezeti felépítés segítségével értük el, míg a másodikat a gyakran ismétlődő szerkezeti elemek rövid elnevezésével, és csak a feltétlenül szükséges attribútumok kötelező megadásával valósítottuk meg. A DTD tervezése során kiemelt figyelmet fordítottunk arra, hogy az összes szöveges mező elemek tartalmaként kerüljön feldolgozásra, és kizárólag az egyéb metainformációk kerüljenek az egyes elemek attribútumaiba. Ezzel egyrészt a konvertáló programok számára kívántuk feldolgozási konvenciót előírni, s ezáltal a kon-

[3] fulldoc.dtd – Jelenlegi elérhetősége: http://dodona.tmit.bme.hu/~tikk/fulldoc.dtd

[4] http://www.newsml.org/pages/spec_main.php, http://www.tei-c.org/P4X/ST.html

verziót megkönnyíteni, másrészt ily módon oldottuk meg a szövegjellegű információk és a metaadatok keveredésének kiküszöbölését. A szöveges adatok képezik az indexelő modul primer bemenetét, míg az attribútumokban előforduló metaadatok az intelligens szövegbányászati és nyelvtechnológiai módszerek bemeneteként szolgálnak.

A szövegfeldolgozási folyamat lépéseit a 2. ábra ismerteti. Ez alapján látszik, hogy minden dokumentum esetén ugyanaz a feldolgozás folyamata. Az eredeti dokumentumok XML konverziója után (FDX formátum) a Merger/Splitter modul a dokumentumok összefűzésén kívül a karakterkódolás egységesítését is elvégzi. A Text Extractor komponens nyelvtechnológiai eljárások alkalmazását végzi. Ennek során az alábbiakat valósítja meg:



2. ábra. A feldolgozási folyamat lépései

- **Szótövező**: A rendszer két alternatív lehetőséget kínál a feladat elvégzésére. Egyrészt tartalmazza a szabályalapú, ún. óvatos szótövező algoritmust [4], másrészt pedig integrálja a Szószablya projekt[5] HunMorph csomagjának HunStem szótövesítő eljárását [5]. A DTD lehetőséget ad különböző elemzési alternatívák kezelésére (ld. g[rammar] elem), így több szótő megadására is, amit a g elem stem attribútuma tárol.
- **Szófaj meghatározása**: Ezt az eljárást szintén a HunMorph csomag morfológiai elemzőjének segítségével valósítjuk meg. Az indexállományok általános megvalósítása szótő szerinti nyilvántartást végez, ezért homonim szótövek esetén a különböző előfordulások egybevonódnak. Ennek elkerülésére rendszer az indexállományban [szótő, szófaj] párokat tárolunk. A szófaj információt a g elem pos attribútuma tárolja.
- **Szószűrő**: Ahhoz, hogy hatékonyan lehessen a keresések finomításához kulcsszavakat javasolni, elengedhetetlen az általános értelmű ún. funkció- v. stopszavak szűrése. Ezt egyrészt egy előre megadott szótár, illetve minták alapján valósítjuk meg. A szűrőn fennakadó szavak esetén a g elem sw attribútumát igazra állítjuk, ugyanis az indexelésnél ezekre a szavakra is szükség van, így nem törölhetőek.
- **Szótári névelemek felismerése:** Szótári névelemeknek nevezzük azokat a rögzített formájú kifejezéseket (többnyire tulajdonnevek), amelyek alapalakja a szövegben változatlan formában fordul elő. A névelemeknek lehetnek különböző előfordulási alakjaik (pl. Petőfi Sándor és Petőfi vagy Orléans-i szűz és Jeanne d'Arc), amelyek közül egyet kanonikus alaknak jelölünk ki, a többit pedig a kanonikus alak szinonimájaként kezeljük. Ezeket, ahogy elnevezésük is utal rá, egy szótárban tároljuk. A felismerésükhöz a HunMorph csomag morfológiai elemzőjét is felhasználó eljárást alkalmazzuk [6]. A névelemként felismert kifejezéseket e[xpression] címkével látjuk el.
- **Mondathatár-detektáló**: A modul a szövegek mondatszintű szegmentálását végzi, eredményét a keresési eredmények rövid legjellemzőbb részletének meghatározásánál alkalmazzuk. Működése szabályrendszer alapú: mondathatároló jelek előfordulásánál a szabályok alapján eldöntjük, hogy az adott jel ténylegesen mondathatárt jelöl, vagy sem. A szabályokhoz előjeles súlyértékeket rendelünk. Amennyiben egy adott mondathatár-környezetre több szabály illeszkedik, akkor a szabályok súlyának aggregálásával határozzuk meg a végső értéket. A feldolgozás során a szabálytár mellett rövidítéstárat is alkalmazunk. A detektált mondatokat s[entence] címkék közé tesszük.

A felsoroltakon kívül a DTD lehetőséget nyújt tetszőleges nyelvtechnológiai alkalmazás, pl. teljes morfológiai elemzés kimenetének felhasználására is. A rendszer továbbfejlesztése során ezen eljárásokat a keresés támogatásában nyújtott hatékonyságuk alapján integráljuk a rendszerbe.

A dokumentumoknak három különböző mértékben feldolgozott verzióját tároljuk a rendszerben. Az eredeti formátumú dokumentum mellett, a nyers XML dokumentumot is tároljuk, majd a Store modul a feldolgozott XML-t tárolja el, és amennyiben rendelkezésre áll, kategóriainformációt rendel hozzá. A dokumentumok különböző verzióinak elérési útvonalát a document elem megfelelő attribútumaiban tároljuk.

[5] http://mokk.bme.hu/projektek/szoszablya

A dokumentumok feldolgozása ezek után már numerikus alakban történik, az átalakítást a Tokenizer modul végzi el. A Datamine modul már ebben a formátumban keres gyakran ismétlődő tokensorozatokat, amelyeket egyedi indexszel lát el. Végül a belső reprezentációs alakot a Vector Generator komponens látja el, amely az irodalomban leggyakrabban használt vektortér alapú szózsák (*bag of words*) modellel[6] a dokumentumokból két vektort állít elő, egyet az indexeléshez, egyet pedig az osztályozáshoz. Az indexeléshez létrehozott vektor TF-IDF súlyozást alkalmaz és tartalmazza a stopszavakat is; míg az osztályozáshoz készített vektor entrópia-alapú súlyozást használ, és a stopszavakat nem tartalmazza [7].

Ezen a ponton válik el a különböző dokumentumok feldolgozási folyamata, hiszen a tanulóadatokat az osztályozó motor tanítására használjuk (Train), a többi dokumentum kategóriáját pedig a tanulóadatok alapján felépített osztályozási modellel (Inf) határozzuk meg. Ezután kerülnek a dokumentumok indexelésre, majd a felhasználói felület felé a Presenter modul jeleníti meg a szükséges kulcsszó- és kategóriainformációkat, immár nem numerikus (tokenizált), hanem szöveges formában.

4.2 Az aratórobot működése

Az arató modul feladata a célirányos, előre specifikált, illetve keret-megállapodással rendelkező partnerek portáloldalainak (összefoglalóan: gyűjtési tartomány) folyamatos követése, archiválása és címkézése. Az aratásnak jellegzetesen két fő funkciót kell kielégíteni:

1. Az oldalak folyamatos gyűjtését és háttértárra mentését (röviden szüretelés).
2. Az elmentett oldalak előfeldolgozását és szerkezeti címkézését.

A gyűjtés során adott, jól meghatározott forrásokat kell üzemszerűen meglátogatni. A gyűjtést egy ún. dæmon kell végezze – nevezzük a továbbiakban Aratónak –, amelyet indítani, azonnali gyűjtésre ösztönözni, valamint leállítani és késleltetni lehet.

Az Arató elindítja a letöltési folyamatot, amelynek bemenete a specifikált, gyakran meglátogatandó URL – tipikusan egy portál főoldala, vagy egy RSS-csatorna[7]. Az URL-t meglátogatva, az oldal tartalmát letöltve, rövid analizálás és címkézés után az oldal új dokumentumait le kell töltenie, és dokumentumarchívumba el kell helyeznie, majd a letöltött dokumentumból el kell távolítania a nem releváns részeket.

A megoldást nem kívántuk egyetlen tématerületre limitálni, így a specifikációban a legáltalánosabb megoldást választottuk. Ugyanakkor látni kell, hogy a releváns szövegek tartalmi elválasztása a dokumentum többi részétől nem biztosítható egyetlen univerzális algoritmus segítségével. A tartalmilag összefüggő, a tényleges információt hordozó szöveg kiválasztását legfeljebb nagyon mélyreható szemantikai elemzéssel lehetne a 100%-os pontosság közelébe juttatni. (Pontosság alatt értve azt, hogy a gazdasági hír, mint tartalom, teljes anyaga szerepel a kiválasztott szövegrészletben, és kizárólag az szerepel benne.) E tekintetben a statisztikai megoldások sem lehetnek segítségünkre, hiszen az összefüggő szövegek kiválasztására ma még nem ismert statisztikai alapú módszer.

---

[6] A modell a dokumentumokat a bennük szereplő szavak, illetve kifejezések (általánosan: tokenek) halmazának tekintik, ez tehát figyelmen kívül hagyja a tokenek pozícióját és sorrendjét a szövegben.

[7] Real Simple Syndication – http://blogs.law.harvard.edu/tech/rss

Felfigyeltünk ugyanakkor arra, hogy a cikk megjelenített és tényleges címének azonossága, illetve a cím ismerete esetén, valamint öt kulcsjellemző (dátum, szerző, cím, kivonat, szövegtörzs) egymáshoz viszonyított elhelyezkedésének ismeretében a ténylegesen releváns szöveg, mintegy 90%-os pontossággal azonosítható.

A gyakorlatban a különböző portáloldalak szerkezeti címkézését oldalanként egy-egy kis segédprogram – *plugin* – elkészítésével oldottuk meg. Ezek a segédprogramok az adott oldal szerkezeti jellegzetességeit figyelembe véve a HTML forrást fulldoc sémára illeszkedő XML-lé alakítják.

Mivel a segédprogramok elkészítésénél csak az egyes portálok aktuális jellegzetességeit ismertük, fel kellett készülnünk arra, hogy egy esetleges portál-motor váltáskor, vagy az oldal szerkezetének nagyobb léptékű változásakor a régi szerkezet figyelembe vételével készített segédprogram rossz kimenetet kezd produkálni. Ezért minden, az aratórobot által előállított XML-t megvizsgálunk, szintaktikailag ellenőrzünk. Egy oldal változása miatt elavult segédprogrammal előállított XML-ből többnyire hiányoznak a legfontosabb, kötelezően kitöltendő mezők, (pl. cím, szövegtörzs), ezért a fájl a szintaktikai ellenőrzésen fennakad. Az egyes portálok tartalmából előállított XML dokumentumok ilyen módon vizsgált tulajdonságairól statisztikát vezetünk, ami lehetővé teszi, hogy a figyelt portálok szerkezeti változtatásairól értesüljünk.

Ugyanakkor elképzelhető, hogy egy portál szerkezete úgy változik, hogy a szintaktikai ellenőrzés helyes marad, de a tartalom nem, pl. nem gazdasági témájú cikket gyűjtünk be. Az aratórobot az ilyen problémákat egyelőre nem tudja automatikusan kiküszöbölni.

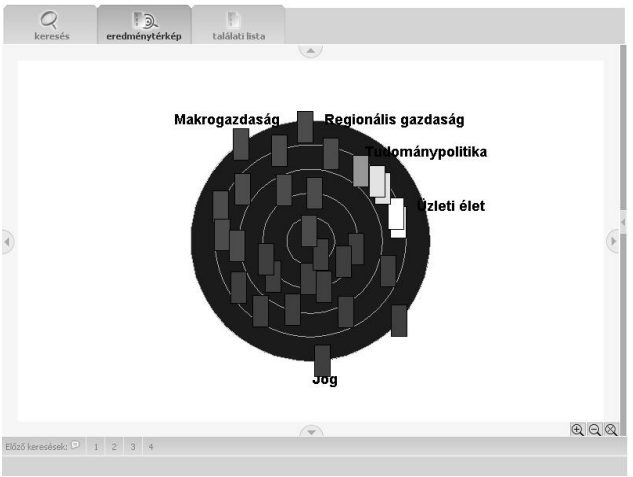
4.3 A felhasználói felület

A felhasználói felület a tervezésénél a funkcionális szempontok mellett a vizuális megjelenítésre is nagy hangsúlyt fektettünk. Terjedelmi okokból csak a találatok egyik megjelenítése formáját, az eredménytérképet tudjuk megmutatni. A szolgáltatás prototípusának beüzemelése és a szolgáltatás publikussá tétele 2006 év végére fog megvalósulni.

5 Hasonló kezdeményezések

Az elmúlt 5 évben az otthonukból internetezők tábora 6%-ról közel 30%-ra emelkedett és a havi legalább 1 órát internetezők száma elérte a 1,5 milliót. A felhasználók számának növekedése a tartalomipar expanzióját vonta maga után, amelyhez napjainkban az üzleti oldalú tartalom-előállítás mellett – a technológiai fejlődés és támogatás eredményeként – a felhasználó oldali tartalom-létrehozás is hozzájárul. Ennek a bővülésnek köszönhetően megnőtt az igény a keresőszolgáltatások iránt, amelyek kiépítésére a szolgáltatásból adódó üzleti lehetőség, a hazai online hirdetési piac dinamikus növekedése is serkentően hatott. Az üzleti oldalt megelőzve a tudományos szféra hamarabb felismerte az internetes keresésben rejlő tudományos kihívásokat és 2000-től – elsősorban az NKFP IKTA program finanszírozásában – tudományos kutatóműhelyek kezdtek különböző keresőalgoritmusok és -intelligenciák fejlesztésében. Az elmúlt 6 évben mind a tudományos, mind az üzleti szférában voltak törekvések olyan újfajta keresőeljárások kidolgozására, amelyekkel a szövegek gépi meg-

értésén keresztül próbálták a keresést pontosabbá és hatékonyabbá tenni, ám e kezdeményezések gyakorlati hasznosulása és hasznosítása jellemzően nem történt meg, így indokoltnak tartottuk olyan kutatás-fejlesztési projekt megvalósítását, amely nemcsak új eredményeket képes felmutatni a keresőintelligencia-kutatás területén, de az üzleti hasznosítást is képes biztosítani. A következőkben összefoglaljuk a hasonló hazai kezdeményezések.



3. ábra. A keresőszolgáltatás eredménytérkép oldala

Az *Információ és Tudás Tárház* (IKTA3-181/2000) projekt fő célkitűzése új intelligens tudás tárházak analízise, tervezése és megvalósítása volt, melyek lehetővé teszik a fejlett tudás- és üzletiinformáció-menedzsmentet [8, 9]. A projekt tudásalapú információ visszakeresési rendszert fejlesztett ki a pénzügyi szféra cégei számára, amely különböző forrásokból (Internet, belső adatbázisok, külső adattárházak, stb.) merít információt az alkalmazási környezet igényeinek megfelelően, majd ezt strukturált formában tárja a felhasználó elé.

A szavak hálójában (NKFP 0019/2002) projekt célja egy komplex internetes kereső/kérdező eszköz létrehozása volt, amely mind az Interneten elérhető online adatbázisok szöveges tartalmaiban – azaz a *mélyhálón,* a hagyományos keresőkkel nem indexelhető tartalmak összességén –, mind képek közti keresések terén új technológiákat tartalmaz [10, 11]. A képi keresés támogatására egy vizuális tezaurusz került kifejlesztésre, ami a képi tartalmak jellemzésére és indexelésére használható szöveges leírások, mint tartalmi kategóriák rendszere, strukturált szótára. A mélyhálótartalmakban történő keresésnél a rendszer támogatja magyar nyelven megadott teljes mondatok keresőkifejezésként történő használatát.

A *Szemantikailag szervezett lexikai hálózat és internetes tartalomkeresés* (IKTA5-123/02) projekt célja egy szemantikai szerveződésű, lexikai hálózat kifejlesztésére épülő, internetes tartalomkeresésre alkalmazható, újfajta technológia létrehozása volt. A projekt a célját a lexikai hálózat alapegységének tekintett, már kifejlesztett, ún. jelentésközpontok egy lehetséges kapcsolódásainak kutatásával és a kapcsolóelemek kiépítésével kívánta elérni (a jelentésközpont az azonos jelentés köré szerveződő, értelmezett természetes nyelvi kijelölők – szavak, szószerkezetek, mondatértékű kifejezések – egy struktúrában összefogott és kezelt egysége). A jelentés-

központok egymással való összekötésével, ún. linkek létrehozásával létrejövő, szemantikailag szervezett, kommunikatív lexikai hálót a projekt olyan kutatási szempontok alapján fejlesztett ki, hogy az képes legyen nyelvtechnológiai alkalmazásokban (természetes nyelvi szövegfeldolgozó-rendszerek, értelmezett információ-keresés elektronikus szövegekben és strukturált szövegtestekben, tartalomfigyelés, gépi fordítás, kontextus- és stílusérzékeny helyesírás-ellenőrző) értelmezetten és hatékonyan működni.

Az Országos Baleseti és Sürgősségi Intézet vezetésével valósult meg a *Tudásalapú magyar nyelvű szemantikus kereső rendszer kifejlesztése és alkalmazása a sürgősségi betegellátásban* (IKTA 00148/2002) projekt. A projekt tartalmazza az adatok statisztikai kontrollja mellett az adatok fogalmilag rokon csoportjainak (klaszterezés), valamint a logikai kapcsolatok extenzionális összefüggéseinek megállapítását. Ehhez a kidolgozott technológia a „tudásfeltárás" gépi tanulási és neuron-hálós eljárásokon alapuló módszereit ajánlja a „klasszikus" adatbányászati módszerekkel (drilling-up, drilling-down, stb.) együtt. A rendszer éles kipróbálása az Országos Traumatológia Intézet Információs rendszerébe ágyazva történt meg, ahol a szükséges orvosi ontológia rendelkezésre áll, és a megfelelő dokumentumok gyors megtalálása életbevágóan fontos. A kidolgozott tudásalapú keresési technológia teljesen általános, és széleskörűen használható könyvtárak, archívumok, orvosi, jogi és vállalati adat- és ismeretbázisok keresőmotorjaként, és mindazoknál a kereskedelmi alkalmazásoknál, amelyekben a célorientált keresés fontos szerepet játszik.

A *WebKat*[8] az első magyar fejlesztésű tématérképen alapuló modell, amelyet a Neumann-ház egy pályázat keretében hozott létre 2002-ben. Ez a tezaurusz a kereséseket a meglévő tárgyszórendszer relációinak tématérkép alapú vizuális megjelenítésével támogatja. A szolgáltatás nem az internetes tartalmakban, hanem a saját adatbázisában keres.

A *PolyMeta*[9] egy általános célú metakereső, amely lehetőséget nyújt tetszőleges számú Interneten keresztül elérhető kereső (adatbázis, forrás) egyidejű keresésére. Az eredményekből közös találati lista készül, amelyben az elemek fontossági sorrendbe rendezettek. Megjelenítésre kerül egy „tartalomjegyzék" is, ami segítséget ad a felhasználónak a témához tartozó résztémák, kapcsolódó fogalmak azonosításában, az azokhoz tartozó találatok megjelenítésében.

A Vipkereső[10] nevű legújabb kezdeményezés jelenleg még teljes funkcionalitásában nem érhető el, de az előzetes információk alapján a szabadszavas webkereső mellett képkereső, hírkereső és blogkereső funkciókat is nyújt majd. Várhatóan az Index szolgáltatásaként jelenik majd meg.

6 Összefoglalás

Cikkünkben ismertetjük egy tematikus keresőmotor felépítését és megvalósításának fő lépéseit. A megvalósított keresőszolgáltatás prototípusa az Interneten fellelhető magyar nyelvű gazdasági tartalmakat gyűjti, indexeli és teszi egy helyen kereshetővé.

[8] http://www.webkat.hu

[9] http://www.polymeta.hu/polymeta/meta.html

[10] http://www.vipkereso.hu/ – az alkalmazás a cikk írásakor még nem érhető el.

A találati dokumentumokat, amennyiben a tartalomszolgáltató ezt engedélyezi, a keresőfelületen keresztül is elérhetővé válik.

Köszönetnyilvánítás

A cikk a Gazdasági Versenyképesség Operatív Program GVOP-3.1.1.-2004-05-0130/3.0 jelű projektjének támogatásával készült.

Bibliográfia

[1] D. Tikk, Gy. Biró, and J. D. Yang. Experiments with a hierarchical text categorization method on WIPO patent collections. In N. O. Attok-Okine and B. M. Ayyub, editors, *Applied Research in Uncertainty Modelling and Analysis*, number 20 in Int. Series in Intelligent Technologies, pages 283–302. Springer, 2005.

[2] D. Tikk, J. D. Yang, and S. L. Bang. Hierarchical text categorization using fuzzy relational thesaurus. *Kybernetika*, **39**(5):583–600, 2003.

[3] Z. Alexin and D. Csendes, editors, *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY'05)*, Szeged, Hungary, December 8–9, 2005. SZTE, Informatikai Tsz.csoport.

[4] D. Tikk, A. Törcsvári, Gy. Biró, and Z. Bánsághi. Szótövező eljárások hatása magyar szövegek automatikus kategorizálásánál. In [3], pages 430–434.

[5] V. Trón, P. Halácsy, P Rebrus, A. Rung, E. Simon, E, and P. Vajda: morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In [3], pages 169–179.

[6] D. Tikk, F. P. Szidarovszky, Zs. T. Kardkovács, and G. Magyar. Ismert névelemek felismerése és morfológiai annotálása szabad szövegben. In [3], pages 190–199.

[7] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Information Processing and Management, **24**(5):513–523, 1998.

[8] Cs. Dezsényi, P. Varga, T. Mészáros, Gy. Stratusz, T. Dobrowiecki: Ontológia-alapú Tudástárház Rendszerek. http://nws.iif.hu/ncd2003/docs/ahu/AHU-118.htm

[9] Cs. Dezsényi et. al: Tudásalapú információk kinyerése: az IKF projekt. In: Tudományos és Műszaki Tájékoztatás, 2004/5. http://www.neumann-haz.hu/tei/publikaciok/2004/biro_ref_ikf_hu.html

[10] D. Tikk, Zs. T. Kardkovács, et al: Natural language question processing for Hungarian deep web searcher. In *Proc. of the IEEE Int. Conf. on Computational Cybernetics (ICCC'04)*, pages 303–308, Vienna, Austria, Aug 30–Sept 1.

[11] D. Tikk, Zs. T. Kardkovács, and G. Magyar. Searching the deep web: the WOW project. In *Proc. of the 15th Int. Conf. on Intelligent Systems Development (ISD'06)*, Budapest, Hungary, Aug 31–Sept 2, 2006.

Igei vonzatkeretek az MNSZ tagmondataiban

Sass Bálint

MTA Nyelvtudományi Intézet, Nyelvtechnológiai Osztály,
PPKE, Információtechnológiai Kar, MMT Doktori Iskola
joker@nytud.hu

Kivonat: A vonzatkeretekkel foglalkozó mulkálatok célja a magyar vonzatkeretrendszer korpuszalapú feltérképezése. A Magyar Nemzeti Szövegtárból származó egyszerű mondatok vizsgálata után most alkalmassá vált a rendszer tetszőleges morfoszintaktikailag elemzett szöveg feldolgozására az új tagmondatra bontó modul segítségével. A tagmondatok egy igét és annak bővítményeit tartalmazzák, így alkalmas bemenetét képezik a vonzatkeret-illesztő modulnak. A magyar vonzatok és bővítmények tanulmányozására létrejött egy internetes alkalmazás, mely a fenti emailcímen igényelt jelszóval elérhető a http://corpus.nytud.hu/mazsola címen.

1 Bevezetés

A már több mint egy éve folyó munkálatok távlati célja a Magyar Nemzeti Szövegtárban lévő igei vonzatkeretek feltérképezése. A III. MSZNY konferencián a vonzatkeret-táblázatban összegyűjtött, kézzel kódolt, ismert keretek azonosításáról számoltam be [6].

A készülő magyar-angol gépi fordítási projekt számára szükség volt további vonzatkeretekre, a vonzatkeret-táblázat kiegészítésére. Kidolgoztam egy módszert az eddig ismeretlen keretek azonosítására, melynek segítségével korpuszból nyerhetünk ki új vonzatkereteteket [5]. *A vonzatok: nem kompozicionális bővítmények* elvére alapozva az igék és a mellettük álló NP-k alkotta keretjelöltek idiómaságát mértem a *DF (distributed frequency)* idiómasági mértékkel [7]. Eszerint a mérték szerint az a keret az idiomatikusabb, melynek bővítményei az adott formában kevés (szélső esetben egyetlen) igével fordulnak elő (pl. *fittyet hány*). A magasabb idiómasági érték azt jelenti, hogy a keretjelölt valódi vonzatkeret. Az így nyert új keretek kézi ellenőrzést követően a magyar-angol gépi fordító rendszer lexikai erőforrásába kerültek.

Maga a vonzatkeret-felismerés a következőkben fejlődött: az igei vonzatkeretek feldolgozása a cél, ezért csak olyan tagmondatokkal foglalkozom, melyekben található ige vagy főnévi igenév. Természetesen utóbbi is elegendő, mert a főnévi igenév mindig hordozza a neki megfelelő ige vonzatkeretét.

Továbbfejlesztettem az igekötő- és igetőazonosítást, az elváló igekötőket az igetőhöz ragasztottam, a gyakori képzőket (pl. *-hAt*) levágtam, mivel nem befolyásolják a vonzatkeretet.

2 Tagmondatra bontás

2.1 Motiváció

Az az optimális, ha a vonzatkeret-illesztő modul számára mindig pontosan egy egy vonzatkeretet tartalmazó szövegdarabot tudunk bemenetként adni. A korábbiakban az MNSZ egy preparált részkorpuszával dolgoztam: egyszerű heurisztikával kiválasztottam a korpuszból a rövid (maximum tíz szavas), írásjel nélküli mondatokat. Ezek a mondatok nagy valószínűséggel egy keretet tartalmaznak, azonban nyilvánvalóan nem reprezentálják a valós nyelvhasználatot, így a belőlük nyert gyakorisági adatok nem kellően megalapozottak [6]. Egyetlen előny a könnyű feldolgozhatóság.

Jelen cikk lényegi előrelépése, hogy elkészült egy előfeldolgozó, tagmondatra bontó modul, így a rendszer alkalmassá vált tetszőleges szöveg feldolgozására. Most a szöveg tagmondatai képviselik a nagy valószínűséggel egy vonzatkeretet tartalmazó egységet. A tagmondat kifejezést tehát ebben az értelemben használom: a mondat egy vonzatkeretet tartalmazó része, így lényeges követelmény lesz annak garantálása, hogy a tagmondat egy igét tartalmazzon. Sok helyen találkozhatunk a mondatok bizonyos szempontból könnyebben elemezhető, kisebb részekre darabolásával [4], itt is erről van szó.

2.2 Korábbi megoldások

Nincs tudomásom magyar nyelvre alkalmazható, részletesen ismertetett, reprodukálható tagmondatra bontó módszerről. Létezik egy az INTEX/NooJ nyelvfeldolgozó rendszerben implementált eljárás, melynek vázlatos leírása a [8] cikkben olvasható. Ezenkívül két kéziratos leíráshoz jutottam hozzá, melyeket itt most röviden ismertetek: a fenti eljárás részleit tartalmazó kézirat [3], illetve egy másik megközelítést tartalmazó kézirat [9].

A [3]-ban leírt *tagmondathatár-azonosító* rendszer az alábbi tizenegy darab szabályból áll. A szabályokat reguláris kifejezésre emlékeztető szintaxissal írom le, adott szabály illeszkedése esetén a '@' jel helyére kerül be egy tagmondathatár.

1. [,|-] @ [kötőszó|határozószó]? [vonatkozó névmás]
2. [-] @ [kötőszó|határozószó]? [vonatkozó névmás] [bármi]+ [-] [,]? @
3. [,|-] @ [bármi|NP|AdjP]? [*pedig*|*akár*|*azonban*|*viszont*|*ellenben*|*mihelyt*|*tehát*|*ugyanis*]
4. [,|-] @ [NP]? [*meg*]
5. [,|-] @ [határozószó]? [*nehogy*|*mintha*]
6. [,] @ [kötőszó, kivéve: *de*|*illetve*|*illetőleg*|*mintegy*]
7. [,|-] @ [múlt idejű, egyes szám harmadik személyű ige]
8. @ [kötőszó] [kötőszó]
9. @ [kötőszó] [kérdőszó]
10. [,] @ [határozószó|NP]? [kérdőszó]
11. [,] @ [*az* szótőként] [határozói igenév] [,] [meglévő tagmondathatár] [*hogy*]

Az eljáráshoz tartozik még egy a szabályalkalmazások után futó program, mely lehetséges tagmondathatárként megjelöli az összes kötőszót, mely két olyan finit ige között helyezkedik, melyek között még nincs tagmondathatár. Látjuk, hogy a szabályrendszerben részletesen benne foglaltatik, hogy az egyes kötőszók hányadik pozícióban szoktak állni a tagmondathatárhoz képest, és milyen típusú elemek előzhetik meg őket. A cikk közvetlen tagmondatkezdő kötőszókat tartalmazó listáját már sikerrel alkalmaztuk korábban is [1].

A [9]-ben leírt, de nem implementált eljárás igazi célja, hogy megállapítsa a szöveg *kötőszavairól*, hogy szerkezeteket koordinálnak vagy esetleg tagmondatokat kötnek össze, így mintegy melléktermékként kapjuk meg a tagmondatokat.

Az eljárás a következő előfeldolgozó lépésekkel kezdődik:

– az első finit ige előtti és az utolsó finit ige utáni kötőszavakat jelöljük meg koordináló kötőszavaknak;
– ha két finit ige között egyetlen kötőszót találunk, akkor az jelöljük meg tagmondathatárnak.

Ez után hajtandó végre a következő utasítássorozat a szöveg összes kötőszaván:

d) ha a kötőszó két oldalán lévő két frázis különböző típusú, a kötőszó tagmondathatárt képvisel;
di) különben AdjP és AdvP esetén koordinációról van szó;
dii) NP esetén, ha egyezik az eset és az NP-k monotonicitása, akkor koordináció, ha nem egyezik, akkor tagmondathatár;
diii) VP esetén, ha egyezik a szám és a személy, akkor koordináció, ha nem egyezik, akkor tagmondathatár;
div) egyébként tagmondathatár.

A kéziratban [2,9] megfogalmazott fontos elv szerint: a finit ige vonzatai az igét tartalmazó tagmondaton belül vannak. A kézirat említést tesz a magyar névszói állítmányról, mégis a továbbiakban már mindig csak finit igékre hivatkozik, így lehetőséget ad arra, hogy esetleg hibásan bevegyük a finit ige vonzatai közé a szomszédos névszói prédikátum vonzatait is. Mivel ez az eljárás a kötőszavak alapján hoz döntést, nem ad számot azokról a tagmondathatárokról, ahol nincs jelen kötőszó. A kiértékeléskor használt korpuszból származó példamondatok minkét problémára rávilágítanak:

Meglehet: mindkettőre igen a válasz.

Nagy tanulság, hogy győzelmem Kwasniewski ellen azt jelentette volna, Lengyelországban még forradalom zajlik.

2.3 Az inkrementális nyelvtanfejlesztés

Módszeremben lényegében a fent ismertetett első eljárásra építek [3]. A végső szabályrendszer kialakítása során az *inkrementális nyelvtanfejlesztésnek* nevezett módszert alkalmaztam, az alábbiak szerint.

Minden elemi fejlesztési lépés (újabb szabály hozzávétele vagy strukurális változtatás) után lefuttattam a programot egy 600 mondatos tesztkorpuszon, és megvizsgáltam, hogy mit változtatott a kimeneten ez az egy lépés. Manuálisan értékeltem az eredményt. Ha nagyrészt megfelelőnek ítéltem – néhányszor szinte hibátlan működésre is volt példa – akkor megtartottam, ha sok esetben rossz eredményt hozott, akkor elvetettem az adott fejlesztési lépést.

Ehhez az informális értékeléshez nagyon fontos, hogy mindig megfelelő számú olyan eset álljon rendelkezésre, amikor az adott jelenség, amit a fejlesztési lépés kezel, előfordul. Ahelyett, hogy külön az egyes szabályokhoz preparált kisméretű korpuszokat készítenék, megpdóbálok mindig akkora korpuszt venni, amelyben legalább kb. 20 példányban előfordul a jelenség. Esetemben, ha nem volt elég példa, vagy nem volt egyértelmű az értékelés, akkor egy 4500 mondatos tesztkorpuszon végeztem el újból a vizsgálatot. Ha a nagyobb korpuszban sem volt példa a jelenségre, akkor túl ritka lévén, elhagytam az adott lépést. Így folyamatos korpuszkontroll mellett és viszonylag alacsony időráfordítással tudtam a fejlesztést végezni.

Ahhoz, hogy ezt a fejlesztési módszert alkalmazni tudjam, szükséges, hogy a szabályok függetlenek legyenek egymástól. Erre törekedtem is, ezért nincs is a rendszerben olyan, hogy egy már meglévő tagmondathatárra hivatkozom, ahogy az a fenti 11. szabályban történik.

2.4 A tagmondatra bontó eljárás

A fejlesztés során a következő megfigyeléseket tettem:

1. kettőspontnál illetve pontosvesszőnél minden mástól függetlenül meg lehet jelölni a tagmondathatárt;
2. a *meg* kötőszóra épülő (4.) szabályt elhagytam, mert a szó szófaji egyértelműsítés itt bizonytalan, az esetek nagy részében ténylegesen igekötő a kötőszónak jelölt *meg*;
3. a két egymást követő kötőszót váró 8. és kötőszó + kérdőszót váró 9. szabály az esetek nagy részében rossz helyen jelölt ki tagmondathatárt, így ezt a két szabályt elhagytam;
4. kérdőszó elé ékelődő határozószóra nem volt példa, a szabályt elhagytam (10. szabály első fele);
5. kérdőszó elé ékelődő frázis esetén ismét a szófaji egyértelműsítés korlátjába ütköztem, itt legtöbbször a kérdőszónak jelölt *ki* igekötő kapcsán működött a szabály (10. szabály második fele).

Ez a tagmondatra bontó modul előfeldolgozóként a részleges szintaktikai elemzés előtt fut, ezért a szabályokban szereplő frázisokat opcionális 1-2-3 darab tetszőleges szóval helyettesítettem.

Tudjuk, hogy az ige vonzatai vele egy tagmondatban vannak [2,9]. Ezt kiegészíthatjük azzal, hogy csak a tagmondat igéjének a vonzatai vannak a tagmondatban. Ebből következik, hogy az ige-koordinációt nem engedjük meg, tehát két finit ige között mindig van tagmondathatár. Ilyenkor megjelölhetjük az összes közbeeső kötőszót, mint *lehetséges* tagmondathatárt [8]. Egyszerű továbblépés, hogy ha egyetlen közbeeső kötőszó van, akkor az lesz a tagmondathatár [9].

Azt figyeltem meg, hogy nemcsak kötőszó, hanem legalább ugyanolyan gyakran közbeeső központozás (vessző, pontosvessző, kötőjel) is lehet tagmondathatár. Tehát két finit ige között (finit igét közvetlenül követő *volna* nem számít annak) megjelölöm ezen írásjelek *utáni* és a kötőszavak *előtti* összes pozíciót, mint lehetséges tagmondathatárt. Majd ezek közül – heurisztikus döntéssel – mindig a leginkább jobbra esőt választom ki, bízva abban, hogy így a legkisebb az esélye annak, hogy egy felsorolás közepére helyezem el a tagmondathatárt.

3 Kiértékelés

A fentiek szerint kialakított magyar tagmondatra bontó eljárás tesztelésére és kiértékelésére az MNSZ részét képező *Magyar Nemzet* napilap anyagából választottam ki véletlenszerűen 200 mondatot. A következő nagyon egyszerű taggelési útmutató szerint végeztem a tagmondatok manuális bejelölését:

1. Jelöljük be a szövegben a tagmondatokat.
2. Minden finit ige külön tagmondatba kerüljön.
3. A tagmondatvégi központozás minden esetben a megelőző tagmondathoz tartozzon.

Sok esetben nehezen tudtam eldönteni, hogy adott ponton valóban van-e tagmondathatár, vagy nincs. Az alábbi szövegrészben a vessző után például végül nem jelöltem tagmondathatárt:

> *Ami a szabad demokratákat illeti,*
> *Pető Iván lemondása tipikusan jelzésértékű ...*

Az eredmények a következők lettek: a 171 bejelölt tagmondathatárból a program 148-at talált meg (23-at hagyott ki), helytelen tagmondathatárt 29-et jelölt meg, azaz:

pontosság = **83,6%** lefedés = **86,5%**

Ezen mérőszámokat befolyásoló tényező lehet az, hogy a szöveg egy viszonylag bonyolult jogi nyelvezeten írt részletet: egy rendeletszöveget tartalmazott. Az eredeti korpuszban sokszor helytelen volt a mondatok határainak megállapítása. Egyszerűbb szerkezetű szöveg esetén, valamint jobb mondatrabontás alkalmazásával minden bizonnyal még növelhetők ezek az értékek.

Amint várható volt, a hibák főleg olyan pontokon jelentkeznek, ahol szinte semmi konkrét jel nem utal arra, hogy ott egy tagmondat kezdődik, nincs kötőszó (sőt esetleg központozás sem), illetve névszói állítmány van valamelyik tagmondatban. Ilyen példa:

> *A kérdés második felére azt felelném,*
> *minden lehetséges s minden az erőviszonyoktól függ.*

4 Lekérdező

A munkálatok egyik eredményeként létrehoztam egy internetes felületen hozzáférhető nyelvészeti kutatóeszközt, melynek segítségével a magyar igei vonzatkereteket, az igék bővítményszerkezetét tudjuk kvantitatívan tanumányozni. A fenti emailcímen igényelt jelszóval érhető el az alábbi címen:

http://corpus.nytud.hu/mazsola

Adott igetőhöz megadhatunk két darab esetraggal (vagy névutóval) meghatározott bővítményt, az ezekhez tartozó konkrét szótövet is megköthetjük, illetve megadhatunk szóközzel elválasztott szótőlistát. Lehetőség van esetek és szótövek kizárására, az a tagadásra. A felület lehetőséget ad kiegészítő szabadszavas keresésre is, itt tetszőleges kiterjesztett reguláris kifejezést használhatunk. Kérhetjük, hogy a találatok az egyik szempont (az igető vagy valamelyik bővítmény) szerint csoportosítva, gyakoriság szerinti sorrendben jelenjenek meg.

Az eredményoldalon a találatok száma alatt a gyakorisági lista található, alatta pedig a korpuszból származó megfelelő példamondatok sorakoznak, egy kattintással könnyen elérhető elrendezésben.

A korábban kidolgozott idiómasági mérést [5] nagy műveletigénye miatt egyelőre nem integráltam az eszközbe, helyette az adott bővítménynek (*y,* pl. *megdöbbenésnek*) a vonzatkeret többi részéhez (*x*, pl. *ad hangot*) viszonyított MI-értéke jelenik meg.

$$\mathrm{MI}(x,y) = \log_2 N\, \mathrm{f}(x,y) / \mathrm{f}(x)\mathrm{f}(y)$$

Mivel egy lekérdezésben x állandó, ez felírható:

$$= \log_2 C\, \mathrm{f}(x,y) / \mathrm{f}(y)$$

alakban, ez pedig, mivel nem érdekesek a konkrét MI értékek, csak össze akarjuk hasonlítani őket, felírható így:

$$= \log_2 \mathrm{f}(x,y) / \mathrm{f}(y)$$

Ebből az értékből számítódik ki a megjelenítéskor a betűméret: nagyobb betűméret, nagyobb MI-értéket, szokatlanabb vonzatkeretet jelent.

Példák:

vesz + ACC
vesz + ACC(*rész*) + INE
hány + ACC
néz + nemACC

Az eszközt aktívan használjuk a magyar-angol gépi fordító projektben az egyes szabad keretek különféle fix lemmákkal való leköthetőségének vizsgálatakor.

5 Fejlesztési lehetőségek

Tervezem a modul olyan továbbfejlesztését, hogy részleges szintaktikai elemzést követően, vagy azt követően *is* lehessen vele tagmondatokat azonosítani.

A tipikus hibák kiküszöbölésére egyik lehetőség egy általános állítmány-azonosító eljárás kifejlesztése, mely a névszói állítmányokat is felöleli. Ekkor valóban teljesülhetne az az elv, miszerint két állítmány között mindig van tagmondathatár.

Köszönöm Gábor Katának és Varasdi Károlynak, hogy rendelkezésemre bocsátották kéziratait.

Bibliográfia

1. Bottyán, G., Sass, B.: Conjugated infinitives in the Hungarian National Corpus. In Garabik, Radovan (ed.): Computer Treatment of Slavic and East European Languages, 3rd International Seminar (SLOVKO2005), Szlovák Tudományos Akadémia, Pozsony (2005) 27-30
2. Gábor K., Héja E., Mészáros Á.: Kötőszók korpusz-alapú vizsgálata. In Alexin Z., Csendes D. (szerk.): MSZNY2003, SZTE, Szeged (2003) 305-306
3. Gábor, K.: Tagmondathatár-kijelölő rendszer. Kézirat. MTA, Nyelvtudományi Intézet.
4. Kim Ch., Hong M.: A Korean Syntactic Parser Customized for Korean-English Patent MT System. In: Salakoski T. et al. (eds.): Advances in Natural Language Processing. LNCS, Vol. 4139. Springer-Verlag, Berlin Heidelberg New York (2006) 44-55
5. Sass B.: Extracting Idiomatic Hungarian Verb Frames. In: Salakoski T. et al. (eds.): Advances in Natural Language Processing. LNCS, Vol. 4139. Springer-Verlag, Berlin Heidelberg New York (2006) 303–309
6. Sass B.: Vonzatkeretek a Magyar Nemzeti Szövegtárban. In: Alexin Z., Csendes D. (szerk.): MSZNY2005, SZTE, Szeged (2005) 257-264
7. Tapanainen, P., Piitulainen J., Järvinen T.: Idiomatic Object Usage and Support Verbs, In: Proceedings of the 17th COLING – 36th ACL, Montreal, Canada (1998) 1289-1293
8. Váradi T., Gábor K.: A magyar INTEX fejlesztésről. In Alexin Z., Csendes D. (szerk.): MSZNY2004, SZTE, Szeged (2004) 3-10
9. Varasdi K.: Coordination. Kézirat. MTA, Nyelvtudományi Intézet.

Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre

Farkas Richárd[1], Szarvas György[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2.
{rfarkas, szarvas}@inf.u-szeged.hu

Kivonat: Cikkünkben bemutatunk egy, számos alkalmazásban kiemelkedő pontosságot elérő statisztikai tulajdonnév-felismerő rendszert. A modell, elsősorban az összegyűjtött nagyméretű tulajdonsághalmaz, illetve az abban rejlő lehetőségek hatékony kiaknázásának köszönhetően több összehasonlításban is versenyképesnek bizonyult a hasonló problémákra ismert legjobb módszerekkel.. Részletesen bemutatjuk a tulajdonnevek azonosításában és kategorizálásában elért eredményeinket magyar nyelvű gazdasági híreken, valamint angol nyelvű újságcikkeken és orvosi kórlapok szövegein. Az eredetileg magyar nyelvre kifejlesztett statisztikai modell apró módosításokkal, minden eddig publikáltnál jobb eredményt ért el a standard angol nyelvű tulajdonnév adatbázison, valamint első helyen végzett egy orvosi kórlapok anonimizálására kiírt nyílt nemzetközi versenyen.

1 Bevezetés

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.

E cikkben bemutatjuk statisztikai tulajdonnév-felismerő rendszerünket, amelynek hatékonyságát három különböző feladaton is vizsgáltuk. Az első tesztfeladat magyar nyelvű gazdasági szövegek feldolgozása volt. Méréseinkhez a Szeged Treebank [3] gazdasági rövidhíreit használtuk. Ugyanazt a rendszert alkalmaztuk angol nyelvű újsághírekben (sport, politikai, gazdasági témákból) szereplő tulajdonnevek felismerésére, melyhez a CoNLL-2003 konferencia adatbázisát használtuk [15], illetve angol nyelvű orvosi zárójelentések anonimizálására. Anonimizálás alatt páciensek, doktorok, kórházak stb. neveinek felismerését és véletlenszerű azonosítókkal való lecserélését értjük [11].

A következő fejezetben bemutatjuk a tulajdonnév-felismerési feladatot, ezután a harmadik fejezetben kerül részletes ismertetésre az általunk készített rendszer és

annak főbb építőelemei. A negyedik részben bemutatjuk a három speciális problémát, amelyen rendszerünket teszteltük, majd elemezzük az elért eredményeket. Végül az utolsó fejezetben összefoglaljuk, értékeljük munkánkat.

2 Tulajdonnév-felismerés

Cikkünkben a *tulajdonnév* kifejezést az angol *named entity* (névkifejezés) magyar megfelelőjeként fogjuk használni. A *named entity* megnevezés magában foglal olyan kategóriákat is, amelyek nem tulajdonnevek (mint például telefonszámok, mennyiségek stb.), de a felismerés elsődleges célpontjai mégis a tulajdonnevek, a többi osztály általában egyszerű szabályok segítségével azonosítható. Hasonló módon elképzelhető az is, hogy bizonyos tulajdonneveket az adott problémánál nem áll szándékunkban jelölni (például az orvosi alkalmazásoknál), mi mégis – az egyszerűség kedvéért – a magyar *tulajdonnév* elnevezés mellett döntöttünk.

2.1 A tulajdonnév-felismerési feladat

A megoldandó feladat kétszintű: egyrészt fel kell ismerni a szöveg(ek)ben az előre definiált kategóriákba tartozó tokensorozatokat, másrészt be kell azokat sorolni a megfelelő osztályokba. Az osztályozás során meg kell különböztetni a tulajdonnevek kezdő tokenjeit, és a tulajdonnév részét képező belső elemeket. Ennek elsősorban akkor van jelentősége, amikor a szövegben egymást követően több azonos kategóriába tartozó tulajdonnév található, mert ilyenkor ezek segítségével állapítható meg az elemek kezdőpozíciója.

Az adott problémákhoz rendelkezésre áll egy-egy tanító adathalmaz, ahol a tulajdonneveket kézzel bejelölték. A cél ennek felhasználásával egy olyan modell tanítása, amely ismeretlen szövegen is hatékonyan felismeri az adott kategóriákat (induktív tanulás). Fontos megjegyeznünk, hogy a tanult modell csak a tanító halmazhoz hasonló jellemzőkkel rendelkező szövegeken működik pontosan.

2.2 A tulajdonnév-felismerés története

A tulajdonnév-felismerés problémájával a 90-es évek közepétől foglalkoznak. A Message Understanding Conference (MUC) [2] sorozat angol nyelvű újsághírek automatikus feldolgozását tűzte ki célul. A MUC-7 során a tulajdonnevek azonosítása és a *személynév, földrajzi név, szervezet, egyéb* kategóriákba sorolása, valamint egyéb, időt, mennyiséget stb. leíró kifejezések felismerése volt a feladat. Az utóbbi években további nyelvekre fókuszált a kutatás, mint például a spanyol, a német, a kínai. A Conference on Computational Natural Language Learning (CoNNL) által meghirdetett nyílt versenysorozaton 2003-ban a tulajdonnevek felismerése volt a feladat egyazon modellel angol és német nyelvű szövegekben [15].

Ebbe a trendbe jól illeszkednek a magyar nyelvvel kapcsolatos kutatásaink: létrehoztuk az első releváns méretű (200.000 szó) magyar nyelvű tulajdonnévi korpuszt [12], valamint implementáltuk az első statisztikai alapú magyar tulajdonnév-

felismerő modellt, melynek eredményei az angolra publikált eredményekkel összevethetőek.

A szakterület egy másik folyamatosan bővülő iránya a felismerő rendszerek alkalmazása különböző domainekre. Az első rendszerek (MUC, CoNNL) általános újságcikkekre koncentráltak, napjainkra azonban előtérbe került például a bioinformatikai vagy orvosi szövegek feldolgozása is [16][11]. Annak érdekében, hogy megvizsgáljuk, hogy az újsághírekre kifejlesztett rendszerünk hogyan viselkedik más domainen, idén részt vettünk egy orvosi kórlapok anonimizálását célul kitűző versenyen [14]. A versenyen első helyezést értünk el, ami bizonyítja, hogy sikerült a tulajdonnév-felismerés problémájára általánosan felhasználható eszközt építenünk.

3 A tulajdonnév-felismerő rendszer felépítése

Rendszerünk három főbb összetevőből áll: a szavakhoz tartozó tulajdonságvektorok kinyeréséből, több statisztikai modell betanításából, majd azok összekombinálásából az ismeretlen szöveg bejelölésekor.

3.1 Megközelítésünk

Az általunk alkalmazott gépi tanulási módszer eltér a problémára leggyakrabban alkalmazott, legsikeresebbnek tartott technikáktól. Szekvenciák tanulása helyett (mint például Conditional Random Fields, Maximum Entropy Models stb. [1]) szóalapú osztályozásként kezeljük a problémát. Ilyen típusú modelleket korábban is alkalmaztak tulajdonnevek felismerésére – leggyakrabban Support Vector Machine osztályozót [8][11] –, de a szekvenciális tanulók az elmúlt években sokkal „divatosabbá" váltak.

Az általunk választott döntésifa-alapú megközelítés gyors tanítást és kiértékelést biztosít, ami lehetővé tette hatalmas jellemzőkészlet használatát. Természetesen – a szóalapú osztályozás ellenére – a környezetre, kontextusra vonatkozó információkat mi sem hagyjuk figyelmen kívül: jellemzőként beépülnek a modellbe a megelőző és rákövetkező szavak főbb tulajdonságai, valamint a megelőző szavakra a modell által javasolt tulajdonnévi címkék.

3.2 Nagyméretű tulajdonsághalmaz

Egy igen bő tulajdonsághalmazt gyűjtöttünk össze, amely leírja az egyes szavakat, illetve azok rögzített szélességű környezetét. A következő kategóriákba sorolhatjuk ezeket a jellemzőket:

- **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 (webről gyüjtött frekvenciaszótárban), 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 (szűrésüket a szóalakok egyes osztályok közti entrópiája alapján végeztük el).
- **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, amelyeket mind az Internetről töltöttünk le.

Az egyes feladatoknál felhasználtunk még néhány problémaspecifikus jellemzőt, ezeket a következő fejezetben a problémák tárgyalásánál mutatjuk be.

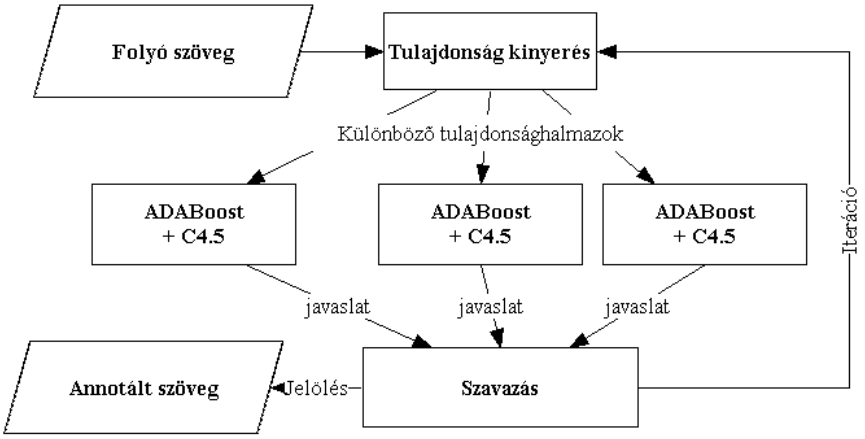
Fontos megemlítenünk, hogy mind a magyar, mind az angol nyelvű feladatoknál próbálkoztunk nyelvtani információk felhasználásával (POS kódok és chunk kódok), de ez egyik esetben sem javított értékelhetőt a modell pontosságán, sőt néhol összekavarta a rendszert, ezért teljes mellőzésük mellett döntöttünk.

3.3 A tulajdonsághalmaz felosztása és újrakombinálása

Az egyes tulajdonságok és azok halmazai más és más szemszögből közelítik/magyarázzák a problémát. Úgy gondoltuk, az általunk összegyűjtött jellemzők nagy száma lehetővé teszi, hogy részekre bontsuk a teljes halmazt úgy, hogy a részhalmazokon külön-külön tanítva is erős modellek legyenek építhetők. Ezeknek a modelleknek a kombinálása később jobb eredményre vezethet, mint az eredeti teljes halmazon tanított egyedi modell.

Hipotézisünk helyességének vizsgálatára a következő újszerű eljárást – amelyet a az 1. ábra szemléltet – dolgoztuk ki: a jellemzőket csoportosítottuk úgy, hogy egy csoportba hasonló jellegű tulajdonságok kerüljenek. Az így kialakított csoportokat – bonthatatlan egységnek tekintve – azok összes lehetséges kombinációját kiértékeltük a tanító halmazon (futásigény miatt egyszerű döntési fákat tanítottunk csak) és az öt legjobban teljesítő kombinációt tartottuk csak meg (tehát az eredeti teljes tulajdonsághalmaz öt nem diszjunkt részhalmazát).

Ezekre külön-külön tanítottunk egy-egy teljes modellt (boostingolt döntési fákat) majd a tesztelési fázisban a modellek jelölési javaslatait szavaztattuk a következő egyszerű döntési szabály segítségével: *ha van három megegyező javaslat, akkor fogadjuk azt el, ellenkező esetben adjunk nem-tulajdonnévi címkét*. Tehát ha a modellek nem képesek „egyezségre jutni” akkor inkább nem jelölünk tulajdonnevet. Ez a stratégia statisztikailag jobb eredményt ad, mintha találomra választanánk az öt javaslat közül, ugyanis a hibásan jelölt tulajdonnév két büntetőpontot von maga után (a pontosság és a fedés is sérül), míg a nem jelölt tulajdonnév csak egyet (a fedés csökken).



1. Ábra A tulajdonnév-felismerő rendszerünk sematikus váza

3.4 Iteratív tanulás

Az orvosi kórlapokon szerepelnek a – folyó szövegrészeken túl – rekordba rendezett információk is. Ezekben a tulajdonnevek felismerése és kategorizálása lényegesen egyszerűbb feladat, mint a folyó szövegekben. Ezért ennél a problémánál a következő iteratív tanítást hajtottuk végre: az első modell célja csak a rekordokban szereplő információk kinyerése volt. Az itt megtalált neveket, azok részeit (pl. személyneveknél hasznos a családi és keresztnév különválasztása) felhasználtuk egy következő tanítási fázisban, ahol már folyó szöveget is elemeztünk, de csak a biztos helyeken jelöltünk tulajdonneveket. Ezeket ismét hozzáadtuk az „ismert" egyedek listájához, amivel egy végső tanítási lépésben már az összes tulajdonnév-felismerése volt a cél.

A fent leírt módszert általánosíthatjuk: minden iterációban csak a biztos egyedeket jelöljük, hogy ezekből tulajdonságokat nyerünk ki, amelyeket a következő iterációbeli tanítás folyamán felhasználunk. Tehát egy nagyon pontos (de lefedésben gyenge) modellből kiindulva, iterálva jutunk el egy végső modellig, aminek a pontossága és fedése a kívánt szintű. Ennek az általánosított módszernek az empirikus vizsgálatát a jövőben fogjuk elvégezni.

3.5 Gépi tanuló algoritmusok

Számos klasszifikációs technikát kipróbáltunk a probléma megoldása folyamán (Neurális hálók, Support Vector Machines stb), de legjobbnak a döntésifa-alapú technikák bizonyultak [4]. Ez elsősorban annak köszönhető, hogy az ID3 algoritmus [9] diszkrét tulajdonságok – és a modellünkben túlnyomórészt diszkrétek a tulajdonságok – kezelésére lett kidolgozva (a numerikus tanulókkal ellentétben), másodsorban a fák tanításának és predikciójának sebessége kezelhetővé tette a hatalmas tulajdonsághalmazt.

Az egyszerű döntési fák által elért pontosság javítására Shapire AdaBoostM1 [10] algoritmusát használtuk. A módszer több iteráción keresztül javít egy tetszőleges, gyenge tanulót (mint például esetünkben a döntési fát) úgy, hogy minden iterációban a tanító halmazon eltévesztett minták súlyát növeli, a helyesen jelöltekét csökkenti a következő modell mintaválasztásához.

A kísérletekhez a WEKA keretrendszer [7] egy kiegészített változatát használtuk. A döntési fánál mindig az alapértelmezett paramétereket, a boostingnál 30 iterációt használtunk. A paraméterek finomhangolásával minden bizonnyal talán modelleink további javulását érhetnénk el.

4 Eredmények

A három tulajdonnév-felismerési problémát és a kapcsolódó eredményeinket (mindenhol frázisszintű $F_{\beta=1}$ metrikát használva[11]) kronológiai sorrendben mutatjuk be.

4.1 Magyar nyelvű gazdasági rövidhírek elemzése

A Szeged Korpusz 200 ezer szóból álló, gazdasági rövidhíreket tartalmazó szegmensében (NewsML) bejelöltük a *szervezet, személy, hely* és *egyéb* kategóriákba tartozó tulajdonneveket (SzegedNE korpusz). Az annotátorok közti végső egyetértési szint 99,89% lett [12]. Ez a korpusz az első magyar nyelvű tulajdonnévi korpusz[12].

A korpusz jellegzetessége a szervezetek túlsúlya a tulajdonneveken belül. Ráadásul ezek felismerése általában egyszerűbb a többi osztályhoz képest a cégformára utaló frázisvégződések (*kft., rt*. stb.) miatt. Ez magyarázza az angolra publikált eredményeknél lényegesen jobb eredményeinket.

1. Táblázat: Eredmények a SzegedNE korpuszon

	$F_{\beta=1}$
Szervezet	95,84%
Személy	94,67%
Hely	95,07%
Egyéb	85,96%
Mindösszesen	94,77%

Az előző fejezetben bemutatott alaprendszer – a korábban még nem látott szövegrészleteken mért – osztályszintű eredményeit az 1. Táblázat tartalmazza [13]. Sajnos ezen eredményeinket nem tudjuk más rendszerekkel összehasonlítani, ez idáig csak szabályalapú rendszerek készültek magyar nyelvű tulajdonnevek felismerésére [6], a mi rendszerünk az első statisztikai modell.

[11] A kiértékelést a CoNNL versenyhez kiadott szkripttel végeztük, ami letölthető a http://www.cnts.ua.ac.be/conll2003/ner/ oldalról

[12] A korpusz kutatási célokra ingyenesen hozzáférhető a http://www.inf.u-szeged.hu/~hlt/index.html oldalon

4.2 Angol nyelvű újságcikkek elemzése

A fent említett négy tulajdonnévi kategória felismerése volt a célpontja a CoNLL által kiírt nyílt versenynek 2003-ban [15]. A tanító adatbázis Reuters híreket[13] tartalmazott 1996-ból, amelyek felöleltek sport, politikai és gazdasági témákat egyaránt.

A korpuszon a *szervezet* kategória pontossága szignifikánsan rosszabb, mint a magyar szövegen volt (mint ahogy az a 2. Táblázatból is leolvasható). Ez annak köszönhető, hogy a sporthírekben a csapatnevek (amik természetesen *szervezet*ek) illetve a városok nevei nagyon nehezen különböztethetőek meg (pl. *Los Angeles beat Boston*).

A feladat specialitásait kihasználandó az alap tulajdonsághalmazt két további tulajdonsággal bővítettük: először is a legtöbb hír három jól elkülönülő részre volt bontható (cím/rövid összefoglaló a cikk elején; riporter, helyszín, dátum; maga az újsághír), másrészt a Reuters témakódok alá sorolja a híreket, ezeket a kódokat is felvettük külön jellemzőként. Ez utóbbitól reméltük a fent említett város-csapat többértelműség feloldását. Sajnos ebben csalódnunk kellett.

2. Táblázat: Eredmények a CoNLL-2003 adatbázison

	egyéni	hibrid
Szervezet	84,53%	88,32%
Személy	93,55%	96,27%
Hely	92,90%	93,43%
Egyéb	79,67%	82,29%
Mindösszesen	89,02%	91,41%

A versenyen győztes egyéni modellhez [5] képest rendszerünk 2,3%-kal kisebb relatív hibával működik a kiértékelési adatbázison, de igazi haszna hibrid rendszerekben mutatkozik meg: mivel modellünk más megközelítésen alapszik, mint a versenyen induló modellek, eredményeinket kombinálva az ott szereplő legjobb rendszerekkel lényegi javulás érhető el. A versenyen szereplő három legjobb modell többségi szavazásos kombinációja 90,3%-os eredményt hozott. Ha a mi rendszerünk lép a győztes modell helyébe a szavazás utáni eredmény 91,41%, ami már számottevő, 11,44%-os relatív hibacsökkenést jelent [13].

4.3 Orvosi kórlapok anonimizálása

Az orvosi szakszövegek adatbányászati célú felhasználásához elengedhetetlen az abban szereplő személyes adatok védelmének biztosítása. Ezért mielőtt publikussá válik egy orvosi szövegekből álló adatbázis, az abban előforduló személyek neveit (orvos, páciens), telefonszámát, lakhelyét, a kórház nevét stb. anonimizálni kell. A feladat tehát itt is tulajdonnevek bizonyos jól körülhatárolt osztályainak felismerése és kategóriákba sorolása.

Erre az információkinyerési feladatra az MIT Computer Science and Artificial Intelligence Laboratory, Informatics for Integrating Biology and the Bedside (i2b2) kutatóintézet nyílt versenyt írt ki idén nyáron [11], amelyre beneveztük rendszerünket

[13] http://www.reuters.com/researchandstandards/

is. A szervezők biztosítottak egy 200 ezer szavas annotált tanító adathalmazt, majd egy körülbelül 50 ezer szavas halmazon értékelték ki a rendszereket.

Az orvosi kórlapok – melyek szerkezete ugyan nem kötött, de – tartalmaznak strukturált, rekordokba rendezett egységeket. A rekordok határait és azok belső szerkezetét egyszerű szabályokkal azonítottuk, és a fejléceket új tulajdonságként hozzátettük az alap jellemzőkészlethez (ezeket a fejléceket használtuk fel a 3.4 fejezetben bemutatott iteratív tanulás első iterációjában is). A másik tulajdonság, amivel a halmazt bővítettük az a könnyebben felismerhető osztályokra (*dátum, életkor, telefonszám, azonosítók*) felírt reguláris kifejezések voltak.

Az egyik legfontosabb jellemzőnek a szó környezetének – implikáló – szóalakjai bizonyultak. Arra, hogy a környezetet pontosan hogyan használjuk fel, három különböző módszert dolgoztunk ki. Ennél a feladatnál az idő rövidsége miatt nem hajtottuk végre a teljes tulajdonsághalmaz-felbontást és újrakombinálást (amit részletesen a 3.3. fejezetben mutattunk be), helyette az erre a három módszerre épített modelleket kombináltuk többségi szavazással.

3. Táblázat: Eredmények az i2b2 adatbázison

	$F_{\beta=1}$
Kórház	92,69%
Doktor	95,88%
Páciens	96,21%
Hely	63,79%
Életkor	100,00%
Dátum	99,25%
Azonosító	99,33%
Telefonszám	98,31%
Mindösszesen	97,41%

A 3. Táblázat tartalmazza a teszthalmazon, az egyes osztályokon elért eredményeinket. Ahogy azt vártuk, a négy egyszerűen felismerhető osztály pontossága látványosan jobb, mint a klasszikus tulajdonnévosztályoké. A *hely* kategória erősen alulreprezentált, a tanító halmazban mindössze 150 hely kifejezés szerepelt, ez magyarázza az igen gyenge felismerési eredményét. A másik három „érdekes” osztályon elért eredményeink összehasonlíthatók a korábbi, újsághírekkel foglalkozó feladatok hasonló osztályain (*személy*, *szervezet*) elért pontosságokkal [14].

A versenyen 16 rendszer vett részt. A testre szabott általános tulajdonnév-felismerő rendszerünkkel elért 97,41%-os pontossággal a versenyen első helyezést értünk el.

4.3 Az egyes feladatok modelljei közti eltérések

E fejezetben bemutattuk, hogyan alkalmaztuk tulajdonnév-felismerő rendszerünket három különböző probléma megoldására. A magyar nyelvű gazdasági hírekre fejlesztett modell angolra adaptálásakor csak a szótárakat cseréltük le azok angol nyelvű megfelelőire, valamint felhasználtuk az újsághírek speciális tulajdonságait (témakód, dokumentumon belüli rész).

A kórlapok anonimizálásánál szintén kihasználtuk a dokumentumok szerkezetét (rekord fejlécei és reguláris kifejezések), azon felül csak a cégvégződés listát (*ltd.* stb.) cseréltük le kórháznévvégződés-listára (mint pl. *Hospital*). Minden más tekintetben ugyanazokat a tulajdonságokat, ugyanazokat a tanulókat, paraméterbeállításokat és technikákat használtunk mindhárom statisztikai modell építése folyamán.

5 Diszkusszió

Az előző fejezetekben bemutattuk statisztikai tulajdonnév-felismerő rendszerünket, amelyet sikeresen alkalmaztunk magyar nyelvű gazdasági rövidhírekben található tulajdonnevek felismerésére és kategorizálására. A kisebb változtatásokon keresztülment modellel minden korábbinál jobb eredményt értünk el a standard angol nyelvű tulajdonnévfelismerési adatbázison (CoNLL), és megnyertük az i2b2 orvosi kórlapok anonimizálására kiírt nemzetközi nyílt versenyt is.

Rendszerünk sikerét elsősorban az összegyűjtött nagyméretű tulajdonsághalmaznak és abban rejlő potenciálok hatékony kiaknázásának (tulajdonságmegosztás, majd rekombináció, jól megválasztott tanuló modell) köszönheti. Szeretnénk még egyszer kiemelni, hogy az elemzéshez csak felszíni jegyeket, illetve a tanító adatbázisból kinyerhető statisztikai jellemzőket használtunk fel. A rendszer nem függ semmilyen külső modelltől – mint például POS-tagger – és nincs szüksége semmilyen nyelv-, illetve domainfüggő szakértői tudásra (az i2b2 versenyen például számos más rendszer használta az orvosi *Medical Subject Headings* kódokat).

Természetesen – mint minden induktív tanulási modell – rendszerünk csak akkor alkalmazható, ha rendelkezésre áll megfelelő méretű tanító adatbázis (mindhárom esetben a körülbelül 200 ezer szavas halmaz kielégítőnek bizonyult), és a jelölendő szöveg főbb jegyeiben megegyezik a tanító halmazzal.

A jövőben ezért szeretnénk megvizsgálni, hogy milyen lehetőségeink vannak, ha nem áll rendelkezésre elégséges méretű előre bejelölt példákat tartalmazó adatbázis (ugyanis annak előállítása általában igen költséges). Ezen az úton meg is tettük az első lépéseket, folyamatban vannak olyan kísérletek, amelyekben azt vizsgáljuk, hogy jelöletlen szövegek, illetve internetes keresőmotorok hogyan segíthetik a jelölt szövegeken tanuló modelleket.

Bibliográfia

1. Hai L. Chieu and Hwee T. Ng.: Named Entity Recognition with a Maximum Entropy Approach. Proceedings of CoNLL-2003 (2003)
2. Nancy Chinchor.: MUC-7 Named Entity Task Definition. Proceedings of Seventh Message Understanding Conference (1998)
3. Dóra Csendes, János Csirik and Tibor Gyimóthy: The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus. Proceedings of TSD 2004, vol. 3206 (2004)
4. Richárd Farkas, György Szarvas, András Kocsor: Named Entity Recognition for Hungarian using various Machine Learning Algorithms. Acta Cybernetica (2006)

5. Radu Florian, Abe Ittycheriah, Hongyan Jing and Tong Zhang: Named Entity Recognition through Classifier Combination. Proceedings of CoNLL-2003 (2003)
6. Kata Gábor, Enikő Héja, Ágnes Mészáros, Bálint Sass: Nyílt tokenosztályok reprezentációjának technológiája. IKTA-00037/2002, Budapest, Hungary (2002)
7. S. Garner. Weka: The waikato environment for knowledge analysis. (1995)
8. C. Lee, W.-J. Hou and Chen, H.-H. Annotating multiple types of biomedical entities: A single word classification approach. In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004).
9. Ross Quinlan: C4.5: Programs for machine learning, Morgan Kaufmann (1993)
10. Rob E. Shapire: The Strength of Weak Learnability. Machine Learnings, Vol. 5 197-227 (1990)
11. Tawanda Sibanda, Ozlem Uzuner, and Ozlem Uzuner. Role of local context in automatic deidentification of ungrammatical, fragmented text. In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, pp 65-73, New York City, USA, June (2006)
12. György Szarvas, Richárd Farkas, László Felföldi, András Kocsor, János Csirik: A highly accurate Named Entity corpus for Hungarian. Proceedings of International Conference on Language Resources and Evaluation (2006)
13. György Szarvas, Richárd Farkas, and András Kocsor: A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. DS2006, LNAI 4265, pp. 267-278 (2006)
14. György Szarvas, Richárd Farkas, Szilárd Iván, András Kocsor, Róbert Busa-Fekete: An Iterative Method for the De-identifcation of Structured Medical Text. In Proceedings of American Medical Informatics Association, (2006)
15. Erik F. Tjong Kim Sang, and Fien De Meulder: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Proceedings of CoNLL-2003 (2003)
16. Y. Tsuruoka J-D. Kim, T. Ohta and Y. Tateisi. Introduction to the bio-entity recognition task at jnlpba. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA),Geneva, Switzerland, 2004.

Magyar nyelvű tulajdonnév-felismerés maximum entrópia módszerrel

Varga Dániel[1], Simon Eszter[2]

[1] Budapesti Műszaki Egyetem -- Média Oktató és Kutató Központ
daniel@mokk.bme.hu
[2] BME Kognitív Tudományi Tanszék
esimon@cogsci.bme.hu

Kivonat: Cikkünkben bemutatunk egy maximum entrópia módszeren alapuló statisztikai tulajdonnév-felismerő rendszert magyar nyelvre. A rendszer bemenetként morfológiailag elemzett szöveget dolgoz fel, ráépülve a hunpos morfológiai egyértelműsítőre. A felismerés hatásfokát tulajdonnév-címkézett magyar nyelvű korpuszon értékeljük.

1 Bevezetés

A tulajdonnév-felismerés (named entity recognition, NER) a természetes nyelv feldolgozását célzó alkalmazások közül az egyik legnépszerűbb, mivel hatékonyan automatizálható, és eredménye hasznos bemenete különböző magasabb szintű információ-kivonatoló és információ-feldolgozó rendszereknek.

A NER során egy bemeneti tokensorozatban kell tulajdonnevet alkotó intervallumokat (chunk) kijelölünk, ezeket véges sok kategóriába besorolva (például személy, szervezet, hely, egyéb). Egy NER algoritmus kiértékelése manuálisan annotált korpusszal való összevetés útján történik, és szokásosan maga az algoritmus is egy ilyen korpuszból tanulja meg a paramétereit automatikus módon.

Angol nyelvre sok tucat cikket publikáltak NER eljárásokról. Szinte minden ma ismert gépi tanulási módszert felhasználtak már NER címkéző építéséhez. Néhány példát említve: Rejtett Markov modellt használtak a BBN IdentiFinder építői [7], valamint Zhou és Su [3]; maximum entrópia módszert Borthwick [1] [2] és Chieu és Ng [3]; döntési fát Sekine et al. [8] [9]. Ezeknek a rendszereknek kevés a nyelvfüggő elemük, elvileg könnyen adaptálhatóak új nyelvekre. Magyar nyelvre ennek ellenére egyetlen kvantitatív vizsgálatról van tudomásunk: Szarvas et al. [11] publikált eredményeket Boosting és C4.5 döntési fa módszerekre épülő tulajdonnév-felismerő technológiájukról. Az általuk konstruált rendszer state-of-the-art pontosságot ér el angol nyelvre, és magyar nyelvre adaptálva a Szeged NER korpuszon [10] 94.77% CoNLL F-mértékű a pontossága. (A mérték definícióját lásd [13].)

Az alábbiakban felvázoljuk a rendszerünk architektúráját, leírjuk az általunk használt jegyeket, majd a mérési metodológia bemutatása után publikáljuk méréseink eredményeit.

2 A korpusz

Statisztikai alapú tanulórendszerünk tanításához és kiértékeléséhez a Szeged NER korpuszt [10] alkalmaztuk, cikkünk leadásának pillanatában ez volt az egyetlen gépi tanításhoz megfelelő méretű magyar nyelvű tulajdonnév-korpusz.[14]

A Szeged NER korpusz a Szeged Korpusznak [4] egy több mint 220 ezer szövegszónyi tematikusan válogatott és manuálisan tulajdonnév-annotált része. A szöveg legjellemzőbb tulajdonsága tematikus homogenitása: kizárólag gazdasági hírek szerepelnek benne. Ennek következtében nagyon nagy számban találhatók benne intézménynevek (magyar és multinacionális vállalatok, hivatalos szervezetek stb.), ez a kategória gyakoriságban dominálja a többit.

A Szeged NER Korpusz a CoNLL [13] típusrendszerét és címkézési konvencióit követi. Ennek megfelelően a következő kategóriákkal dolgoztunk: személynevek (PER), intézménynevek (ORG), földrajzi nevek (LOC) és egyéb tulajdonnevek (MISC), utóbbiak leggyakrabban márkanevek, címek és tőzsdeindexek megnevezései.

3 Tulajdonnévtárak

Rendszerünk fejlesztéséhez a különböző forrásokból származó tulajdonnévtárakat (gazetteer, a továbbiakban szótár) gyűjtöttünk össze:

- vezetéknevek
- keresztnevek
- magyar településnevek
- magyar nyelvű országnevek
- magyar utcanevek
- magyar cégnevek
- nemzetközi cégnevek
- cégnévvégződések (Rt., Kft., Ltd.)
- utcanévvégződések (út, utca, tér)
- pénzügyi rövidítések

Az első hat esetben forrásunk egy magyar telefonkönyv aggregált változata, illetve egy webes adatbázis volt. A cégnevek és utcanevek listáját automatikus eszközökkel megtisztítottuk a végződésektől, amelyek külön végződéslistába kerültek. A nemzetközi cégneveket tartalmazó listát Szarvas György és munkatársai bocsátották rendelkezésünkre.

Az egyetlen eset, amikor a fejlesztő (devel) korpuszon elkövetett hibák elemzése új szótár felvételéhez vezetett, a pénzügyi rövidítések tára. A devel korpuszban rendkívül gyakran szereplő tőzsdeindex-nevek (DAX, Libor, Nasdaq) MISC helyett sokszor tévesen ORG-nak címkézte az algoritmus. A „felületes szemlélő“ ezeket jó okkal minősítheti formájuk és szövegkörnyezetük alapján ORG-nak. Mint látni fogjuk, rendszerünk nem használ közvetlen módon a tanítókorpuszból épített szótárat,

[14] Reményeink szerint ez a helyzet hamarosan változni fog a HunNER korpusz megépítésével.

így ezeknek az egyedi eseteknek a megtanulása nehézséget okozhat a számára. A probléma megoldására egy weben található tőzsdei rövidítésgyűjteményben automatikus módon azonosítottuk a tanítókorpuszban ORG-ként soha nem szereplő elemeket, és ezekből szótárat alkottunk. Megjegyezzük, hogy a szótár alkalmazása nem javította, sőt kis mértékben rontotta a tesztelési korpuszon elért teljesítményt.

A rendszerünk végső, itt ismertetett változatában található szótárak (a fenti esettől eltekintve) teljesen azonosak a rendszer fejlesztésének megkezdése előtt véglegesítettekkel. Ennek oka a következő: bár a fejlesztés során komoly túl- és alulgenerálásokat találtunk a szótárakban, és ezeket sokszor javítottuk is, de azt tapasztaltuk, hogy az így elért pontosságnövekedés nem számottevő, és módszertani előnyei okán inkább visszatértünk a korpuszra nem rátanult szótárakhoz.

Az általunk épített szótárakat a rendszer forráskódjához hasonlóan szabad forráskódú licensz alatt publikáljuk.

4 Architektúra

Rendszerünk az [1] vagy [3] által ismertetett rendszerek architektúráját követi. Ha statisztikus gépi tanulási implementációt tervezünk, el kell döntenünk, hogy egy címkézési döntés meghozatalához milyen információt használjunk fel. Ez a feladat a jegykinyerés (feature extraction). NER esetében kézenfekvő és standard megoldás, hogy a vizsgált token környezetében levő tokenekről nyerünk ki nagy mennyiségű hasznosnak remélt egyszerű információt, például hogy nagy kezdőbetűs-e, szerepel-e valamely szótárunkban vagy mi a szófaja. Ezután valamilyen gépi tanulási algoritmusra bízzuk, hogy a tanulókorpusz alapján kiválogassa, hogy a nagy mennyiségű összegyűjtött jegyből melyek és hogyan segítenek a címkézési döntésben.

Gépi tanulási algoritmusként a maximum entrópia módszert választottuk. Ennek kimenete valószínűségeloszlás a választható címkéken, amelyből egy úgynevezett simító eljárással választjuk ki a legmegfelelőbbet.

4.1 Jegykinyerés

Megközelítésünk az volt, hogy minél nagyobb méretű, de egyszerűen implementálható jegyhalmazt építünk, kihasználva, hogy a maximum entrópia módszer nagyon nagy számú jegy hatékony feldolgozására képes. Jegyeink nagy része elsősorban a szavak egyszerű formai tulajdonságait írja le. Ugyanakkor azt is kihasználtuk, hogy rendelkezésünkre áll a hunpos morfológiai egyértelműsítő [6], amely ismeretlen szavak elemzését is elvégzi.

Az alábbi információkat építettük bele az algoritmusunkba:

1. Előfordul-e a token környezete valamely tulajdonnévtárunk valamely tételében, és ha igen, akkor a token a kifejezésnek mely pozícióján szerepel: az elején, a végén, a belsejében vagy egyszavasként? A többszavas kifejezések utolsó szavával való illeszkedést nem betű szerinti egyezésként definiáltuk, hanem egy alkalmasan választott kezdőszeleten való egyezésként.

2. Mondat eleje, mondat vége. (A Szeged Korpusz mondatra szegmentált, és a korpusz más felhasználóihoz hasonlóan ezt az információt az algoritmus számára hozzáférhetőnek tekintettük.)
3. A token igen/nem értékű formai jegyei: nagybetűs, csupa nagybetűs, tartalmaz kisbetű-nagybetű szekvenciát (pl. iPod), szám, számmal kezdődik, számot tartalmaz, kötőjelet tartalmaz, pontot tartalmaz.
4. A token felszíni, karakterlánc értékű formai jegyei: a token karakterszáma, a szóalak, a token három és öt hosszú kezdőszelete, a szó összes három karakterből álló összefüggő részkarakterlánca.
5. A hunpos morfológiai egyértelműsítő által szolgáltatott információk: szófaji kategória (NOUN, ART, NUM, ADJ, VERB, stb.), a szónak a hunpos egyértelműsítő által javasolt lemmája. Felismerte-e a hunmorph morfológiai elemző a szóalakot? Az azonosított lemma más kapitalizációjú-e, mint a token maga?

Az aktuálisan vizsgált token tehát megkapja a fent leírt jegyek közül azokat, amelyeket saját tulajdonságai implikálnak. Ezen kívül megkapja azokat a jegyeket is, amelyek a környezetében, de a mondatán belül álló szavak jegyei, mellékelve azt az információt, hogy mekkora eltolásra lévő szóból származik a jegy (pl. *oov.pre4*, *allcaps.post3*, *multi.start.pre1*). A figyelembe vett tokenek ablakának méretét optimalizáltuk; kényelmi okokból csak két paraméterre: egyrészt a karakterlánc értékű formai jegyek összességére, másrészt az összes többi jegyre. Méréseink szerint a karakterlánc értékű formai jegyeknél a 3 sugarú környezet (7 token) jegyeinek figyelembe vétele vezet optimális eredményhez, a többi jegy esetében az optimális az 5 sugarú környezet (11 token). Természetesen ezek az értékek nem univerzálisak, de modellünk többféle paraméterbeállítása mellett is a legkedvezőbbnek bizonyultak.

4.2 Címkék

A NER mint címkézési feladat eredeti formájában kevésbé alkalmas alanya gépi klasszifikálási módszereknek, mint az alábbi átcímkézett változatában: Minden tokent az alábbi 17 osztály egyikébe kell sorolnunk: {0, LOC.egytagú, LOC.eleje, LOC.belseje, LOC.vége, ORG.egytagú, …, MISC.vége}. Ennek a megoldásnak két előnye van a nyers ötelemű címkézéshez képest. Egyrészt a tanulóalgoritmus számára könnyebb felismerni olyan korrelációkat, amelyek speciálisan a tulajdonnév elejére és végére jellemzőek. Másrészt ez a címkézés implicit konzisztenciafeltételeket tartalmaz: például *.belseje után nem következhet *.eleje. Mint látni fogjuk, ezt felhasználhatjuk arra, hogy a gépi tanuló algoritmus kimenetét utófeldolgozva javítsuk.

4.3 Maximum entrópia

Címkézési algoritmusnak a maximum entrópia módszert választottuk. Ez a módszer már jó eredményeket hozott a hunpos morfológiai egyértelműsítő rendszer súlyozott morfológiai elemző (WMA) komponensének megépítésekor. Az általunk választott implementáció ezúttal könnyű beépíthetősége és nagy sebessége okán Zhang Le [15] rendszere volt, amely tanításhoz az L-BFGS algoritmust [17] alkalmazza.

Adathalmazainkon a L-BFGS iteratív tanulóalgoritmus 100 alatti iterációszámnál még nem kezd el konvergálni, a modell teljesítménye a pontos iterációszámtól erősen és kiszámíthatalanul függ. Publikált számaink 300-as iterációszám mellettiek, itt már tipikusan stabilizálódnak a modellek, és azok teljesítménye. Ugyanakkor a 300 iteráció relatíve magas, közel egy órás futásideje miatt az egyes elemi változtatások hasznos mivoltát gyakran kis (30 vagy 100) iterációszámmal épült modelleken vizsgáltuk, ami esetenként téves döntésekhez is vezethetett.

A futásidő kapcsán említjük meg, hogy a kezdőszeletek és karakter-trigramok jegyként való felvétele, és a nagyméretű ablakok alkalmazása miatt a jegyek száma rendkívül magas. A 200,000 példát tartalmazó tanítási korpuszon 250,000 különböző jegy összesen 10 millió előfordulása szerepel.

4.4 Simítás

A maximum entrópia klasszifikáló alkalmas arra, hogy több, helyességi valószínűséggel súlyozott alternatív javaslatot tegyen a címkére. Ennek fontos előnye, hogy erre épülve úgynevezett simítást implementáltunk, felülbírálva olyan lokális döntéseket, amelyek egymással inkonzisztensek. A gépi tanulási szakirodalomban elterjedt [3] módszer lényege, hogy 0 valószínűségű eseménynek tekintjük a lehetetlen átmeneteket (pl. LOC.belseje után ORG.eleje következik), uniform eloszlásúnak tekintjük a valid átmeneteket, és a maximum entrópia módszer által az egyes tokenekre kibocsátott valószínűségeloszlásokat függetlennek tekintve Viterbi módszerrel kiszámoljuk, hogy a mondatra mi a legnagyobb valószínűségű címkézéssorozat. Ez szükségszerűen valid lesz. Méréseink szerint ez a paramétermentes utófeldolgozási lépés egy tipikus mérési konfigurációban 0.5% körüli értékkel javítja meg rendszerünk F-pontszámát.

5 Mérések

5.1 Módszertan

A tanulóalgoritmus hatékonyságának méréséhez egy tesztkorpuszt kell címkéznie az algoritmusnak. A korpuszban rendelkezésre álló aranymérték (gold standard) címkék alapján mérhető az algoritmus pontossága és fedése. A NER rendszerek hatásfokának mérésére hagyományosan alkalmazott CoNLL kiértékelési függvény egy az algoritmus által azonosított tulajdonnevet címkéjének helyessége és a lefedett intervallum pontos megtalálása alapján is pontoz (0, ½ vagy 1 ponttal). Ezen pontszámok összesítése alapján egyesített pontossági és fedési értéket állapít meg, és az ezekből kapott F-pontszám a hatékonyság végső mérőszáma.

A Szeged NER korpusz birtokában egy ad hoc módon választott train-test szétválasztással, és keresztértékeléssel kezdtük meg rendszerünk vizsgálatát és fejlesztését. Később azonban világossá vált, hogy ha eredményeinket össze kívánjuk mérni az egyetlen rendelkezésre álló alternatívával, akkor a Szarvas et al. [11] által használt bontást és train-devel-test metodológiát kell használnunk.

Szarvas György és kollégái a rendelkezésünkre bocsátották az általuk alkalmazott adatszétválasztást, és ettől a ponttól ezt az övékkel azonos metodológiát alkalmaztuk, sőt a teljes összehasonlíthatóság céljából az általuk választott train-devel szétválasztást is átvettük. Minthogy erőforrásainkat a fejlesztés kezdetekor véglegesítettük, és a metodológia-váltás a paraméter-hangolás korai szakaszában történt, ezért meggyőződéssel állítjuk, hogy rendszerünk nem „fertőződött" meg a tesztadatok ismeretével.

A fejlesztés folyamán a tesztadathalmazt és az azon vétett hibákat nem tekintettük meg. A rendszer nagyszámú paraméterének optimalizálásakor szigorúan a devel halmazon elért pontosság (F-pontszám) maximalizálása vezérelt.

5.2 Eredmények

Ismertetett rendszerünk F-pontszáma 96.35% a devel korpuszon, és 95.06% a test korpuszon. Ezek kis mértékben magasabb értékek a [11] által publikált 96.20% illetve 94.77% F-pontszámoknál. Megjegyezzük azonban, hogy a [11] rendszer optimalizálása angol és magyar nyelvre párhuzamosan történt, ezzel szemben mi kizárólag a Szeged NER Korpusz adatain dolgoztunk, és rendszerünk nyilvánvalóan továbbfejlesztést igényelne ahhoz, hogy más nyelvű adatokon is jó eredményt adjon. Ezek a továbbfejlesztések terveink között szerepelnek.

1. Táblázat.

NE-típus	Devel	Test	Szarvas et al. Devel	Szarvas et al. Test
LOC	92.06	96.36		95.07
MISC	93.58	85.12		85.96
ORG	97.62	96.20		95.84
PER	97.44	94.94		94.67
Össz.	**96.35**	**95.06**	**96.20**	**94.77**

6 Köszönetnyilvánítás

Köszönetet mondunk Szarvas Györgynek és Farkas Richárdnak adathalmazaik megosztásáért, és Halácsy Péternek a simítási algoritmus kivitelezéséért.

Bibliográfia

1. A. Borthwick.: A Maximum Entropy Approach to Named Entity Recognition. PhD thesis, New York University, 1999.
2. A. Borthwick, J. Sterling, E. Agichtein, R. Grishman: NYU: Description of the MENE Named Entity System as Used in MUC-7. Proceedings of MUC-7, 1998
3. Hai Leong Chieu, Hwee Tou Ng: Named Entity Recognition with a Maximum Entropy Approach. In: Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp. 160-163.

4. Dóra Csendes, János Csirik and Tibor Gyimóthy: The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus. Proceedings of TSD 2004, vol. 3206 (2004) 41-49.
5. P. Halácsy, A. Kornai, Cs. Oravecz, V. Trón, D. Varga: Using a morphological analyzer in high precision POS tagging of Hungarian. Proceedings of LREC 2006, pp. 2245—2248, 2006.
6. P. Halácsy, A. Kornai, D. Varga: Morfológiai egyértelműsítés maximum entrópia módszerrel Proc. 3rd Hungarian Computational Linguistics Conf., 2005. Szegedi Tudományegyetem.
7. S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel, and the Annotation Group (BBN Technologies): BBN: Description of the SIFT System as Used for MUC-7. Proceedings of MUC-7, 1998.
8. S. Sekine: Description of the Japanese NE System Used for MET-2. Proceedings of MUC-7, 1998.
9. S. Sekine, R. Grishman, and H. Shinou. A decision tree method for finding and classifying names in japanese texts. In Proceedings of the Sixth Workshop on Very Large Corpora, Montreal, Canada, 1998.
10. Gy. Szarvas, R. Farkas, L. Felföldi, A Kocsor, János Csirik: A highly accurate Named Entity corpus for Hungarian. Proceedings of International Conference on Language Resources and Evaluation, 2006.
11. Gy. Szarvas, R. Farkas and A. Kocsor: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In Proceedings of Discovery Science 2006, DS2006, LNAI 4265 pp. 267-278, Springer-Verlag 2006.
12. E. F. Tjong Kim Sang, and Fien De Meulder: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, Proceedings of CoNLL-2003, 2003. 13. http://www.cnts.ua.ac.be/conll2003/ner/
14. Trón, Gy. Gyepesi, P. Halácsy, A. Kornai, L. Németh, D. Varga: Hunmorph: open source word analysis. Proceeding of the ACL 2005 Workshop on Software, 2005.
15. Zhang Le.: Maximum Entropy Modeling Toolkit for Python and C++. http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html
16. G. Dong Zhou, Jian Su: Named Entity Recognition using an HMM-based Chunk Tagger. Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, pp. 473-480, July 2000
17. C. Zhu, R. H. Byrd, P. Lu, J. Nocedal: Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software (TOMS). 1997.

II. Morfológia

ReALIS projekt: a szóképzés általánosítása a számítógépes fordításban

Alberti Gábor[1], Kleiber Judit[1], Ohnmacht Magdolna[2],
Szilágyi Éva[1], Anne Tamm[3], Viszket Anita[1]

[1] Pécsi Tudományegyetem, Bölcsészettudományi Kar, Nyelvtudományi Tanszék,
7624 Pécs, Ifjúság útja 6.
[2] Szegedi Tudományegyetem, Bölcsészettudományi Kar, Nyelvtudományi Doktori Iskola,
6722 Szeged, Egyetem u. 2.
[3] Università degli Studi di Firenze, Dipartimento di Filologia moderna,
Via Santa Reparata 93-95, 50129 Firenze
gelexi@btk.pte.hu

Kivonat: A ReALIS projekt célja egy olyan nyelvelemző program megalkotása, amely minden eddiginél alaposabb szemantikai reprezentációt társít szövegekhez, és azt mint nyelvfüggetlen közvetítőnyelvet alkalmazva gépi fordító rendszerként is hasznosítható. Az elemző elméleti hátteréül a totálisan lexikalista GASG szolgál [3], a szemantikai reprezentációt pedig mint a ReALIS [1] implementációját képzeljük el. A totálisan lexikalista morfológia kiterjesztéseként az inflexiós morfémák mellett a produktív *derivációs morfémákhoz* is közvetlenül lexikai egységeket rendelünk, a szóalaktanilag „láthatatlan" *konverziós* [17] eseteket is figyelembe véve. A nyelvenkénti sokszínűség a nyelvfüggetlen szemantikai reprezentációs szinten eltűnik, így nem igényel bonyolultabb mechanizmust a nagyon különböző nyelvek közötti fordítás sem. A lexikalista elméletekre épülő elemzők és gépi fordító rendszerek egyre nagyobb térhódítása igazolni látszik, hogy szükség van a kifinomult nyelvelméleti eszköztárra a számítógépes nyelvészet olyan igényes területein, mint például a gépi fordítás. Projektünk alapvető célja a végsőkig vitt lexikalizmus kipróbálása ezen a területen.

1 Bevezetés

Mint a szerzők névsora mutatja, a *ReALIS* projekt kutatócsoportja a *GeLexi* (*Ge*nerative *Lexi*con) elméleti és számítógépes nyelvészeti kutatócsoport [3] és a *LiLe* (*Li*nguistic *Le*xicon) adatbázis-készítő team [6] egyesülése révén jött létre, további elméleti nyelvészek bevonásával. Alapvető célunk változatlanul a *kifinomult nyelvelméleti eszköztár* hasznosságának igazolása a számítógépes nyelvészet olyan igényes területein, mint amilyen például a *gépi fordítás* [4].

Azon az úton haladunk tovább, melyet a *totális lexikalizmus* elvének a morfémák szintjén való érvényesítése és egy korszerű DRT alapú szemantikához [16] való közvetlen (azaz nem egy hagyományos generatív szintaxison keresztül történő) hozzáférés fémjelez. A ReALIS (*Reciprocal And Lifelong Interpretation System*) [1] mint

háttér a DRT szemantikának egy olyan pragmatikai kiterjesztésére utal, amelynek alapján „interpretálói tudásbázist / információállapotot” építünk fel a parser támogatására.

Egy konkrétumot szeretnénk ebben a tanulmányban kiemelni. Az inflexiós morfémák mellett a produktív *derivációs morfémákhoz* is lexikai egységeket rendelünk, a szóalaktanilag „láthatatlan” *konverziós* [17] eseteket is figyelembe véve. A cikk második felében a lexikalista alapú gépi fordítás jelenlegi állásának áttekintése után projektünk megközelítését ismertetjük.

2 Elméleti háttér

A nyelvelméletben az utóbbi évtizedekben lexikalista fordulat következett be: a kezdeti szintaxis-központú elméleteket egyre inkább felváltják a szótári komponenst előtérbe helyező megközelítések. Ennek hatására egyre több lexikalista elmélet és számítógépes alkalmazás születik. Kutatócsoportunk alapvető célja megvizsgálni, hogy a végsőkig vitt (totális) lexikalizmus mint elméleti keret mennyire sikeres elméletileg, illetve intelligens nyelvtechnológiai alkalmazások fejlesztése céljára.

2.1 Totális lexikalizmus

A totális lexikalizmus azt jelenti, hogy a mondat összeépüléséhez szükséges minden információt a lexikonban tárolunk, így nincs szükség külön szintaktikai szabályrendszerre (frázisstruktúra-szabályokra). A nyelvtan tehát nem más, mint egy nagy adatbázis, amelyben a lexikai egységeket és azok tulajdonságait tároljuk, illetve egyetlen művelet – az *unifikáció* – mint a mondatok összeépülésének motorja.

Az a hipotézisünk, hogy egy ilyen homogén lexikalista nyelvtan mind elméletben, mind gyakorlatban jól működő, hatékony tud lenni; célunk ennek igazolása. Első lépésként kidolgoztuk a nyelvtant a magyar nyelv egy fragmentumára, majd azt implementáltuk Prolog programnyelven [3]. A parser bemenete egy magyar mondat, kimenete annak morfológiai és szintaktikai elemzése (szintaxison a függőségi viszonyokat értve), illetve a mondathoz rendelhető diskurzus-szemantikai reprezentáció.

Következő lépésként kidolgoztuk a nyelvtant az angol nyelv egy kis szeletére, és – kihasználva azt, hogy amit a Prolog elemezni tud, azt generálni is – kipróbáltuk a totálisan lexikalista megközelítést a gépi fordítás területén is [4]. Elképzelésünk az, hogy az univerzálisnak tartott szemantikai reprezentáción keresztül elemzőnk kétirányú használatával bármely két nyelv között megvalósítható a fordítás. Nem kell megírni minden nyelvpárra külön a fordító mechanizmust, csak rendelkeznünk kell az adott nyelvek nyelvtanának implementációjával.

Működő programunk bizonyította, hogy érdemes a totális lexikalizmus eszméjét a nyelvtechnológia területén alkalmazni. A következő lépés annak vizsgálata, hogy nagyobb adatbázison is működik-e a mechanizmus, illetve hogy bármely nyelvi jelenségről számot tud-e adni. Célunk egy pártízezer lexikai egységet tartalmazó relációs adatbázis és egy azt működtető program megalkotása, totálisan lexikalista alapokon. Szeretnénk továbbá minden eddiginél alaposabb gépi fordítást megvalósítani, amely nemcsak mondatok, hanem összefüggő szövegek fordítására is alkalmas, és számot tud adni többek között a retorikai viszonyról, a diskurzusfunkciókról (topik,

fókusz), a hangsúlyról és az aspektusról is. Ehhez egy minden eddiginél alaposabb szemantikai reprezentációt nyújtó elméletre van szükség.

2.2 ReALIS

A ReALIS (Reciprocal and Lifelong Interpretation System) egy reprezentacionalista, dinamikus szemantikai rendszer. Az interpretáció folyamatának az eddigieknél „reálisabb" szemléletét nyújtja. Négy komponensből áll: tartalmaz egy modellt a „külső" világról, a létező entitásokkal és a köztük lévő relációkkal, egy parciális függvényt, amely folyamatosan változó („felfrissülő") információállapotokat társít minden egyes interpretálóhoz, valamint egy-egy függvényt a dinamikus információállapot-változás és a statikus igazságértékelés formális kifejezésére.

Elsődlegesen magának az interpretálónak az információállapotát hivatott ábrázolni, és csak közvetve a feldolgozott diskurzusokét – ezért mondjuk a ReALIS interpretációs megközelítését *élethossziglaninak* ('lifelong'). A rendszer karakterisztikus tulajdonsága a kölcsönös ('reciprok') információábrázolási technika: az interpretáló egy interpretálási folyamat során nemcsak az adott témáról meglévő saját tudására építhet, hanem a másokról feltételezhető tudásra is. Ez olyan konkrét nyelvi tények magyarázatában is szerepet játszhat, mint a névelők, névmások és más anaforikus elemek használata adott mondatbeli helyzetekben.

Előnye más szemantikai rendszerekhez képest, hogy pragmatikai tényezőket is figyelembe vesz, illetve hogy nagyon alapos szemantikai reprezentációt társít nem csupán egy-egy mondathoz, hanem szövegekhez is. Ezek a tulajdonságai teszik alkalmassá intelligens nyelvtechnológiai alkalmazásokra, mint például a gépi fordítás.

3 Szóképzés

A totálisan lexikalista morfológia kiterjesztéseként az inflexiós morfémák mellett a produktív *derivációs morfémákhoz* is közvetlenül lexikai egységeket rendelünk – Alberti (2006) [2] szellemében, a szóalaktanilag „láthatatlan" *konverziós* [17] eseteket is figyelembe véve. Mi több, az alábbi fordítási problémák a „szóképzés" fogalmának még merészebb általánosítását látszanak szükségessé tenni, amennyiben szeretnénk megőrizni egy olyan kézenfekvő fordítási megfeleltetést (legalábbis kiindulópontként), mely szerint a forrásnyelv t_f relatív tövéből létrehozott $K_f(t_f)$ képzett alak célnyelvi megfelelője egy $K_c(t_c)$ képzett alak lesz, ahol a t_c tő a t_f tő megfeleltetettje, a K_c képzési „függvény" pedig a K_f képzésé.

Nézzünk néhány példát a nyelvekben található változatos eszközökre, ahogy a különféle „képzéseket" megvalósítják!

Míg az angol a *progresszív* alakot hagyományos értelemben vett igenévképzéssel hozza létre (ld. (1a): t_f: *set up* ⇒ $K_f(t_f)$: ***sett**ing up*), addig a magyarban úgy foghatjuk fel, hogy a képzett igealak máshova rendeli az igekötőjét ((1b): t_c: ***fel**állítottuk* ⇒ $K_c(t_c)$: *állítottuk **fel***). Megint másként jár el az észt nyelv [23]: a tárgy esetét módosítja ((1c): a $t_{c'}$ argumentumszerkezetében: *Accusativus* ⇒ a $K_{c'}(t_{c'})$ argumentumszerkezetében: *Partitivus*).

(1) a. We set up the tent. / We were setting up the tent.
b. Felállítottuk a sátrat. / Állítottuk (éppen) fel a sátrat (mikor...).
c. Panime telgi püsti / Panime telki püsti.
put.past.1pl tent.acc up / put.past.1pl tent.part up

Az utóbbi két esetben tehát maga az igei szótest nem módosul (*konverzió* [17]), *általánosított képzésként* foghatjuk azonban fel a vonzatszerkezet-módosulást [2]. Az (1) példasorban bemutatott szófajváltást megvalósító K_c (igenév-) képzésnek tehát a két tekintett célnyelvben a finit igei alakot megőrző operáció feleltethető meg. Az észtbeli $K_{c'}$ operációt olyan argumentumszerkezet-módosító igeképzésnek tekinthetjük, mint a magyarban az *átúszik a folyón → átússza a folyót* példa által szemléltethető változást: más lexikai egység szerepel a kimenetben, mint a bemenetben, hiszen már esetkeret jelölődik ki. A magyarbeli K_c operációt abban az értelemben tekinthetjük képzésnek, hogy a bemeneti alakban prefixummá inkorporálódó igekötői argumentum a kimeneti alakban szóvá önállósul [2]; a különbséget tulajdoníthatjuk eltérő lexikai leírásoknak (ahogy az *eszik egy almát → almát eszik*, *alakult egy kórus → kórus alakult* példapárok esetében is eltérő lexikai tétel megadásával szeretnénk számot adni a bemenet és a kimenet különbségéről, ami az egyik argumentum inkorporált helyzetét illeti).

Ugyanígy, nyelvenként különbözően valósulhat meg a passzív progresszív alakok előállítása (lásd a (2) példasorban: K_f igenévképzés, K_c új igetétel létrehozása egy argumentum inkorporált helyzetének megszüntetése révén, míg $K_{c'}$ új igetétel létrehozása egy argumentum esetének módosítása révén [23]), illetve az aktív passzív képzés (a (3) pontban):

(2) a. The owls were checked. / The owls were being checked.
b. Átvizsgálták a baglyokat. / Vizsgálták (éppen) át a baglyokat (mikor...).
c. Vaadati öökullid üle. / Vaadati öökulle üle.
look.impers owl.nompl over / look.impers owl.partpl over
(3) a. Harry kissed his mother. / Harry was kissed by his mother.
b. Harry megcsókolta az anyját. / Harry-t megcsókolta az anyja.
c. Harri suudles oma ema. / Harrit suudles ta ema.
H.nom kiss.3sg.past his-own mother.part H.part kiss.3sgpast his mother.nom

Ez utóbbi esetében (3) az angol ismét igenévképzéshez folyamodik, míg a magyar és az észt megőrzi az időjeles igeformát, sőt az esetkeretet is, továbbá argumentuminkorporációval kapcsolatos változás sem történik. Ami változik, az az argumentumok szórendi helyzete. Amennyiben úgy tekintjük, hogy a magyarban és az észtben nem *alulspecifikált* az igei régens ellenőrzése alatt tartott argumentumok sorrendje, hanem különféle sorrendek különféle lexikai tételekhez tartoznak, akkor még a (3b-c) pontokban bemutatott „topikalizációt" is tekinthetjük általánosított képzésnek.

Visszatérve az angol *–ing* igenévképzésre, annak egyes esetekben akár a magyarban is igenévképzés feleltethető meg (pl. *runn**ing*** ~ *fut**ó***), máskor azonban (igekötő híján) formai szempontból „identikus függvénynek" kell tekintenünk a képzést ((4): [*ran ~ futott*] [*was runn**ing*** *~ futott*]), vagy egy mátrixigének kell megfeleltetnünk ((5): [*dim ~ ostoba*] [*be**ing*** *dim ~ ostobán* ***viselkedett***].

(4) a. He ran. / He was running.
b. Futott. / Futott (éppen, mikor...).
c. Ta jooksis. / Ta jooksis.
s/he run.3sgpast / s/he run.3sgpast

(5) a. Harry thought Ogden was (being) extremely dim.
b. Harry úgy gondolta, hogy Ogden rendkívül ostoba (... ostobán viselkedik).
c. Harri arvas, et Ogden [on erakordselt tobe]. / [käitub erakordselt tobedalt].
H.nom think.3sgpast that O.nom [be.3sg extremely stupid] / [behave.3sg exteremely stupidly]

Ez utóbbi (5) a szerint a gyakori fordítási megfelelés szerint működik, amit az angol és a magyar viszonylatában a másik irányban tapasztalunk gyakrabban: amikor is magyarbeli szuffixálással megvalósuló képzésnek (pl. K_f: *-(t)At*, *-hAt*, *-Ul*, *-ít, -(V)ll*), mely igei, illetve melléknévi töveken működik, egy mátrixigei régens bevonása feleltethető meg, mely önálló nem finit igealakot, illetve melléknevet szelektál (rendre: *dolgoztat ~ make sy work*, *dolgozhat ~ can/may work*, *barnul ~ become/grow brown*, *barnít ~ make sg brown*, *drágáll ~ consider sg expensive*).

Totálisan lexikalista morfoszintaktikai megközelítésünk kiegészítve a képzés felfogásának fenti erőteljes általánosításával lehetővé teszi tehát, hogy a fordítás azon eseteit is képesek legyünk *szabályalapúan* (és kompozicionális jelentéskalkulációval társítva) kezelni, ahol elvész a szófaji megfelelés, sőt még a szó~szó megfelelés is két nyelv között a lexikai egységek mondattá szerveződésének folyamatában.

4 Gépi fordítás

A gépi fordító rendszereket két nagy csoportra oszthatjuk: szabályalapú [15] és mintaalapú [22] megközelítésekre. A kezdetekre szabályalapú rendszerek fejlesztése volt a jellemző, míg az utóbbi években inkább hatalmas korpuszok és fordítómemóriák segítségével próbálják létrehozni a minél kisebb emberi beavatkozást igénylő gépi fordító programokat.

A szabályalapú rendszereken belül többféle megközelítést alkalmazhatnak. A fordítás történhet (1) direkt módon, kis elemzéssel, (2) közvetítőnyelven keresztül, illetve (3) transzferrel, amikor a forrásnyelvi mondatból egy absztrakt forrásnyelv-közeli reprezentáción, majd egy célnyelv-közeli reprezentáción keresztül állítják elő a mondat célnyelvi megfelelőjét.

A fordításhoz két lépésben jutnak el. Az első lépés az analízis (elemzés), amely a forrásnyelvi szöveg szintaktikai elemzésén túl tartalmazhat morfológiai, és valamilyen mértékű szemantikai elemzést is – ez utóbbira a többértelműségek kezeléséhez van (minimálisan) szükség. A második lépés pedig a szintézis (generálás), amikor a mondat célnyelvi megfelelőjét állítják elő lehetőleg ugyanazzal a mechanizmussal. Így a modularitás és a megfordíthatóság fontos tulajdonságaik ezeknek a rendszereknek.

Miután a szabályalapú megközelítés nem bizonyult kellően hatékonynak, még statisztikai módszerek bevonásával sem, továbbá rengeteg befektetett munkát igényelt a nyelvtaníró részéről, a 90-es évektől kezdtek tért hódítani a mintaalapú fordítórendszerek. Hatalmas korpuszok születtek, amelyeket minél alaposabban annotáltak, annál jobb minőségű elemzőket lehetett rájuk építeni, viszont a ráfordítási idő is jelentősen megnőtt. Közülük leghatékonyabbnak a treebankok (mondatszerkezettel is annotált korpuszok) bizonyultak, hiszen azok tartalmazzák a lehető legtöbb információt; viszont előállításuk nem egyszerű feladat, sok munkaórát igényel. Éppen ezért napjainkban egyre több olyan rendszert fejlesztenek, amely annotálatlan korpuszból képes „megtanulni” a nyelv elemeit és azok tulajdonságait különféle módszerek (analógia, disztribúciós módszer) segítségével, így kevés ráfordítással készíthetnek haté-

kony elemzőket. Nagyon elterjedt továbbá gépi fordításkor a párhuzamos korpuszok, illetve különféle fordítómemóriák használata, melyeket például az interneten fellelhető több nyelven elérhető anyagok szinkronizálásával állítanak elő.

A korpuszalapú megközelítések sem hoztak teljes sikert. Nem minden jól formált kifejezés található meg bennük, különösen a gazdag morfológiájú nyelvek esetében, mint a magyar, és nagyon ritka az igazán alaposan és jól annotált anyag, vagy megfelelően illesztett párhuzamos korpusz. Emiatt születnek különféle hibrid megoldások: alapvetően korpuszra támaszkodó, de valamilyen szintű elemzést is végző, vagy szabályalapú, de a többértelműségek kezelésében korpuszt is használó rendszerek.

Napjainkban egyre többen kezdenek visszatérni a szabályalapú megközelítéshez, de nem sekélyelemzést használnak, ami csak részleges elemzést ad, és hiába robusztus, ha nem annyira precíz [12], így komplexebb célokra (pl. igazán jó minőségű fordítás) nem alkalmas. A mélyelemzést végző rendszerek sokkal alaposabbak és pontosabbak, és az utóbbi időben lefedettségben is felveszik a versenyt a sekélyelemző rendszerekkel. Az ilyen nyelvészeti alapú, kézzel írt nyelvelemzők készítéséhez szükséges kezdeti nagyobb energiabefektetés pedig megtérül később, például amikor új nyelvekre dolgozzák ki [10]. Az igazán jól működő és hatékony mélyelemző rendszerek unifikációs mechanizmusokat használnak, így előnyeik között szerepel, hogy egyszerűbb szabályokat lehetséges megfogalmazni, gyorsabb, egyszintű, lexikalista és megfordítható [15]. A többértelműség kezelésére ezek a rendszerek is használhatnak korpuszt, mivel sokszor szabályba nem foglalható tényezők szükségesek az egyértelműsítéshez (pl. kontextus), illetve optimalitás-jelölő rangokat, hogy a ritka alakzatokat is elfogadja az elemző, de alapesetben a gyakoribb elemzés mellett döntsön [10].

A legtöbb gépi fordító rendszert egy adott nyelvpárra dolgozzák ki, sőt, akár csak az egyik irányban működnek, de léteznek többnyelvű rendszerek is, ahol a mechanizmus univerzalitására törekszenek.

4.1 Létező rendszerek

Egyetértés mutatkozik abban, hogy ha a cél egy igazán intelligens gépi fordító rendszer létrehozása, nem elégségesek a sekélyelemzést végző programok. A mintaalapú és statisztikai megközelítések ellen pedig az szól, hogy a legtöbb nyelvre nem létezik igazán jól használható párhuzamos korpusz [7]. További érv a szabályalapú gépi fordítás mellett az újrahasznosíthatóság: hogy a gépi fordítás fejlesztésének eredményei más nyelvtechnológiai alkalmazásokban is használhatók, illetve más területek eredményeit a gépi fordítás is hasznosíthatja. Ezért ha olyan gépi fordító rendszer fejlesztése a cél, ami bármely két nyelv esetében[15] jó minőségű és pontos fordítást ad, akkor mélyelemzést végző szabályalapú rendszer alkalmazása tűnik a legjobbnak.

Napjainkban erre a célra az unifikációs mechanizmusokat használó lexikalista elméletek látszanak a legmegfelelőbbnek, amik nemcsak a konfigurációs nyelveket (mint az angol) kezelik hatékonyan, hanem a magyarhoz hasonló szabadabb szórendű nyelveket is. Továbbá a jó minőségű fordításhoz szükségesnek látszó valamiféle

[15] Rokon, vagy nagyon hasonló nyelvek közötti fordításkor elég lehet a sekélyelemzés, ha nagyfokú a morfológiai és szintaktikai hasonlóság. Például csehről szlovákra lehet szinte szóról szóra fordítani [14].

szemantikai reprezentáció hozzárendelésében is jobb eredményeket érnek el, mint a frázisstruktúra-nyelvtanok.

A lexikalista nyelvtanok közül a legeredményesebbek LFG vagy HPSG formalizmust használnak. Számos nyelvre léteznek már nagy lefedettségű elemzőik, amik nagyon jó minőségű és alapos elemzést adnak, továbbá szemantikai reprezentációt is társítanak a mondatokhoz. Különböző nyelvekre alkalmazzák ugyanazt az elméleti keretet, így tudják elérni, hogy a különböző nyelvű elemzések szinte teljesen párhuzamosak legyenek, ami jelentősen megkönnyíti az elemzőkre épülő gépi fordító rendszerek fejlesztését.

Ilyen nyelvtan például az ERG (English Resource Grammar), ami a legnagyobb HPSG-alapú nyelvtan angol nyelvre, implementálva az LKB (Linguistic Knowledge Building) rendszerben [9]. HPSG-re kifejlesztettek egy keretrendszert is (Grammar Matrix [5]), ami nem egy nyelvtan, hanem nyelvtanok feletti általánosítások gyűjteménye. Négy függetlenül fejlesztett HPSG-alapú nyelvtanból indult ki: angol, japán, német és norvég. A cél egy egységes nyelvtan megalkotása az egyes nyelvtanok feletti általánosítások alapján, hogy további megszorítások bevezetésével még egységesebb elemzések születhessenek, és hogy újabb nyelvek nyelvtanainak megírását gyorsabban lehessen elkezdeni. A Grammar Mátrixszal több kompatibilis nyelvtant is fejlesztettek. Ilyen például a JACY [21], ami egy nagy lefedettségű, nagy pontosságú japán nyelvtan. HPSG-alapú, MRS-t (Minimal Recursion Semantics[16]) használ a szemantikai reprezentáció hozzárendeléséhez. Eredetileg beszéltnyelvi dialógusok gépi fordítására fejlesztették, később automatikus email-megválaszolásra és egyéb (többnyelvű) nyelvtechnológiai feladatokra is használták.

Az LFG-t mint elméleti keretet használó elemzők közül érdemes megemlíteni a Parallel Grammar (ParGram [8]) projektet. Lényege, hogy a hasonló szerkezetek elemzései a különféle nyelvekben amennyire csak lehet, párhuzamosak. Így hasonló számítógépes alkalmazásokban használhatók, és a gépi fordítás is egyszerűsíthető. Eredeti célja az LFG-formalizmust tesztelni: univerzalitását, lefedettségének határait, és hogy mennyire tartható a párhuzamosság a különféle nyelvek között. A párhuzamosságot az f-struktúra (funkcionális szerkezet) szintjén érik el[17]. Az eredmények bíztatóak, nagyfokú párhuzamosság érhető el, így új nyelvtanok építése is könnyebben, gyorsabban lehetséges. Az elemzőt szándékosan nagyon különböző nyelvekre dolgozták ki először: angolra, németre, japánra, urdura és norvégra. Némelyik nagy lefedettségű, ipari alkalmazás, némelyiknél az s-struktúrát (szemantikát) is kidolgozták. Az elméletet az XLE (Xerox Linguistic Environment) platformon implementálták.

Jónéhány a különféle lexikalista alapú elemzők közül tehát elérte már azt a lefedettséget, amit korábban csak nagy korpuszok vagy statisztikai módszerek alkalmazásával lehetett megvalósítani. Ezek az elemzők mélyelemzést végeznek, kimenetükben szemantikai (vagy szemantika-közeli) reprezentáció is szerepel, továbbá nagyfokú párhuzamosságot képesek megvalósítani akár nagyon különböző nyelvek között is. Mindezek lehetővé teszik, hogy jó minőségű gépi fordító rendszerek épüljenek

[16] Lapos szemantikai reprezentációt rendel szavakhoz és frázisokhoz. Mélyebb szintű, mint csupán a predikátum-argumentum viszonyok. Rendezetlen struktúra, hatókör tekintetében alulspecifikált.

[17] Néha szándékosan nem párhuzamosak az f-struktúrák, pl. a névszói állítmány esetében (az *It is red* mondat esetében az angolban a kopula a fej, a japánban a melléknév), vagy amikor egy nyelvben megvan egy bizonyos jegy (pl. a németben a nem), egy másikban pedig nincs.

rájuk. Legnagyobb részük még kísérleti fázisban tart, de eredményeik nagyon ígéretesek. Több nyelvre is kidolgoztak már működő gépi fordító rendszereket, amelyek ezeket az elemzőket (és generálókat) használják. Legtöbbjük transzfer-alapú, így minden nyelvpárra és mindkét fordítási irányra külön transzfer-komponenst kell kidolgozni. A magyarra ilyen rendszer eddig még nem készült[18].

LFG formalizmust használ például az XTE (Xerox Translation Environment), ami a ParGram projekt fordítórendszere [11]. A nyelvfüggetlennek tartott f-struktúra teszi alkalmassá többnyelvű nyelvtechnológiai alkalmazásokra, mint a gépi fordítás. A hatékony elemzőt és generálót tartalmazó XLE platformot kiterjesztették egy transzfer-komponenssel. A transzfer az f-struktúrán keresztül történik, amiről elismerik, hogy néha nem elég univerzális (pl. a fejváltó fordítás esetében), de a nyelvek közötti különbségek legnagyobb része már ezen a szinten eltűnik. A legjobb az s-struktúrán keresztüli fordítás lenne, viszont ekkor ki kellene dolgozni a szintaxis-szemantika interfészt mind az elemzés, mind a generálás oldaláról (azóta ez irányban folytak a kutatások). A többértelműséget a döntés végsőkig való halasztásával kezeli. HPSG-formalizmust használ például a DELPH-IN (Deep Linguistic Processing with HPSG Initiative [7]), ami egy nyílt forráskódú, szemantikai transzfer-alapú gépi fordító rendszer. Létező forrásokat alkalmaz: elemzőket, generálókat, kétirányú nyelvtanokat és transzfer-motort. Kezdetképpen japánról angolra fejlesztették ki. A cél a mechanizmus működésének megmutatása, ezért egyelőre száznál kevesebb transzfer-szabályt, és csupán párezres transzfer-lexikont tartalmaz. A rendszer a forrásnyelvet (japán) elemzi egy szabályalapú forrásnyelvi nyelvtannal (JACY), a legjobb elemzést valószínűségi rangok alapján kiválasztja; kimenete egy precíz, alulspecifikált (nyelvspecifikus) forrásnyelvi szemantikai reprezentáció (MRS_S). Ebből a transzfer-motor előállítja az alulspecifikált (nyelvspecifikus) célnyelvi (angol) szemantikai reprezentációt (MRS_T) újraíró szabályokkal. Ha több megoldás van a szabályalkalmazásnál, több fordítás lesz. Végül a célnyelvi generátor (ERG) angol nyelvű szöveget készít belőle. További céljaik között a lefedettség növelése szerepel, valamint választ kapni néhány elméleti kérdésre: mennyire lehet egységes a szemantikai reprezentáció, lehet-e (kvázi)automatikusan növelni a nyelvtanokat és lexikonokat, mi lehet a lexikális szemantika szerepe, stb. A fejlesztés egyelőre kísérleti stádiumban van, azonban nagyon ígéretes.

Létezik olyan rendszer is, amely LFG és HPSG alapú elemzőket is használ, a norvég-angol gépi fordítást megvalósító szemantikai transzfer-alapú LOGON [18]. Olyan célokra kezdték fejleszteni, ahol a fordítás minősége fontosabb a nagy lefedettségnél. A fordítás három lépésben történik: első a norvég mondat LFG-alapú grammatikai és szemantikai mélyelemzése, amihez az XLE platformon fejlesztett NorGram-ot használják (amit a ParGram projekten belül fejlesztettek). Kimenete egy nyelvspecifikus logikai szemantikai reprezentáció (MRS). Majd ezeknek a reprezentációknak a transzfere történik nyelvspecifikus angol reprezentációkba (MRS). Végül a szemantikai reprezentációból angol nyelvű mondatot generálnak HPSG alapú elemző (generáló) segítségével (célnyelvtan: ERG, generátor: LKB). A projekt hosszú távú, bár célja egyelőre csak egy demonstráció fejlesztése, amely megmutatja, hogy a mechanizmus működik. A többértelműség kezelésére valószínűségi rangokat használ.

[18] A magyarra működő gépi fordító rendszerek vagy mintaalapúak (Hunlish [13]), vagy alapvetően szabályalapúak (MetaMorpho [19]), de nem lexikalisták. Nem céljuk továbbá olyan minőségű fordítás, mint a cikkben tárgyalt egyéb alkalmazásoknak.

Fontos jellemzője a modularitás, így mindig a legfrissebb elemzőt rakhatják a rendszer mögé.

4.2 Javaslatunk

A ReALIS projekt a végsőkig viszi a lexikalista megközelítést. Azt próbálja igazolni, hogy nincs szükség frázis-struktúrára: a mondatok összeépüléséhez elégséges a lexikai egységek gazdag jegystruktúrája és egyetlen művelet, az unifikáció.

Schneider [21] is amellett érvel, hogy nem feltétlenül kell konfigurációs felszíni reprezentációs szint a nyelv szintaktikai elemzéséhez. Az LFG például főleg azért használ c-struktúrát (összetevős szerkezetet), mert az környezetfüggetlen. Ha csak a függőségi viszonyokat mutató f-struktúrát használná, az nem lenne hatékony, mivel környezetfüggő. Akárcsak a Függőségi Nyelvtan, ahol (eredetileg) szintén csak funkcionális szint van, nem használ összetevős szerkezetet. A szórendnek nincs elsődleges szerepe (néha segít az egyértelműsítésben), de valójában nincsenek megszorítások arra nézve, hogy egy fej hol keresse a bővítményeit. Schneider [21] megmutatja, hogyan lehet az f-struktúrát környezetfüggetlenül előállítani, így nem lenne szükség a teljes c-struktúrára, csak a fontosabb elemeire.

A totálisan lexikalista megközelítés a végsőkig viszi ezt a javaslatot: kipróbálja, lehetséges-e semmivé redukálni a c-struktúrát. Készíthető-e hatékony elemző (és gépi fordító) rendszer akkor, ha nem támaszkodunk a környezetfüggetlen c-struktúrára, vagy egyéb konfigurációs felszíni reprezentációs szintre. Véleményünk szerint a válasz igen: a megközelítés hatékony, mert a függőségi nyelvtanokkal szemben nálunk vannak megszorítások a szórendre, az ún. rangparaméterek formájában [3]. Ezekkel a paraméterekkel elegánsan megragadható az is, hogy egy nyelven belül milyen szórendi variánsok lehetségesek, és számot ad a nyelvek közötti szórendi különbségekről is.

Célunk tehát egy olyan (szabályalapú) elemző és gépi fordító rendszer fejlesztése, amely totálisan lexikalista, nyelvfüggetlen, és minden eddiginél alaposabb szemantikai reprezentációt társít nem csupán mondatokhoz, hanem szövegekhez. Programunk a 3. pontban bemutatott nyelvenkénti sokszínűséget ezt a (ReALIS alapú) szemantikai reprezentációt mint közvetítőnyelvet alkalmazva hidalja át, hiszen ez nyelvfüggetlen közös alapot jelent, amihez képest aztán a felszíni megjelenítés az egynyelvű komponensek feladata. Így a fordításhoz alaposan kidolgozott elemzőkre van csupán szükség, melyek legfontosabb része a szemantikai komponens; hiszen programunk ugyanazt a mechanizmust használná bármely két nyelv esetében, és a fordítási iránytól függetlenül.

Morfémaszintű totális lexikalizmusunk pedig oly módon fejleszti tovább a grammatika *egyszintűségének* a gondolatát, hogy a kategoriális többértelműség alább szemléltetett potenciális túlburjánzása kikerülhetővé válik, hiszen az adott példában egy moduláris rendszer szintaxisának 72 megemlített lehetséges bemenete helyett pusztán azt a három elemzési utat kell bejárnunk, amit az időjeles igeként elemezhető három egység indít el.

(6)

Dobom	az	ír	szánom.
N+poss1sg+Nom/N+poss1sg+Acc/V+1sg	Pron/D	V/A/N_1/N_2	N+poss1sg+Nom/N+poss1sg+Acc/V+1sg
3	·	2 · 4	· 3 = 72

A magyar nyelvre nem készült még olyan alapos elemző és gépi fordító program, mint amilyen a ReALIS projekt céljai között szerepel. Olyan jelenségekről is számot kívánunk adni, mint a retorikai viszonyok (hogyan kapcsolódnak a mondatok egymáshoz), a különféle diskurzus-funkciók (mint a topik és a fókusz), vagy a hangsúly és az aspektus. Mindehhez egy minden eddiginél alaposabban kidolgozott formális szemantikai eszköztár áll rendelkezésünkre.

5 Összegzés

Az elmúlt években a GeLexi projekt egy magyar és angol mondatokat elemző Prolog programot fejlesztett, hogy igazolja a totális lexikalizmus eszméjének a gyakorlatban való alkalmazhatóságát [3]. Megmutattuk, hogy az elemzőnk kétirányú használatával a gépi fordítás is lehetséges, a diskurzus-szemantikai reprezentációt mint közvetítőnyelvet alkalmazva [4]. Eközben a LiLe projekt egy relációs adatbázis létrehozásába fogott, egyéb célok mellett azért, hogy nagyobb adatbázison vizsgálhassuk a totálisan lexikalista megközelítés hatékonyságát [6].

A ReALIS projekt keretében a két kutatócsoport egyesült, és további nyelvészek bevonásával azt a célt tűzte maga elé, hogy minden eddiginél pontosabb gépi fordítást valósítson meg, lényegesen nagyobb adatbázison. Rendszerünk szabályalapú, mélynyelvészeti elemzést végez, a totálisan lexikalista GASG (Generatív Argumentumstruktúra Nyelvtan) alapján, amely nem rendel összetevős struktúrát a mondatokhoz, csupán a lexikai egységek tulajdonságaira támaszkodik az elemzés során, a szórendről pedig rangparaméterek segítségével ad számot. A mondatokhoz rendelt diskurzus-szemantikai reprezentáció, amely a ReALIS [1] implementációján alapul, az eddigi (LDRT-alapú) reprezentációnál jóval részletesebb, olyan jelenségekről is számot tud adni, mint a retorikai viszonyok, a diskurzusfunkciók vagy az aspektus. A reprezentáció nyelvfüggetlen, így a gépi fordítás során nem okoz problémát a nyelvek sokfélesége: bizonyos lexikai egységek az egyik nyelvben szavak, a másikban szuffixumok, a topikalizációt az egyik nyelv pusztán a szórenddel, a másik képzők hozzáadásával fejezi ki, vagy a progresszív aspektust az egyik nyelvben segédige jelzi, a másikban pedig egy vonzat esetének módosítása.

Az utóbbi években a mélynyelvészeti, lexikalista, unifikációra épülő elemzők és gépi fordító rendszerek egyre nagyobb teret hódítanak. Már nem csupán pontosabbak, mint a mintaalapú vagy sekélyelemzést végző rendszerek, hanem lefedettségben is kezdik felvenni a versenyt velük. Ezeknek a rendszereknek a sikere még inkább arra ösztönöz minket, hogy folytassuk a totálisan lexikalista megközelítés nyelvtechnológiai alkalmazhatóságának vizsgálatát, és létrehozzunk egy minden eddiginél pontosabb gépi fordító rendszert.

Bibliográfia

1. Alberti, G.: ReAL Interpretation System. L. Hunyadi – Gy. Rákosi – E. Tóth eds.: Preliminary Papers of the Eighth Symposium on Logic and Language, University of Debrecen (2004) 1–12
2. Alberti, G.: Changes in Argument Structure in the course of Derivation in Hungarian. Acta Linguistica Hungarica (2006)

3. Alberti, G., Kleiber, J.,Viszket, A.: GeLexi project: Sentence Parsing Based on a GEnerative LEXIcon. Acta Cybernetica 16 (2004) 587–600
4. Alberti G., Kleiber J., Viszket A.: GeLexi projekt: Fordítás totálisan lexikalista alapokon. In: Alexin Z. – Csendes D. (szerk.): II. Magyar Számítógépes Nyelvészeti Konferencia, Juhász Nyomda, Szeged (2004) 73-80
5. Bender, E. M., Flickinger, D., Oepen, S.: The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In Proc. COLING 2002 Workshop on Grammar Engineering and Evaluation. Taipei (2002)
6. Bódis Z., Kleiber J., Szilágyi É., Viszket A.: LiLe projekt: Adatbázis mint „dinamikus korpusz". In: Alexin Z. – Csendes D. (szerk.): II. Magyar Számítógépes Nyelvészeti Konferencia, Juhász Nyomda, Szeged (2004) 11–18
7. Bond, F., Oepen, S., Siegel, M., Copestake, A., Flickinger, D.: Open source machine translation with DELPH-IN. In Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit, Phuket, Thailand (2005) 15-22
8. Butt, M., Dyvik, H., King, T. H., Masuichi, H., Rohrer, C.: The Parallel Grammar project. In Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan (2002)
9. Copestake, A.: Implementing Typed Feature Structure Grammars. Stanford, CA: CSLI Publications (2002)
10. Forst, M., Kuhn, J., Rohrer, C.: Corpus-Based Learning of OT Constraint Rankings for Large-Scale LFG Grammars. In: Butt, M. – King, T. H. (eds.): Proceedings of the LFG'05 Conference, University of Bergen, CSLI Publications (2005) 154-165
11. Frank, A.: From parallel grammar development towards machine translation. In Proceeding of MT Summit VII (1999) 134–142
12. Frank, A.: Projecting LFG F-structures from Chunks --- or (Non-)Configurationality from a different Viewpoint. In: Butt, M. – King, T. H. (eds.): The Proceedings of the LFG'03 Conference, University at Albany, State University of New York, CSLI Publications (2003) 217-237
13. Halácsy P., Kornai A., Németh L., Sass B., Varga D., Váradi T., Vonyó A.: A hunglish korpusz és szótár. In: Alexin Z. – Csendes D. (szerk.): III. Magyar Számítógépes Nyelvészeti Konferencia, Juhász Nyomda, Szeged (2005) 134-143
14. Homola, P., Kuboň, V.: A Translation Model for Languages of Acceding Countries. In: John Hutchins (ed.): Broadening Horizons of Machine Translation and its Applications. Proceedings of the Ninth EAMT workshop. Foundation for International Studies, University of Malta, Valletta (2004) 90-97

Tisztán statisztikai alapú szófaji címkéző használata a Szeged Korpuszon

Kiss Géza, Németh Géza

Budapesti Műszaki Egyetem, Távközlési és Médiainformatikai Tanszék
{kgeza, nemeth}tmit.bme.hu

Kivonat: A cikkben bemutatott munka egy, eredetileg nyelvazonosításra kidolgozott, tisztán statisztikai alapú (morfológiai analízist igénybe nem vevő), automatikus gépi tanuláson alapuló címkéző eljárás alkalmazása szófaji címkézésre a Szeged Korpusz 2.0-ra. A megközelítés magyar nyelvre önmagában nem ad olyan pontos eredményt, mintha morfológiai elemzőt is használnánk, épp a tisztán statisztikai megközelítés miatt. Előnye, hogy nem látott szavak címkéjére is használhatóan jó becslést ad, viszonylag kis számításigényű, valamint tisztán statisztikai jellege miatt megfelelő tanítóhalmaz birtokában egy ismeretlen nyelvre is egyes gyakorlati alkalmazásokhoz elegendő pontossággal képes a szófaji címkézés feladatát ellátni. Ezek miatt más módszerek kiegészítőjeként is alkalmazható, pl. az ismeretlen szavak szófajának becslése céljából.

1 Bevezetés

A cikkben bemutatott munka egy tisztán statisztikai alapú, morfológiai analízist igénybe nem vevő, automatikus gépi tanuláson alapuló címkéző eljárás, amely egy korábban nyelvazonosítás céljára alkalmazott módszerünk [2] alkalmazása szófaji címkézésre a Szeged Korpusz 2.0-ra [1].

Mint látni fogjuk, ez a megközelítés magyar nyelvre önmagában nem ad olyan pontos eredményt, mintha morfológiai elemzőt is használnánk, épp a tisztán statisztikai megközelítés miatt. Viszont vannak olyan előnyei, amelyek akár más módszerek kiegészítőjeként való alkalmazásra is érdemessé teszik.

Egyrészt a tisztán statisztikai megközelítés miatt megfelelő tanítóhalmaz birtokában egy ismeretlen nyelvre is elegendő pontossággal képes a szófaji címkézés feladatát ellátni. Másrészt ezt viszonylag kis memória- és számítási igénnyel teszi. Példának okáért szöveg-beszéd átalakítás során is fontos információ a szintetizálandó szöveg szavainak szófaja, viszont valós idejű alkalmazásában ezt az információt a lehető legkisebb erőforrás igényű eszköz használatával szeretnénk megkapni.

Morfológiai elemzők alkalmazásakor is szükségszerűen találkozunk ismeretlen (out of vocabulary, OOV) szavakkal, amelyek szófajának megállapításához szükséges kiegészítő megoldáshoz folyamodni. Az itt bemutatott módszer ezt a problémát is kezeli. A módszer másik összetevője a szó környezetének figyelembevételével a legvalószínűbb nyelvi címkesor közelítése.

A munka során a korpuszt XML formából egy egyszerűsített, csak a megoldás szempontjából releváns információkat tartalmazó formára konvertáltuk, közben konzisztencia ellenőrzést is végezve. Megvizsgáltuk a korpuszban egy egységnek címkézett kifejezéseket is, és úgy módosítottuk a saját címkéző algoritmusunkat, hogy a detektált kifejezések minél inkább egyezzenek a korpuszban található felbontással, bár a szóközt is tartalmazó egységek miatt ezt nem tudtuk teljesen megvalósítani. Utóbbi probléma kezelésére készítettünk egy olyan szövegváltozatot, amely minden szót (szóközökkel szeparált részt) külön címkével címkéz; ehhez a több szóból álló kifejezések minden szava külön-külön a teljes kifejezés címkéjét kapta. Ezt használtuk a betanítás során, az eredmények ellenőrzéséhez ezt és az eredeti formát is használtuk; az utóbbin való kiértékelés rosszabb eredményt ad.

A korpusznak az X (ismeretlen szavak), Z (korpuszhibák), valamint O (nyílt tokenosztály) kategóriájú elemeket is tartalmazó mondatait különválasztottuk. A mondatokat összekevertük, és különválasztottunk ennek tizedét teszthalmaznak, a fennmaradó részt pedig tanítóhalmazként használtuk. Az ellenőrzéseket a teszthalmazon és a teljes halmazon is elvégeztük.

A következőkben leírjuk a módszer elméleti hátterét, részletesebben bemutatjuk a Szeged Korpuszon való munkánkat, majd számszerűsített eredményeket adunk a módszer hatékonyságáról, végül néhány következtetést is megfogalmazunk.

2 A módszer elméleti háttere

A szófaji címkézés problémáját az irodalomban több különböző eszközkészlettel veszik munkába, amelyek alkalmazását magyar nyelvre már többen számba vették [4][5]. A nagyobb pontosságra törekvő módszerekben két összetevő található meg a kívánt végeredmény eléréséhez: az elsőben megállapítjuk, hogy adott szó melyik kategóriákba eshet, vagy legalább egy kiindulási címkét kap (pl. TBL esetén); a másodikban a szó környezetét figyelembe véve döntünk a szóbajövő alternatívák között.

A szó lehetséges szófajainak megállapítása történhet egyszerűen szólista alapján: a tanítóhalmazban előforduló szavakhoz tároljuk a lehetséges kategóriákat. Ennél lényegesen kifinomultabb módszer, amikor morfológiai elemzőt veszünk igénybe a szavak elemzéséhez. Mindkét esetben gondot okoznak azok a szavak, amelyek nem szerepeltek a listában (nem látott szavak), illetve az elemző szótárában (OOV szavak). Ennek a nehézségnek az orvoslására is több módszer létezik, melyek a szó egyes jellemzői (jellemzően az utolsó karakterei, stb.) alapján próbálnak egy/több jó becslést adni a szó szófajára, pl. a maximum entrópia módszer [3].

A [2]-ben bemutatott módszerünkben egy szó különböző osztályokhoz való tartozására a P(szó | osztály) valószínűségét döntési fával becsüljük. A becslés pontosságára adott küszöb az, hogy a tanítóhalmaz adott részét helyesen osztályozza, a szó leggyakoribb kategóriájának adva a legnagyobb valószínűséget.

Így nem szükséges előzetes feltételezés alapján kiválasztani a szó azon részét (pl. utolsó *n* karakter), amelynek várhatóan hatása lesz az osztályozás szempontjából, hanem az rátanul a probléma megoldására a rendelkezésre álló minták alapján: olyan irányban bővíti a döntési fát, hogy az a lehető legkevesebb feltétel vizsgálatával adott pontosságú helyes szétválasztást érjen el.

A második lépés elvégzéséhez is számos módszert használnak, pl. Markov-láncot [6], vagy TBL-t (Transformation Based Learning) [7]. A mi módszerünk leginkább a

Markov-lánccal való megközelítése hasonlít, mivel a szófajnak az adott helyen való előfordulása valószínűségét becsli a környezetében lévő szavak címkéje alapján. Viszont nem csupán a szót megelőző, hanem az azt követő szavak osztályát is figyelembe veheti. Ehhez szabály-sablonokat definiálunk, amelyekből a tanítóhalmaz alapján szabályokat hozunk létre: ezek azt adják meg, hogy adott környezetben milyen valószínűséggel fordulnak elő a különböző címkék. A szóra kapott címkevalószínűséget ezzel az értékkel módosítva egy pontosabb becslést kapunk a címkék valószínűségére, ami segít megtalálni a legvalószínűbb címkesort.

A módszer részletesebb leírását az érdeklődő olvasó megtalálhatja pl. a [2]-ben. A következőkben ennek a szófaji címkézésre való alkalmazását mutatjuk be.

3 Munka a Szeged Korpusz 2.0-val

A módszert a Szeged Korpusz 2.0-n [1] próbáltuk ki (2005. december 5-i állapot). Ez tartalmazza a szöveghatárok jelölését, ezen belül a mondatokat, a mondaton belül pedig annak kifejezéseit MSD morfo-szintaktikai címkékkel ellátva, a szintaktikai egységek jelölésével együtt. Ezt a korpuszt használtuk a betanításhoz és a teszteléshez.

3.1 A korpusz előzetes vizsgálata

A korpuszban az egyes szófaji kategóriákba eső szavakat megvizsgálva azt láttuk, hogy a már korábban felsorolt három kategóriába (X, Z, O) tartozó szavak automatikus címkézésével nem tudunk foglalkozni, mivel nincs megadva javított kód a korpuszhibákhoz, valamint az O kategória elemei időnként más osztályoktól nehezen megkülönböztethetők (pl. létezik számnév kategória, de ebben is van szám típus). A Type-Token Ratio (a kategóriába tartozó szavak és ezek különböző fajtáinak aránya) is jelzi a feladat bonyolultságát: erre a három osztályra a legnagyobb, mivel nagyon kevés közöttük a többször ismétlődő.

A gépi címkézés úgy működik, hogy szeparátor karakterek mentén egységekre tagolja, majd címkézni a szöveget. A szóköznek mindenképpen szeparátor karakternek kell lennie, a csak alfanumerikus karakterekből álló szövegrészeknek pedig mindenképpen egy egységet kell alkotnia. Ezen túl viszont további munkát igényelt a a szeparáló karakterek körének meghatározása úgy, hogy minél inkább a korpuszban előforduló egységekkel dolgozzon az algoritmusunk. Ehhez kigyűjtöttük a nem csak alfanumerikus karaktereket tartalmazó kifejezéseket, és azoknak a központozásoknak az előfordulását megengedtük az egységben, amelyek ezeket a kifejezéseket nem bontották meg. Így a ’.’, ’-’, ’—’, ’,’, ’/’, ’"’, ’&’, ’+’, ’:’, ’\’, ’@’ karakterek egy előfordulását alfanumerikus karakterek között nem vesszük szeparátornak.

A korpusz egy jellegzetessége, hogy szóközt is tartalmazó egységeket is tartalmaz (pl. „OTP Bank Rt.” mint főnév), ami a címkézés eredményének kiértékelését nehezíti, hiszen ezeket a címkéző nem tudja egy egységként értelmezni (bár egy utólagos feldolgozási lépés során ezekből előállíthatók a szintaktikai egységek, de ez számottevően megnehezíti a feladatot). Hogy lehetőleg mégis jól összehasonlítható eredményt kapjunk, használható megközelítés egy olyan szövegváltozat létrehozása, amelyben a kifejezések szófaji címkéjét az azt alkotó, szóközzel vagy más szeparátor

karakterrel határolt egységek mindegyikének elejére elhelyezzük (a fenti példával: „{N}OTP {N}Bank {N}Rt.”). Jóllehet ez nem ad minden esetben helyes címkét a különálló szavakra, mégis esélyt teremt arra, hogy több adattal tanítsuk az algoritmusunkat, valamint hogy a címkézés eredményét értékeljük. Az értékelést ezzel a szöveggel, és az eredeti címke-elhelyezést megtartó szöveggel is elvégeztük.

3.2 A tanító és teszt halmaz előállítása

A kiértékeléshez a korpuszt szétválasztottuk az X, Z, O kategóriákat nem tartalmazó és tartalmazó részekre, ezeknek mondatait összekevertük, és 9 részt tanításra, 1 részt tesztelésre használunk. Ennek megvalósítása úgy történt, hogy a korpusz XML formátumú tartalmából létrehoztunk egy címkék nélküli egyszerű szöveget, amely az automatikus címkéző bemenete lesz, valamint egy egyszerűsített, csak a szófaji címkékkel címkézett kifejezéseket tartalmazó szöveget. Kiértékeléskor a kimenetet ezzel hasonlítjuk össze, annak érdekében, hogy csökkentsük a munkához használt anyag méretét, és ne legyen további szükség XML feldolgozásra.

A tanító és teszt szövegek létrehozásakor ellenőriztük, hogy az XML fájlokban megadott mondatokat kiadja-e a mondathoz tartozó szavak listája. A korpusz használt változatában csak néhány apróbb eltérést találtunk, általában abból fakadóan, hogy a szerkesztők a mondatot eredeti állapotában, érintetlenül hagyták, míg a mondatot alkotó kifejezések felsorolásában helyenként javították a szöveg szóköz és központozás hibáit. Az eltérések tehát a szóközök más elhelyezésében (pl. „D. J.” helyett „D.J.”), illetve három egymást követő pontnak a „…” speciális karakterre való cseréjében jelentek meg; egy esetben jelent meg egy plusz ’.’ központozás a mondatvégi „stb.” kifejezés után. A címkéket tartalmazó szöveget az XML fájlok címkézett kifejezéseiből állítottuk elő. Itt is külön figyelmet igényeltek a szóköz-szeparátorok, mivel ezek az XML címkék között nem jelennek meg, ezért csak az eredeti mondatnak és a kifejezések sorozatának az egybevetéséből tudtuk őket kinyerni, hogy a tanító és a teszt-halmaz csak a címkékben térjen el egymástól. A feldolgozás során végzett konzisztencia-ellenőrzésekkel sikerült néhány hibás HTML címkét, felesleges XML címkét, hibás ill. rosszul zárójelezett MSD kódot, valamint véletlenül beszúrt karaktert is megtalálni, ami remélhetőleg segített még tovább javítani az addig is nagyon jó minőségű korpuszt.

3.3 Kiértékelés módja

A kiértékelés során megvizsgáljuk, hogy az algoritmus milyen arányú egyezést ad a tanítóhalmazban látott szavakra, a tanítóhalmazban nem látott szavakra, valamint ezek együttesére (a korpusz egészére). A teljes MSD címkékre való tanítás mellett fontosnak találtuk megnézni, hogy az MSD fő (szófaji) kategóriáira mennyire hatékonyan működik, hiszen egyes alkalmazásokban ez a kevésbé részletes jellemzés is elégséges lehet (pl. szöveg-beszéd átalakítás hangsúlyozásának javítására).

Ha egy szóra megállapított címke hibás, akkor is érdekes lehet az algoritmusunk jóságának értékelése szempontjából, hogy más környezetben előfordulhat-e a szó ezzel a címkével. Más szóval az is fontos, hogy olyan szófajúnak címkéztük-e, amilyen szerepben előfordulhat, csak a szövegkörnyezet másfajta egyértelműsítést tett volna szükségessé, vagy egyáltalán nem fordulhat elő a szó a megállapított szerep-

ben. Ennek megállapításához azt is ellenőrizzük, hogy a megadott alternatív MSD címkék között előfordul-e az általunk adott besorolás. Ha szólistákkal illetve morfológiai elemzővel dolgozunk, akkor ez az eset nem fordulhat elő, viszont ezek önmagukban nem is képesek ismeretlen szóhoz szófajt javasolni.

Az ellenőrzést elvégezzük a feldolgozási egységekre címkézett és az eredeti címke-pozíciókat tartalmazó szövegekre is.

4 Számszerű eredmények

Az eredményeket a következő pontokban adjuk meg, először a szavak önmagukban való címkézésének feladatára, azután a szavak környezetét is figyelembe vevő megoldásra. Mindenütt megadjuk annak az eredményét, ha csak az MSD főkategóriákkal (a szófajokkal) dolgozunk, ill. ha a teljes morfo-szintaktikai címkékkel, valamint hogy milyen arányban találtuk el a helyes címkét, és milyen arányban csak a lehetséges címkék egyikére döntöttünk.

4.1 A módszer erőforrásigénye

A szükséges erőforrásigény mind memóriakapacitásban, mint számításigényben csekély. A főkategóriák becslésére használt adatbázisunk nem egészen 1 megabájt méretű, és átlagosan 4,1 mélységű döntési fa bejárását igényli karakterenként, valamint a címke-kategóriák számának (11) megfelelő számú összeadást. A szófajra való döntés szavanként a legnagyobb szám megtalálásával történik (10 összehasonlítás). Ezért az algoritmus az itt megadott eredményt kis futási idővel éri el. A részletes címkékhez 37,2 megabájtos, átlagosan 2,9 mélységű döntési fával végeztük a tanító halmazban előfordult 980 kategóriára.

A környezet hatásának figyelembevétele megnöveli a futási időt, de ez sem számottevően. Itt is egy sekély mélységű döntési fa bejárására van szükség a címke valószínűségek kiszámításához használandó szabály megtalálásához, ez esetben lépésenként egyszer minden szóra; a végleges címkesorhoz a szavak számával arányos számú lépésben eljuthatunk.

4.2 Gyakorlati és elvi korlátok

Az, hogy legfeljebb milyen mélységűre nőhet az alkalmazott döntési fa, a tanítás elvégzése előtt eldöntendő. Mi az MSD főkategóriára való tanítás esetén maximálisan 5 mélységet, a teljes MSD címkére való tanításkor 4 mélységet engedtünk meg. Ilyen szintű korlátozásra jelen esetben elsősorban a tanítás memóriában való limitálásához volt szükség, bár természetesen a túltanítás (az általánosító képesség elvesztése) elkerülése végett is érdemes küszöböt megszabni erre a mélyégre. Ha a teljes MSD címkék betanítása esetén nagyobb küszöböt engedtünk volna, számottevően megnőtt volna a betanítás erőforrásigénye, viszont az elkészült adatbázisé nem, vagy nem számottevően. Ekkor a helyes azonosításának aránya megnőtt volna, legkevesebb, hogy a látott szavakon való javulás miatt. Az itt látható eredmények tehát nem a módszer képességeinek elvi korlátját jelentik.

A látott szavak helyes azonosításának arányát a tanítóhalmazban előforduló leggyakoribb címke használatával magasabbra lehetne emelni. Ezt a lehetőség valóban szükséges lehet kihasználni a pontosabb eredmény eléréséhez, bár ez számottevő adatbázisméret növekedést jelent, ha explicit címke valószínűségeket is tárolunk a környezethez igazodó címkék kiszámításához. Emellett a szólistában elő nem forduló szavakra két feldolgozási lépést: egy meghiúsult keresést egy szófában, majd egy új számítási lépést a nem látott szó szófajának becslésére. Azonban látni fogjuk, hogy pl. csak a főkategóriákra való döntés esetén a szólista kiküszöbölésével is egy azzal közel egyenértékű megoldásra jutunk.

4.3 Szavak szófajának környezettől független becslése

Az 1. táblázatban látjuk a MSD főkategóriák becslésének eredményét. A táblázat első két sorában azt az esetet látjuk, amikor több szóból álló kifejezéseknél minden szóra a teljes kifejezés címkéjét helyezzük, erre végezzük a tanítást és az ellenőrzést. A harmadik és negyedik sorban az ellenőrzést a korpusz eredeti címkéire végezzük, ami számottevően rosszabb eredményt ad, hiszen az több szóból álló kifejezések címkéje ilyenkor nem lehet helyes. Láthatjuk, hogy a tanítóhalmazban nem látott szavak több, mint 91%-ára helyes címkét helyezünk azoknak a látott szavakhoz való hasonlósága alapján, és csak kb. 7% kapott olyan címkét, amivel nem fordulhat elő.

A 2. táblázatban a teljes MSD címkék becsléséről láthatunk adatokat. Nem látott szavak több, mint 54%-ához a helyes címkét sikerült rendelni, a szó környezetének figyelembe vétele nélkül. Ahogy fentebb írtuk, a látott szavak helyes azonosításának arányát nagyobb mélységű döntési fa készítésével, vagy egyszerűen szólista használatával is megoldhatjuk, ami az utolsó előtti oszlopban 100%-os eredményt adna, a többi oszlop adatai pedig ennek az aránynak megfelelően változnának.

1. Táblázat: azonosítási arányok szavanként az MSD főkategóriákra

	ellenőrizve valódi címkékre			ellenőrizve alternatív címkékre		
	nem látottra	látottra	összesre	nem látottra	látottra	összesre
szavakra címkézett, csak teszt	91.73	95.70	95.40	94.07	99.10	98.72
szavakra címkézett, egész	91.39	95.94	95.90	93.81	99.06	99.02
eredetire címkézett, csak teszt	88.61	93.97	93.57	90.90	97.38	96.88
eredetire címkézett, egész	82.65	94.03	93.93	84.78	97.15	97.05

2. Táblázat: azonosítási arányok szavanként a teljes MSD kategóriákra

	ellenőrizve valódi címkékre			ellenőrizve alternatív címkékre		
	nem látottra	látottra	összesre	nem látottra	látottra	összesre
szavakra címkézett, csak teszt	54.27	81.82	79.67	58.09	91.15	88.58
szavakra címkézett, egész	54.02	82.66	82.43	57.86	91.29	91.02
eredetire címkézett, csak teszt	52.92	80.80	78.62	56.63	90.14	87.51
eredetire címkézett, egész	49.48	81.54	81.27	52.96	90.16	89.85

3. Táblázat: azonosítási arányok környezettel az MSD főkategóriákra

	ellenőrizve valódi címkékre			ellenőrizve alternatív címkékre		
	nem látottra	látottra	összesre	nem látottra	látottra	összesre
szavakra címkézett, csak teszt	92.64	96.22	95.95	94.64	98.79	98.48
szavakra címkézett, egész	92.34	96.45	96.42	94.39	98.79	98.76
eredetire címkézett, csak teszt	89.50	94.48	94.10	91.44	97.05	96.62
eredetire címkézett, egész	83.46	94.52	94.43	85.27	96.86	96.77

4. Táblázat: azonosítási arányok környezettel a teljes MSD kategóriákra

	ellenőrizve valódi címkékre			ellenőrizve alternatív címkékre		
	nem látottra	látottra	összesre	nem látottra	látottra	összesre
szavakra címkézett, csak teszt	57.37	90.45	87.88	60.74	95.18	92.49
szavakra címkézett, egész	57.11	91.16	90.89	60.50	95.32	95.05
eredetire címkézett, csak teszt	55.60	88.97	86.35	58.87	93.69	90.96
eredetire címkézett, egész	51.99	89.63	89.31	55.06	93.79	93.47

4.4 Szavak szófajának környezetet figyelembe vevő becslése

Ebben a részben a szavak környezetét is figyelembe vevő módszer kiértékelését adjuk meg. Láthatjuk a 3. és 4. táblázatban, hogy az előző pontban megadottakhoz képest számottevő javulást kapunk, különösen a teljes MSD címkékkel való munka esetén. A nem látott szavak helyes azonosításának eredménye kb. 3%-kal megnőtt, a látott szavaké 9%-kal. Ennek a különbségnek az oka, hogy a látott szavak valószínűségét sokkal pontosabban tudjuk becsülni, így a címke valószínűség szabályok használata is jobb eredményeket ad.

Az eredmények nem a módszerrel elérhető legjobb eredményt tükrözik (hiszen szólista használatával az összesített eredmények számottevően jobbak lennének), hanem hogy használható becslést kapunk vele a nem látott szavakra, ill. a látott szavakra is magas recall mértéket kapunk azok külön számontartása nélkül.

4 Konklúziók

A cikkben bemutattuk egy, eredetileg nyelvazonosításra kidolgozott automatikus címkéző módszerünk szófaji címkézésre való alkalmazását a Szeged Korpusz 2.0-n.

A kapott eredmények elmaradnak az egyéb, magyar nyelvre publikált azonosítási rátáktól. Ennek oka egyrészt, hogy nem használtunk morfológiai elemzőt, valamint a teszt során nem használtuk a tanítóhalmazban látott szavak listáját sem, hanem a látott és nem látott szavak címkéjére is közelítő címkét alkalmaztunk. Emellett a nyelvi címkézés nyelvi kategóriái sokkal jobban illeszkednek az alkalmazott valószínűségszámítási modellhez, míg a morfológiai címkézésre a morfológiai elem-

ző és a szólisták használata elvileg jobban illeszkedik a probléma pontos megoldásához.

Az eljárás azonban több előnyös tulajdonsággal rendelkezik: az MSD főkategóriák (szófaji címkék) előállítását csekély erőforrás igénnyel is egyes alkalmazások számára elégséges pontossággal végzi, valamint a nem látott szavak címkéire jó becslést ad. Ezek miatt alkalmas lehet valósidejű alkalmazásokban való használatra, pl. a szövegbeszéd átalakítás problémakörében, illetve más módszerek kiegészítésére.

A kutatásokat részben a Promóció (GVOP – 3.1.1. – 2004 – 05-0245/3.0) projekt támogatásával végeztük.

Bibliográfia

1. Csendes, D., Hatvani, Cs., Alexin, Z., Csirik, J., Gyimóthy, T., Prószéky, G., Váradi, T.: Kézzel annotált magyar nyelvi korpusz: a szeged korpusz. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged. (2003) 238-245
2. Kiss, G., Németh, G.: Machine learning algorithm for automatic labeling and its application in text-to-speech conversion. Híradástechinka, Vol. LXI, Scientific Association for Infocommunications. (2006) 28–35
3. Halácsy, P., Kornai, A., Varga, D.: Morfológiai egyértelműsítés maximum entrópia módszerrel. Magyar Számítógépes Nyelvészeti Konferencia 2005 (2005) 180–189
4. Kuba, A., Felföldi, L., Kocsor, A.: Pos tagger combinations on Hungarian text. 2nd International Joint Conference on Natural Language Processing, IJCNLP (2005)
5. Horváth, T., Alexin, Z., Gyimóthy, T., Wrobel, S.: Application of Different Learning Methods to Hungarian Part-of-speech Tagging. Proceedings of Ninth Workshop on Inductive Logic Programming (ILP99) Bled, Slovenia (1999)
6. Tsujii, S., J., Rim, H.: Part-of-Speech Tagging Based on Hidden Markov Model Assuming Joint Independence. Proceedings of the 38th Annual Meeting of the ACL (2000)
7. Ramshaw, L. A., Marcus, M. P.: Text chunking using transformation-based learning. Proceedings of the Third Annual Workshop on Very Large Corpora (1995)

Milyen a még jobb Humor?

Novák Attila[1] és M. Pintér Tibor[2]

[1]MorphoLogic Kft., 1126 Budapest, Orbánhegyi út 5.,
novak@morphologic.hu
[2]MTA Nyelvtudományi Intézete, 1068 Budapest, Benczúr u 33.,
tpinter@nytud.hu

Kivonat: A számítógépes nyelvfeldolgozás egyik alapvető és általában nélkülözhetetlen eszköze a morfológiai elemző. Az elemzőnek képesnek kell lennie az összes produktív szóalaktani jelenség (ragozás, képzés, szóösszetétel) kezelésére. Bár a lexikai többértelműség viszonylag gyakori jelenség, a többértelmű szavak különböző lehetséges elemzéseihez gyakran nagyon különböző valószínűség rendelhető. A morfológiai elemzőre épülő alkalmazások hatékony működése, illetve bizonyos esetekben helyes funkcionalitása szempontjából is hasznos, ha a lexikai többértelműségek körét sikerül csökkenteni. A valószínűtlen elemzések kiszűréséhez az utóbbi hónapokban szisztematikus korpuszalapú vizsgálatot végeztünk olyan morfológiai jelenségekkel kapcsolatban a magyarban, amelyek rendszeresen vezetnek elemzési többértelműségekhez. Cikkünkben ezeket a vizsgálatokat és az elvégzett lexikográfiai munkát mutatjuk be.

1 Lexikai többértelműségek

A magyarhoz hasonlóan bonyolult morfológiájú nyelvek számítógépes feldolgozása során a nyelvben előforduló lehetséges szóalakok igen magas száma miatt a morfológiai elemzés alkalmazása gyakorlatilag elkerülhetetlen. A morfológiai elemzőnek képesnek kell lennie az összes produktív szóalaktani jelenség (ragozás, képzés, szóösszetétel) kezelésére.

Egy morfológiai elemző kimenetét szemlélve feltűnik, hogy a lexikai többértelműség (azaz az a jelenség, hogy egy szóalaknak egynél több lehetséges elemzése van) viszonylag gyakori jelenség. A többértelműségeket leggyakrabban homonímiák, a különböző paradigmák véletlenszerű vagy rendszerszerű átfedései, illetve a paradigmán belüli rendszeres átfedések okozzák. A magyarban például rendszeres többértelműségre vezetnek az alábbi paradigmatikus átfedések:

az igei paradigmán belül pl.:
vettem, mostam (*én valamit* vagy *én azt*)
vennétek, mosnátok (*ti valamit* vagy *ti azt*)
vennék (*én valamit* vagy *ők azt* – elöl képzett harmóniájú igék esetében)
eszik (*ő* vagy *ők azt* – elöl képzett harmóniájú tárgyas ikes igék esetében)

˜ *néztek*[19] (*ti most* vagy *ők akkor* – elöl képzett harmóniájú igék esetében)
˜ *festette* (*ő azt akkor* vagy *ő azt valakivel akkor* (műveltető) – csak msh.+t tövű elöl képzett harmóniájú igék esetében)

a névszói paradigmán belül:
˜ gyerekével (*az ő gyerekével* vagy *a gyerek valamijével*) – csak az elöl képzett tövű szavak esetében, de minden tővégnyúlást kiváltó esetraggal

A fenti paradigmán belüli többértelműségek és az egyedi tőhomonímiák mellett (melyek gyakran számos toldalékolt alak egybeesésével is együtt járnak: pl. *vár(nak), nyúl(nak)* stb.) sok olyan eset is van, ahol különböző szavak paradigmáinak csak egyes tagjai esnek egybe, a lemma nem:

mentek (*én valamit* (lemma=*ment*), *ti most* vagy *ők akkor* (lemma=*megy*))
csend (*főnév* vagy *te azt*)

Bár a fenti példák esetében különböző valószínűség rendelhető a többértelmű szavak különböző lehetséges értelmezéseihez (elemzéseihez): pl. a *csend* szóalak jóval gyakrabban fordul elő főnévként, mint igealakként, és a sima birtokos alakok (*gyereke*) is gyakoribbak az anaforikus birtokos alakoknál (*gyereké*), a felsorolt példák mindegyike minden említett értelmében viszonylag gyakran előfordul magyar nyelvű szövegekben.

2 Valószínűtlen elemzések

Vannak azonban olyan esetek is, amikor lehetséges értelmezések valamelyikének valószínűsége a többi elemzéséhez képest lényegében elhanyagolható. A korábbi példákban szereplő *vettem* szóalak elvileg elemezhető a *vesz* ige befejezett melléknévi igeneve (*vett*) birtokos alakjaként is, valójában azonban a befejezett melléknévi igenevek birtokos alakjai nemigen használatosak (tehát az adott elemzés valószínűsége szinte zérus, bár grammatikailag elvileg lehetséges: *–Van egy nyers gyémántom. – Na és van csiszoltad is?*).

Hasonló módon a *mentek* szóalak is lehet a *megy* ige befejezett melléknévi igeneve (*ment*) többes számú alakja is, azonban talán ez az elemzés is csak kevéssel valószínűbb, mint a fenti *vettem* alak melléknévi igeneves elemzése.

Amellett, hogy bizonyos általános nyelvi konstrukciók a gyakorlatban soha nem fordulnak elő (mint pl. a befejezett melléknévi igenevek birtokos alakjai), vannak más esetek is, amikor az elemző által több morfémából összeállított egyes elemzések valószínűsége lényegében elhanyagolható az alternatív elemzésekéhez képest. Ilyenek egyrészt azok az esetek, ahol a sokmorfémás elemzés kompozicionálisan kiszámítható jelentése abszurd, másrészt azok, ahol az alternatív elemzés olyan lexikalizált jelentésű tövet tartalmaz, amelyhez képest – bár a kompozicionális elemzés jelentése

[19] A *néztek* és *festette* típusú többértelműségek csak a nyílt *e*-zárt *ë* megkülönböztetést nem ismerő standard nyelvváltozatban (és persze az írott nyelvben) állnak fenn. Az *ë*-ző dialektusban ezek a szavak nem többértelműek: *néztek (ők akkor)-néztëk (ti), fëstëtte (ő azt akkor)-fëstette (ő azt valakivel akkor).*

korántsem képtelenség – előfordulási valószínűsége mégis elhanyagolható. Ezek azok az esetek, amelyekben az anyanyelvi beszélőben a valószínűtlen elemzés lehetősége általában fel sem merül, a (jelentés-összetevőt nem tartalmazó) formális nyelvi modellt alkalmazó gépi algoritmus számára ezek az elemzések azonban éppoly lehetségesek, mint az összes többi.

A morfológiai elemzőre épülő alkalmazások hatékony működése, illetve bizonyos esetekben helyes funkcionalitása szempontjából is hasznos, ha a lexikai többértelműségek körét sikerül csökkenteni. Ehhez célszerű a morfológiai elemzőben implementált produktív szóalaktani folyamatokat úgy megszorítani, hogy a valószínűtlen elemzések minél nagyobb körét ki tudjuk zárni, lehetőleg anélkül, hogy ugyanakkor sok érvényes elemzést elveszítsünk. A MorphoLogic Humor elemzőprogramjához készült morfológiai lexikonok készítéséhez az utóbbi években használt morfológiaiadatbázis-készítő keretrendszer [3] jegyalapú formalizmusa lehetővé teszi az egyes morfológiai jelenségek produktivitásának pontos szabályozását, illetve az elemző is tartalmaz olyan mechanizmusokat, amelyekkel bizonyos elemzések vagy azok alternatívái már a morfológiai elemzés szintjén kiszűrhetők.

A valószínűtlen elemzések kiszűréséhez szisztematikus korpuszalapú vizsgálatot végeztünk számos olyan morfológiai jelenséggel kapcsolatban a magyarban, amelyek rendszeresen vezetnek elemzési többértelműségekhez. Mivel mind a lexikalizálódás, mind a komponált jelentés abszurditásának lehetősége elsősorban a szóképzés és a szóösszetétel körében áll fenn, vizsgálatainkat elsősorban ebben a körben végeztük.

3 Többértelműségek a képzett szavak körében

A magyar képzőkészlet egyes elemei önmagukban is meglehetősen sok többértelműséget vetnek fel. Az alábbi képzők például mind önmagukban is többértelműek:

-s	*harcosak, barackosak* (melléknév), *harcosok, barackosok* (főnév)
-ó	*abban bizakodó* (melléknévi igenév), *bizakodóak* (melléknév), *ablakmosó* (főnév), *ablakmosó gép* (itt jelzői helyzetben)
-z(ik)	*ftp-zik, (le)ftp-z: ftp-zett/ftp-zett le*
-(t)at(ik)	*kihirdettet, kihirdettetik: kihirdettetett*

Ráadásul néhány képzősorozatnak látszó képződmény önálló képzőként önálló jelentéssel is rendelkezik. Ilyenkor a több képzőként való elemzés gyakran (bár nem minden esetben) szintén nagyon valószínűtlen (hibás):

nyávogós:	1. *nyávog+ó+s* (amiben van nyávogó)	2. *nyávog+ós* (hajlamos a nyávogásra)
katonáskodik	1. *katoná+skodik* (katonaként tölti az idejét)	2. ?*katoná+s+kodik* (?katonásan tölti az idejét)
elmagyarosodik	1. *magyar+osodik* (=egyre magyarabb lesz)	2. ?*magyar+os+odik* (=?egyre magyarosabb lesz)

Ezek a többértelműségek (még ha nyelvészetileg megalapozottak is), komoly hatékonysági problémákhoz vezethetnek a morfológia kimenetére épülő szintaktikai elemzés szintjén, különösen azokban az esetekben, ahol több többértelmű képző

együttes előfordulása esetén a többértelműségek összeszorzódnak. Ezért egyik célunk a képzett alakok potenciális többértelműségének csökkentése volt. Módszerünk elsősorban korpusz alapú vizsgálatokon alapult. Korpuszként a Szószablya projektum keretében létrehozott Webkorpuszból [1], [2] készült szóalak-gyakorisági listát használtuk, amelyet a Humor elemzőprogrammal elemeztünk.

3.1 Az *-s* képző

Elsőként az *-s* képző korpuszbeli előfordulásait vizsgáltuk. Feltételezésünk az alábbi volt: a melléknévképző *-s* lényegében teljesen produktív, a főnévképző *-s* azonban nem igazán az: bár sok foglalkozásnév *-s* képzős (órás, lakatos stb.), illetve sok növénynév *-s* képzős alakja használatos az adott növénnyel benőtt terület jelölésére (kukoricás, akácos, gyümölcsös stb.), de teljes körű produktivitásról még ezeken a zárt szemantikai osztályokon belül sem beszélhetünk. Ezért megvizsgáltuk a nem nyitótőként viselkedő *-s* képzősként (is) elemezhető szavakat. Célunk a lexikalizálódott *-s* képzős főnevek kiszótárazása volt, hogy az *-s* képző többértelmű elemzését megszüntethessük. A vizsgálat eredményeképpen meggyőződtünk róla, hogy kiinduló hipotézisünk téves volt: a nem nyitótőként ragozott *-s* képző éppolyan produktív, mint a nyitótőként viselkedő, mert '...-os ember' értelemben éppoly tág körben használható, mint általában a melléknévképző *-s*. Ugyanakkor arra jutottunk, hogy míg egyes alkalmazásokban érdemes a nyitótőként viselkedő (melléknévi) és a nem nyitótő (főnévi) *-s* képzőt megkülönböztetni (például a szóalak-generátorban, ahol így a melléknévi és főnévi *-s* képzős szavak generált alakjai jól elkülönülnek), addig más alkalmazásokban (például a magyar–angol fordítóprogramban) nem érdemes megtartani ezt a megkülönböztetést, mert az *-s* képzős alakok fordításakor a nyitótőség nem játszik szerepet, hanem az számít, hogy az *-s* képzős szerkezet jelzői vagy állítmányi szerepet tölt-e be:

a sárga sisakos férfi	the man with a yellow helmet
a férfi sárga sisakos	the man has a yellow helmet
a sárga sisakosok/sisakosak	the ones with yellow helmets

3.2 Az *-ó* képző

Az *-ó* képzővel kapcsolatban a következő megállapításokat tehetjük:

1. Mivel a melléknévi igenevek mind az alapige vonzatkeretének nagy részét öröklik, mind az igék egyéb bővítési lehetőségeivel rendelkeznek általában, ezért mindenképpen érdemes különválasztani azokat az eseteket, ahol melléknévi igenévi elemzés kizárható. Ebbe a körbe egyértelműen csak a *főnév+ige+Ó* alakú *-ó* képzős szavak tartoznak. Konkrétan ez a szerkezet azonban a morfológiai elemző kimenete alapján akkor is egyértelműen megkülönböztethető, ha az ezt a konstrukciót megtestesítő szavakban semmilyen speciális módon nem annotáljuk magát az *-ó* képzőt.
2. A *főnév+ige+Ó* alakú *-ó* képzős szavak önálló főnévként szerepelhetnek (pl. *kőaprító*), ugyanakkor az érvényes helyesírási norma megengedi azt is, hogy ezek a szavak mintegy jelzőként külön írt szóként épüljenek be nagyobb névszói szerkezetekbe (pl. *kőaprító gép*). Ettől persze ezek a szerkezetek nem lesznek jelzős

szerkezetek, mindenesetre a morfológiai elemző kimenetét felhasználó mondattani elemzőnek képesnek kell lennie ezeknek a szerkezeteknek a kezelésére is. Ehhez azonban nem szükséges, hogy ezeket a szavakat kétféleképpen annotáljuk, mert nem a mondattani elemző általános jelzős szerkezeteket leíró mintája, hanem egy specifikus *főnév+ige+Ó főnév* alakú minta kezeli őket. Ugyanakkor számos ilyen alakú szó lexikalizálódott melléknév (pl. *bizalomgerjesztő*).

3. Az *-ó* képzős melléknevek a főnevekkel és a melléknévi igenevekkel ellentétben általában nyitótőként viselkednek (ugyanakkor a tényleges nyelvhasználatban ez opcionális: *a drogériák bizakodók*), képezhető belőlük határozószó (*bizakodóan*), fokozhatóak (*bizakodóbb*), állhatnak állítmányként. Az egyértelműen nyitótövet tartalmazó szóalakok, a határozói és fokozott alakok csak melléknevek lehetnek.
4. Amennyiben minden *-ó* képzőt azonos módon annotál a morfológiai elemző, akkor az egyetlen megkülönböztetés, ami az elemző által szolgáltatott jegyekből nem rekonstruálható, az egyértelműen nyitótövet tartalmazó szóalakok határozott melléknévi besorolása. Ebben az egy esetben a többi esettől eltérő annotációt kell alkalmazni. A többi esetben elegendő ha egyetlen elemzést ad a morfológiai elemző, a szintaxisnak ugyanakkor ebben az esetben ezt az egy elemzést az adott szókonstrukciónak megfelelő módon kell értelmeznie.

4 Valószínűtlen összetételek

A többértelmű névszóképzőkkel kapcsolatban elvégzett vizsgálatok ugyanakkor rengeteg egyéb problémára vetettek fényt. Ezek közül kiemelten kezelendőek a mára lexikalizálódott *-s* képzős főnevek, amelyek produktív *-s* képzős elemzése egyéb alkalmazásokban problémát okozhat. Fordítóprogramban problematikusak lehetnek például az *anyós=anyó+s, mártás=Márta+s, csapás=csapa+s, emlős=emlő+s* alakok, amelyek képzett szóként történő kezelése hibás fordítást eredményezhet. Sajátos problémaként jelentkeznek az olyan esetek, amikor a szó egyszerre elemezhető *-s* képzős, nyílt szótagra végződő tőként, illetve *-ás/-és* képzős zárt szótagra végződő tőként: *kígyómarás=kígyó+mar+ás ~ kígyó+mara+s, adás=ad+ás ~ Ada+s*, *csapás=csap+ás ~ csapa+s*.

A formális elemzések és a szemantika találkozásakor létrejövő kompozicionális és derivációs sajátosságok közül említést érdemelnek az abszurdnak tűnő jelentésű (egyébként morfológiailag normális) összetételek. Ilyen összetételek szép számmal előfordultak már az *-s* képzős szavaknál is. Kedvenc példáink között tartjuk számon az alábbi szavakat: *anyósom=anyó+som ~ anyós+om, apósom=apó+som ~ após+om, altatás=altat+ás ~ al+tatás* (ilyenek még: *alanyuk=alany+uk ~ al+anyuk, alapjuk=alap+juk ~ al+apjuk*). A felsorolt példák lényege, hogy bennük a többértelműség úgy jön létre, hogy az elemzés egyik eleme összetett szó (ami a „túlelemzett" alak), a másik egyszerű, toldalékos szó. Ennek az alcsoportnak további sajátossága, hogy az összetett alakok mindkét tagja fogalmi jelentéssel bíró „tartalmas" szó (általában főnév). Sajátos csoportot alkotnak azonban azok az összetételként viselkedő szavak, amelyek második tagja hangutánzó szó, indulatszó vagy valamilyen szóteremtéssel keletkezett határozott fogalmi tartalom nélküli szó: *lényekké=lé+nyekk+é, farkukkal=far+kukk+kal.* Egy következő csoport tagjai azok az összetételek, amelyek a mai standardban nem használatosak, egyértelmű archaizmusok vagy historizmusok,

s elfogadásuk csak nehezítené a program működését: pl. az elavult *szaka* szóval képzett összetételként: *korszaka=kor+szak+a ~ kor+szaka*.

Mindhárom csoport kezelése szerencsére egyszerűen – és más-más módon – megoldható még a morfológiában. A morfológiai elemző tartalmaz egy olyan mechanizmust amelynek segítségével egy adott morf többmorfémás alternatív elemzései letilthatók. Az első csoport esetében (*anyósom*) ezt a mechanizmust használva oszthatatlanná lehet tenni a toldalékos szó szótövét (*anyós*), így elkerülhető a szemantikailag hibás vagy nem normatív alakok keletkezése. Más megoldást kíván a második csoport (*lényekké*): mivel szemantikailag nem teljes értékű szavak alkotják az összetétel második tagját, amelyek egyébként a nyelvhasználatban sem alkotnak összetételeket, ezért ezeket olyan jeggyel láttuk el a szótárban, amely előírja, hogy nem állhatnak összetételek második tagjaként (és első tagként sem). Ugyanez a megoldás a harmadik csoport (*szaka*) esetében is alkalmazható. Ezek esetében azonban akár azt a megoldást is választhatjuk, hogy mivel ezek a szavak a mai nyelvhasználatban már csak ritkán fordulnak elő (ezt mutatják a korpuszalapú vizsgálatok, többek között az Értelmező kéziszótár második kiadása is), ezért – a duplicitások elkerülése végett – nem vesszük fel őket a szótárba.

Egy további csoportot alkotnak azok az alakok, amelyek rendszeres többértelműsége a következőképpen alakul: egyik lehetséges elemzésükkor szerves részét alkotják a névszói ragozási paradigmának, másik elemzésükben ugyanakkor összetételként szerepelnek (mivel szótövük utolsó szótagja egybeesik valamelyik névszóraggal, ugyanakkor a toldalékékként szereplő szótag előtti rész is értelmes magyar szó). Tipikus példák az *ének*, *ében*, *ára, inak, imre, kánon, román, szemét stb.* végződésű szavak: *telepének=telep+é+nek ~ telep+ének*, *szerepében=szerep+é+ben ~ szerep+ében*, *csatára=csatár+a ~ csat+ár+a*, *tanulóinak=tanuló+i+nak ~ tanuló+in+ak*, *Pestszentimre=Pest+szent+imre ~ Pest+szent+i+m+re*, *misekánon=mise+kánon ~ mise+kán+on*, *nagyromán= nagy+román ~ nagy+roma+n*, *fémszemét= fém+szemét ~ fém+szem+é+t*. A felsoroltak egy része jól kezelhető a morfológiai elemzőn belül (pl. az *ének* és *ében* főnevek esetében előírhatjuk, hogy összetételben nem állhatnak elöl képzett harmóniájú tövek után, a ténylegesen előforduló ilyen eseteket pedig felvesszük a szótárba), más részük kezelése már túlmutat a morfológián, egyértelműsíteni ezeket leginkább szintaktikai szűrőszabályokkal lehet: pl. a tömeges *-ára* végű többértelműségeket a mondattan szintjén úgy is kiküszöbölhetjük, hogy nem kell a morfológiában az *ár* végű összetételek produktív képzését teljesen letiltani: csak a főnév+*ár+a*[PSe3] elemzéseket érvényteleníti egy esetleges *...a+ra*[SUB] elemzés. Ily módon az ilyen típusú morfológiai túlelemzések az elemzett, vagy fordított szövegben már nem jelennek meg.

Néhány további példa szemantikai anomáliákat mutató túlelemzésekre:

vadbarom=vad+bar+om,
anyagcsere-állapot=anyagcsere+-+ál+lap+ot,
ultramarinkék=ultra[FN]+mar[FN]+i[_IKEP]+nk[PSt1]+ék[FAM]+[NOM]
korcsmáros=korc+smár+os

Ezek az elemzések a magyar alaktan formális szabályainak megfelelnek, azonban a magyar nyelvi kompetenciával rendelkező beszélők ezeket a szavakat az esetek szinte kizárólagos többségében a magyar nyelvközösségben normatívvá vált jelentésben használják. Mivel a program elemzéskor az összes lehetséges morfemikus határon szegmetál (és nem rendelkezik kommunikatív kompetenciával), ezért

esetenként „túlelemzéseket” hoz létre. Korpuszalapú elemzések során csak a főneves összetételeknél több, mint 220 olyan típust találtunk, amelyek a fentiekhez hasonló morfológiai többértelműséghez vezetnek – pontosabban vezettek, mivel a vizsgálat közben felmerült jelenségeket időközben javítottuk. A következőkben ezek közül, a már kijavított esetek közül mutatnánk be párat, természetesen a teljesség igénye nélkül:

gyönyörhullám=gyönyör+hulla[FN][PSe1][NOM],
nőcinek=nőci+nek ~ nő+cin+ek,
gleccserhasadék=gleccser+hasadék ~ gleccser+has+ad+ék,
kakasukkal=kaka+sukk[FN][INS][20],
tengerfenék=tengerfene[FN][PL].

A „túlelemzések” között nem csak összetett főnevek vannak:

fémdoboz=fém+dob+oz, *combizom=combizik[IGE][Te1]*

Az utolsó példa jól szemlélteti a szótáralapú elemzés (és generálás) problémáját: a szótövesítéskor keletkezett szótő (ami attól szótő, mert a szótár tartalmazza) és a hozzá kapcsolódó ismert toldalékok kapcsolatát a program mindaddig „valós” szóként értelmezi, amíg bizonyos eseteket vagy szabályokat meg nem tiltunk neki (például, hogy a *dob* főnévből képzett ige nem lehet *-z* képzős, amit legegyszerűbb módon a doboz további részekre elemzésének letiltásával érhetünk el).

5 Igei többértelműségek

Az igék esetében igazából két nagy problémakör merült fel, amelyek mélyebb elemzések után további alkategóriákra oszlottak.

5.1 Az ikes–iktelen többértelműség

Az egyik felmerülő problémakör az igék ikességét érintette. Mivel az egymástól csak ikességben különböző igék paradigmája a szótári alak kivételével (mely az ikes igék esetében *-ik*re, a nem ikesek esetében Ø-ra végződik) általában teljesen egybeesik (bár az igepár ikes tagja gyakran nem tárgyas), ezért az ikes–iktelen párok a lexikai többértelműségek egyik legmasszívabb forrását adják, különös tekintettel arra a tényre, hogy a *-z* képzőnek mind az ikes mind az iktelen változata rendkívül produktív.

Az ikesség elkülönítésére szabály alapú mechanizmust sajnos nem lehet kialakítani. Ehelyett az elemzett korpuszból kiválogatott szóalakokat tartalmazó listákat kellett átnéznünk, és az abban látottak alapján szabályokat felállítani, amelyeket a listán lévő igékhez rendeltünk. Fontos megemlíteni, hogy mint azt a cikk elején felsorolt paradigmán belüli többértelműsége között már említettük, az *-ik* toldalék homonim alak, azaz egymástól független jelentésben több típusú *-ik* is él a magyar nyelvben: egyrészt mint kijelentő mód, jelen idő, cselekvő, egyes szám harmadik személyű

[20] Az Éksz[2] szerint a *sukk* nem más, mint “a két ököl és az egymás felé fordított két hüvelykujj együttes hossza mint (ácsok használta) hosszmérték”.

alak (ő *eszik* valamit), másrészt mint kijelentő mód, jelen idő, cselekvő, határozott tárgyas, többes szám harmadik személyű alak (ők *eszik* azt). Korpuszalapú vizsgálatunkban (amely 10 805 db igealakot ölelt fel) természetesen az első esetből indultunk ki, ami az igék esetében a szótári alak. Ez alapján az ikes–iktelen párokat átvizsgáltuk a valószínűtlen és lehetőleg megszüntetendő többértelműségek feltárása érdekében. Megjelöltük azokat az alakokat, amelyek

1. *morfológiailag hibásak* (pl. *felület* vagy *magzat* mint műveltető alak),
2. morfológiailag helyes, de szemantikailag nem, ezért pusztán elvileg *lehetséges alaknak* tekintettük (pl. *beleillet*, *megmérettet*),
3. *csak igekötővel és tárggyal fordulnak elő* (pl. *szar*, *száraz*),
4. olyan *produktív -z képzős* alakok, amelynek töve egybeesik egy lexikai *z(ik)* tövű igéével és többmorfémás elemzése rendkívül valószínűtlen (pl. *fáz, távoz, szerezik, fogalmazik*),
5. amelyek egyben *más szófajúak* is (pl. *feladat, aszat*).

A műveltetővel gyakori többértelműséget okozó *-(t)atik* passzív képző esetében – mivel ez a képző már nemigen produktív –, legjobb megoldásnak a korpuszban talált lexikalizálódott alakok felvételét, és a képző törlését láttuk. Ez viszonylag könnyen kivitelezhető a hátul képzett tövek esetében (itt az *-atik* végű alakok egyértelműen ennek a képzőnek az előfordulásai), az elöl képzett tövek esetében azonban a műveltető képző többes szám 3. személyű alakjával való egybeesés miatt a tényleges passzív alakok kiszűrése nagyon nehéz (pl. az *emlékeztetik* nyilvánvalóan nem passzív igealak – hiszen az *emlékezik* nem tárgyas –, mégis úgy néz ki, mintha az lenne).

5.2 Az igetárgyasság

Bár korábban elsősorban a szóképzést és a szóösszetételt emeltük ki, mint a valószínűtlen túlelemzések forrását, a ragozás körében találunk ilyen eseteket.

Korábbi morfológiai leírásunk egyik hiányossága volt, hogy nagyrészt hiányzott belőle az igetöveknek a tárgyasság szempontjából való besorolása. Ez bizonyos esetekben túlelemzésekhez vezetett: határozott tárgyas elemzést kaptunk olyan igék esetében is, amelyek nem lehetnek tárgyasak. Úgy találtuk azonban, hogy a tárgyasság szempontjából annotált létező leírások sok esetben téves besorolásokat tartalmaznak, amelyek átvételével ténylegesen jó szóalakok jó elemzéseit veszítenénk el. Ezeknek a leírásoknak a készítői nem olyan szempontrendszer figyelembe vételével készítették el ugyanis a leírásukat, amelyek alapján számunkra is használható leírás készülhetett volna.

Ennek oka az, hogy az igetárgyasság más nyelvtani jelenségekkel viszonylag komplex módon interakcióba lép: bizonyos igekötős konstrukciók az egyébként tárgyatlan igék egy igen széles körét is tárgyassá tehetik, ugyanakkor a szintaktikai igevivőinverzió következtében az igekötőnek nem muszáj az adott szóalakon megjelennie:

vitatkozik (nem tárgyas ige), de: *míg ki nem vitatkozza magát...*

Ezért a korábbiaknál pontosabb modelleket kellett kidolgoznunk, illetve nem szorítkozhattunk szótárakban megadott információkra, mert ellenkező esetben megszorításaink túl szigorúak lettek volna. Az igéket az alábbi alapvető osztályokba soroltuk:

S soha nem lehet tárgyas
I csak speciális igekötős konstrukciókban lehet tárgyas: *át...-za az éjszakát*, *ki...-za magát*, vagy más igekötős változata tárgyas
N tárgyas is lehet meg nem is minden speciális mellékkörülmény nélkül (jó a *...-za azt* konstrukció, de lehet csak úgy simán *...-zni* is.)
T csak tárgyas

Az igéket első körben a korpusz szóanyagát elemezve egy heurisztika segítségével soroltuk a megfelelő osztályba. Minden igetőhöz a korpuszban szereplő alakok alapján kiszámítottunk egy mérőszámot, amely azt mutatta, hogy azok közül a szóalakok közül, amelyek elemezhetők az adott tő előfordulásaként hány olyan alak volt, ami kizárólag határozott tárgyas igealakként elemezhető (az utóbbi osztva az előbbivel):

1=minden alakja egyértelműen határozott tárgyas volt.
0=nem volt egyértelműen határozott tárgyas alakja.

Egyértelműen határozott tárgyasnak akkor minősítettünk egy alakot, ha egyáltalán nem volt határozott tárgyas igétől különböző elemzése (tehát egy esetleges névszói vagy más szófajú elemzés is kizáró ok volt). Számos más mérőszámmal is kísérleteztünk, végül azonban ezt a viszonylag egyszerű indikátort elég hatékonynak találtuk. Az osztályozó heurisztika az említett indikátor mellett az ikességet vette figyelembe, az alábbi határértékek használatával:

T<0.0008	S
T<0.011, nem ikes	I
T<0.16, ikes	I
T<0.33	N
egyébként	T

A besorolást kézzel javítottuk, és tovább finomítottuk:

I2 Lehet a szónak (pl. speciális igekötős konstrukcióban) 2. személyű tárgyas alakja, pl. *eljöttelek meglátogatni*.
IN Van olyan igekötős változata (az általános *át/végig...-za az éjszakát*, *ki...-za magát* konstrukciókon kívül), ami gyakori és tárgyas, de igekötő nélkül egyértelműen tárgyatlan.
IT Az ige eleve csak igekötővel létezik, (az elválhat, de teljesen igekötő nélkül nincs, ilyen pl. a *(felül)múl*) és az egyértelműen tárgyas.

A morfológiai elemzőben csak az S (soha nem tárgyas) osztályba besorolt igéket zárhatjuk ki a határozott tárgyas elemzést kapó igék köréből. Ugyanakkor a szintaktikai elemzés szintjén az I (csak igekötős konstrukciókban tárgyas) osztályba sorolt igék többértelműségei is hatékonyan csökkenthetők a vonzatkeretek ellenőrzésével.

5.3 Az ikes igék E/1 alakja

Az ikes igék körében egy másik megoldandó probléma volt a hagyományos (de mára lényegében kihalt) önálló ikes paradigmából megmaradt nem határozott tárgyas jelen idő egyes szám első személyű *-m* toldalék kezelése (*eszem valamit*, *alszom egyet* stb.). Ez ugyanis korántsem minden ikes ige esetében lehetséges: **baszom/fingom/nőzöm egyet, *bánom vele, *meghízom* stb. Ezért külön lexikográfiai munkát jelentett ezeknek az ikes igéknek a feltérképezése, amelyet szintén korpusz alapú vizsgálatokból nyert különböző indikátorok szerint rendezett listák többszöri átnézésével, és kézi javításával végeztünk el. A használt indikátorok az *-m* ragos és *-k* ragos alakok száma, ezek egymáshoz és az adott ige összes alakjához viszonyított előfordulási gyakorisága voltak.

6 Összefoglalás

Cikkünkben olyan rendszeresen elemzési többértelműségekhez vezető morfológiai jelenségekkel kapcsolatos szisztematikus korpuszalapú vizsgálatokat mutattunk be magyarban, amelyek eredményeképpen a Humor morfológiai elemzőben a lexikai többértelműségek körét sikerült csökkenteni. A vizsgálatok eredményei mellett bemutattuk az elvégzett lexikográfiai munkát és a túlgenerálás csökkentésének módszereit.

Bibliográfia

1. Halácsy Péter, Kornai András, Németh László, Rung András, Szakadát István, Trón Viktor. Creating open language resources for Hungarian. In: Proceedings of LREC2004, 2004
2. Kornai, A, Halácsy, P, Nagy, V, 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, 1–9.
3. Novák Attila. Milyen a jó humor? In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), pp. 138–145, Szegedi Tudományegyetem, 2003.

III. Ontológia

Az általános ontológia egy új modellje

Varasdi Károly[1], Gyarmathy Zsófia[1], Simonyi András[2], Szeredi Dániel[2]

[1] MTA Nyelvtudományi Intézet
[2] BME – Média Oktató és Kutató Központ
e-mail:{varasdi, gyzsof}@nytud.hu, andras.simonyi@gmail.com, daniel@szeredi.hu

Kivonat Tanulmányunkban az általános ontológia egy olyan új modelljét mutatjuk be, amely a kortárs filozófia és a kognitív tudomány eredményeinek figyelembe vételével került kidolgozásra. A kutatás a Magyar Egységes Ontológia (MEO) projekt [1] keretében jelenleg is folytatott munka részét képezi.[3]

Kulcsszavak: általános ontológia, konceptuális terek, modális dependencia

1. Bevezetés

A gyakran idézett, Thomas Grubertől származó tömör megfogalmazás szerint az ontológia „egy (kölcsönösen elfogadott) fogalmi rendszer explicit, formális specifikációja” [6]. Míg ennek a meghatározásnak a relatíve szűk lefedést feltételező szakontológiák esetében többnyire viszonylag problémamentesen eleget lehet tenni, a tartományfüggetlen *általános ontológiák* esetében olyan kérdések válnak hangsúlyossá, amelyek a szakontológiák esetében nem vagy alig kapnak súlyt. Ilyenek például a következők:

- Mit kell érteni „kölcsönös elfogadottságon” az általános ontológia esetében?
- Hogyan értendő a „fogalmi rendszer” kifejezés az általános ontológia viszonylatában?
- Hogyan kell elképzelni az általános ontológia „explicit, formális specifikációját”?

2. A Konceptuális Terek modellje

Az általános ontológia esetében a *kölcsönös elfogadottságot* alapvetően az emberi faj különböző egyedeihez tartozó kognitív struktúrák nagyfokú hasonlósága

[3] A jelen tanulmányban leírt elképzeléseket a MEO projekt (NKFP-2/042/04.sz.) támogatásával dolgoztuk ki, és kialakításukban a szerzőkön kívül aktív szerepe volt Héja Enikőnek, Mittelholz Ivánnak, Szakadát Istvánnak, Szőts Miklósnak és Ungváry Rudolfnak. Nekik a szerzők ezúton is meg szeretnék köszönni a segítséget. Természetesen egyikük sem felelős a szövegben esetleg előforduló tévedésekért.

garantálja. E struktúrák általános elméletéért a kognitív tudományhoz fordulhatunk. A kognitív tudomány egyik erőteljesen fejlődő részterülete szakított a pusztán szimbolikus alapú reprezentációkat feltételező mechanizmusokkal, és bevonta a tárgyalásba az emberi megismerést legalább olyan szinten jellemző téri aspektust is. E megközelítés az elsősorban Peter Gärdenfors nevéhez fűződő *Konceptuális Terek* modelljében érhető tetten, amelynek filozófiai előzményeit Robert Stalnaker dolgozta ki [10]. Stalnaker eredetileg a modális logika egy alternatív megalapozására tett javaslatot, s elképzelésének alapgondolata szerint egy tetszőleges (ténylegesen vagy csak lehetségesen létező) entitást egy helyvektor képvisel egy olyan térben, amelynek dimenziói az entitást jellemző tulajdonságok (konkrét értékei). Így például egy konkrét, 3 cm sugarú piros g gömböt egy olyan absztrakt térben tudunk lokalizálni, amelynek egyik dimenziója a lehetséges színeket, a másik pedig a gömb rádiuszát képviseli. E térben g-t egy olyan vektor azonosítja, amelynek szín dimenzióra vetített értéke a piros tartományba, a rádiusz dimenzión vett vetülete pedig a 3 cm értékre esik. Könnyen látható, hogy ebben a térben minden egyes helyvektor egy-egy lehetséges — adott színű és méretű — gömböt reprezentál. Ez a rendkívül egyszerű példa már jól szemlélteti a megközelítés főbb vonásait is, illetve kapcsolatát a hagyományos szimbolikus alapú reprezentációkkal. Így például a „4 cm-nél kisebb sugarú piros gömbnek lenni" összetett tulajdonság (fogalom) e kétdimenziós tér egy bizonyos R tartományának feleltethető meg, és az a kijelentés, hogy g rendelkezik ezzel a tulajdonsággal egyszerűen annak ellenőrzését jelenti, hogy g helyvektora az R régióba mutat-e. Hasonlóképpen, a „valamilyen sugarú piros gömbnek lenni" fogalomhoz egy R' tartomány rendelhető, és R valamint R' tartalmazási viszonyai ($R \subset R'$) annak a következtetésnek az ellenőrzését is lehetővé teszik, miszerint minden 4 cm-nél kisebb sugarú piros gömb egyben piros gömb.

Ennek az elképzelésnek a Stalnaker-félén kívül más előzményei is léteznek a filozófiában. Az a felfogás, amely szerint bármely fizikai entitás felfogható mint saját tulajdonságai összesége az ún. troposzelmélet (*trope theory*) néven ismert a filozófiatörténetben [9]. Adott entitás troposzai azok a csakis hozzá tartozó „konkrét, partikuláris, térben és időben lokalizált tulajdonságdarabkák," amelyek azt jellemzik; pl. egy adott rózsa konkrét színe. Mivel különböző entitásokhoz különböző troposzok tartoznak, azt a tényt, hogy két rózsa pontosan ugyanolyan színű úgy tudjuk kifejezni, hogy azt mondjuk, hogy a szóbanforgó rózsákhoz tartozó színtroposzok tökéletesen hasonlítanak (anélkül azonban, hogy azonosak lennének). A jelen megközelítésben a troposzokat olyan primitív entitásoknak tekintjük, amikből az összetett entitások (pl. fizikai tárgyak) felépülnek. (A troposzok ontológiába való beemelésével a a DOLCE ontológiát követjük [8].)

Gärdenfors tehát a fenti filozófiai koncepcióra építette saját elméletét, amelyet [5] mutat be részletesen. Gärdenfors Stalnaker filozófiai keretelméletét empirikus tartalommal kívánta megtölteni. Feltételezése szerint az egyes dimenziók belső szervező elve a *hasonlóság*, azaz az egyes tulajdonságértékek távolsága arányos a hasonlóságukkal, amit egy egyszerű választásos teszt segítségével kísérelt

meg kimérni.[4] Gärdenfors a releváns dimenziók kiválasztására is tett egy vázlatos javaslatot — lényegében egy a faktoranalízisre épülő eljárásról van szó —, amelynek részleteiről az olvasó az idézett műben tájékozódhat. A MEO-projekt keretében végzett kutatásunkban e faktoranalízisre építő eljárás helyett a lexikai szemantikában megszokott jegyekre bontás eljárását választottuk, mert céljainknak valamint a rendelkezésünkre álló idő- és energiakorlátoknak ez felelt meg jobban. Megjegyezzük azonban, hogy véleményünk szerint a releváns dimenziók meghatározására hosszabb távon igen reménykeltőek a Formal Concept Analysis eljárásának eredményei is (ennek rövid bemutatását ld. pl. [2] 3. fejezetében).

Gärdenfors — Stalnaker nyomán — a fogalmakat e konceptuális tér régióival azonosítja. Mivel azonban a dimenziók szervezőelvét a hasonlóságban találja meg, levezethetővé válik számára az emberi fogalmi készlet azon általános (és a fenti példában hallgatólagosan fel is használt) sajátossága, hogy a tanulható fogalmak e tér konvex résztartományait jelölik ki (ez véleménye szerint a dimenziók hasonlósági alapon történő szerveződéséből vezethető le az ún. Voronoi-parkettázás eljárásán keresztül).

Az egyes dimenziók alapvetően skálaszerkezetűek, azaz a dimenzió értékei lineárisan elrendezettek. Ez nem szükségszerű feltétel ugyan — elvileg lehetségesek lennének ciklikus dimenziók is —, de mi egyszerűségi okoknál fogva előnyben részesítettük a lineáris szerkezetű dimenziókat.

A fentiekben a DOLCE általános ontológiához hasonlóan jártunk el, bár az általunk fejlesztett ontológia igen sok szempontból eltér a DOLCE-féle kerettől: míg például a DOLCE lényegében csak a csúcskategóriák elmélete, mi alacsonyabb szintű fogalomleírások elkészítését és rendszerbe szervezését is célul tűztük ki, így olyan kérdésekre is választ kellett találnunk, amikkel a DOLCE készítői nem szembesültek.

3. Az általános ontológia szerkezete

Az általános ontológia általános fogalmak közötti kapcsolatok leírása, azaz fogalmi rendszer. *Fogalmi rendszer* alatt tehát a fogalmakból kialakuló relációs struktúrát értjük. Ennek a szerveződésnek két alaptípusát különböztetjük meg: a horizontálist és a vertikálist. Kezdjük az előbbivel.

3.1. Horizontális tagozódás

A *horizontális* szerveződés szintjén a fogalmak által megnevezett entitástípusok közötti szükségszerű (esszenciális) kapcsolatokat, közelebbről dependenciaviszonyokat értjük. Ilyen az például, hogy színelőforduláshoz szükségszerűen tartozik egy konkrét felületelőfordulás, amelyen megjelenik. A színelőfordulás olyan értelemben dependál a felületelőforduláson, hogy utóbbi nélkül nem tudna megjelenni, azaz ez a kapcsolat szükségszerű is: minden színelőfordulás szükségképpen

[4] A kísérleti személyeknek három tulajdonságértéket mutatnak, majd felteszik a kérdést, hogy melyik kettőt ítéli hasonlóbbnak a három közül.

von maga után egy hozzá tartozó felületelőfordulást. Ebben a példában látható az is, hogy a dependenciaviszony nem kell, hogy aszimmetrikus legyen, hiszen érvelhetünk amellett is, hogy a felületelőfordulások szükségképpen valamely színelőfordulást implikálnak. Egy olyan példa, ahol a szimmetria nyilván nem áll fenn, a következő: minden házasságkötési esemény esszenciálisan dependál a menyasszony létén, ám fordítva természetesen nem áll fenn a dependenciaviszony (menyasszony létezhet házasságkötési esemény nélkül is). Az ilyen összefüggések a legmagasabb fokú (fogalmilag szükségszerű) fogalmi kapcsolatokat írják le, amelyek alól nem engedünk meg kivételeket, s így ezek alkotják az ontológia legáltalánosabb viszonyrendszerét. Az ontológia ezen megközelítése hagyományosan Edmund Husserl nevéhez fűződik [7], és a dependencia fogalmának szigorúbb matematikai alapokra helyezésére az utóbbi időben történtek kísérletek [3]. Az alábbiakban azonban egy viszonylag egyszerűbb közelítéssel fogunk dolgozni, amely céljainknak jobban megfelel, mint a Kit Fine-féle formalizáció.

A dependenciaviszony formális jellemzése. A dependencia fogalmi kapcsolat definiálására jelen tanulmányban nem vállalkozhatunk; az alábbiakban pusztán egy fontos *szükséges feltételt* fogalmazunk meg erre a relációtípusra nézve. Legyen A, B az ontológia két tetszőleges típusa (pl. a felületelőfordulások illetve a színelőfordulások fogalmi típusai). Ha R dependenciareláció A és B között, akkor R ki kell, hogy elégítse az alábbi feltételt:

$$\Box\forall x(x\ \mathsf{instanceOf}\ A \rightarrow \exists!y(y\ \mathsf{instanceOf}\ B \wedge R(x,y))). \tag{1}$$

Szavakban: *szükségszerű*, hogy A bármely x instanciájához található B egy olyan y instanciája, hogy x R viszonyban áll y-nal. A természetes nyelvben a dependenciarelációkat gyakran birtokos esettel fejezzük ki (pl. *színe*, *alakja*, stb.), de — mint az esküvői példa is mutatja — ez inkább csak tendencia, semmint szabály.

A relációkra vonatkozó „dependencia” metapredikátum fogalmi, intenzionális karakterét a ‘$\Box$’ szükségszerűségoperátor jelenléte biztosítja, így a szükségszerűség különböző fokozatainak figyelembe vételével különböző erősségű dependenciaviszonyokhoz jutunk. A fentiekben a szín–felület kapcsolat esetében az ún. metafizikai szükségszerűség egy példáját láttuk. Ez a szükségszerűség rendkívül erős, már-már logikai erejű. Tekintsünk most egy gyengébb szükségszerűségtípust és egy rá alapozott dependenciaviszonyt: ha például $\Box$-t mint „a biológia törvényszerűségei szerint szükségszerű, hogy” jeleként értelmezzük, valamint A-t a férfiak, B-t a nők típusával azonosítjuk, akkor

$$\Box\forall x(x\ \mathsf{instanceOf}\ \mathsf{férfi} \rightarrow \exists!y(y\ \mathsf{instanceOf}\ \mathsf{nő} \wedge \mathsf{anyja}(x,y)))$$

igaz állítás lesz, ám

$$\Box\forall x(x\ \mathsf{instanceOf}\ \mathsf{férfi} \rightarrow \exists!y(y\ \mathsf{instanceOf}\ \mathsf{nő} \wedge \mathsf{testvére}(x,y)))$$

hamis, azaz az anyja reláció dependenciareláció lesz a férfi és nő típusok között — mert biológiailag szükségszerűen minden férfinak (általában pedig: minden

embernek) van anyja —, de a testvére nem, hiszen biológiailag nem szükségszerű, hogy egy férfinak legyen nővére vagy húga. A példa tanulsága az, hogy a '$\Box$' erősség szerinti indexelésével a dependencia különböző fokozataihoz juthatunk, ami lehetővé teszi a zökkenőmentes átmenetet a legáltalánosabb fogalmi struktúrák leírásától a szaktudományok sajátos domainjeinek leírásáig.

Megjegyezzük, hogy ha R-et az azonosságnak választjuk, akkor a kapott

$$\Box\forall x(x \text{ instanceOf } A \to \exists!y(y \text{ instanceOf } B \wedge x = y)) \iff$$
$$\Box\forall x(x \text{ instanceOf } A \to \exists!y(x \text{ instanceOf } B)) \iff$$
$$\Box\forall x(x \text{ instanceOf } A \to x \text{ instanceOf } B)$$

formula a jól ismert generikus (isa) relációt adja A és B között. Valóban, az azonosságot felfoghatjuk a dependencia *triviális* formájának, hiszen bármely entitás tautologikusan dependál saját létezésén.

A fentiekből kiolvashatók a szükséges *formális specifikációra* vonatkozó megfontolások is: a típusok közötti dependenciaviszonyok leírásához egy megfelelően erős gráfleíró nyelvre van szükség. Hogy az egyes típusok leírásához szükséges erőt csökkentsük, az esetleges dependenciakörökre, szimmetrikus függésekre vonatkozó információkat nem az egyes típusleírások, hanem az ontológia egésze tartalmazza. Így az egyes típusleírások DAG-okkal (*directed acyclic graphs*) történnek, de a típusok közötti esetleges ciklikus összefüggések is visszaállíthatók az ontológia különböző típusleírásaiban tárolt információk összevetésén keresztül. Az alábbi példákban tehát DAG-ok leírására alkalmas ún. AVM-ekkel (*Attribute–Value Matrix*) operáló nyelvet használunk, de megjegyezzük, hogy jelenleg is folytatunk kutatásokat a megfelelő nyelv meghatározására (elsősorban a deskripciós logikák [4] területén). Az egyes típusreprezentáló csomópontoknak tehát mátrixok és végső soron értékek — változók, troposzok vagy akár egész részdimenziók — felelnek meg, a gráféleknek pedig attribútumok. Egy A_i–V_j attribútum-érték párt egy T típus mátrixában a fent már említett szükségszerű egzisztenciális implikációként értelmezzük; eszerint pontosan egy olyan V_j típusú érték tartozik T-hez, amely vele A_i relációban áll.

3.2. Esszenciális és kontingens tulajdonságok

Az egy adott típushoz tartozó dependenciaviszonyok szükségszerű megszorításokat tartalmaznak az adott típus egyedeire vonatkozóan. Például, semmilyen esemény *nem lehet* az esküvő esemény instanciája, ha abban nem azonosítható a menyasszony szerep valaki által történő tényleges instanciálása.

A valóságra vonatkozó információink azonban két tág osztályba sorolhatók. Az egyik osztályba a fent említett *a priori* (fogalmilag szükségszerű) összefüggések tartoznak, a másikba azonban *a posteriori*, kontingens összefüggések. Az a priori, szükségszerű összefüggések ahhoz a hálóhoz tartoznak, amit — Wittgensteinnel szólva[5] — a valóságra fektetünk, hogy kezelhetővé váljon az eredetileg formátlan

[5] Although the spots in our picture are geometrical figures, nevertheless geometry can obviously say nothing at all about their actual form and position. The network,

„massza". A háló törvényei tehát a nyelv és logika törvényei, de az, hogy a háló szemei ténylegesen mivel is töltődnek ki, már a tényleges valóság tulajdonságaitól függ, így kontingens.

Az általános ontológiának képesnek kell lennie a mi világunkat jellemző kontingenciák ábrázolására is.

Hagyományosan az individuumok szükségszerű tulajdonságait (jegyeit) eszszenciális tulajdonságoknak szokták nevezni, és a fentiekben már sokat beszéltünk róluk. Az esszenciális és a kontingens vonások azonban természetesen összefüggnek. Például az, hogy minden (makrovilágbeli) konkrét felülethez tartozik egy konkrét szín, szükségszerűen igaz, de ez az állítás természetesen nem mondja ki, hogy annak a színnek pl. éppen a pirosnak kell lennie. Az, hogy melyik szín lesz az adott felület tényleges színe, már a valóságtól függ. Hasonlóan, az esszenciális lehet egy adott típusba tartozó individuum számára, hogy egy tulajdonságának értéke egy bizonyos intervallumba essen, de az, hogy azon belül pontosan melyik értékkel rendelkezik, már esetleg tisztán kontingens.

Az általános fogalmaink segítségével az „a tisztán a priori háló" szemeit kisebb részekre bonthatjuk. Ezekben a kisebb „kompartmentekben" már megjelenik az esetleges tapasztalat is. Ez a tapasztalat persze nem mondhat ellent a „háló geometriájának," ami az a priori állításokban megfogalmazódik, de tartalmazhat olyan elemeket, amik nagy valószínűséggel jellemzik a háló adott szemében található objektumokat. Ezek a tapasztalati és „kivételt ismerő" általánosítások különböző default következtetések elvégzésére tesznek minket alkalmassá, amelyeket persze a konkrét instanciák „megcáfolhatnak". Az ilyen, pusztán default erejű általánosításokat nevezzük a jelen tanulmányban *proprium*nak. Az alábbiakban ezt, illetve a kontingenciához kapcsolódó egyéb fogalmakat pontosítjuk.[6]

3.3. Akcidentális tulajdonság

Akcidencia Egy $\mathcal{A}$ tulajdonság akcidentális c-ben, ha c nem szükségszerűen rendelkezik $\mathcal{A}$-val, vagyis, ha lehetséges az, hogy c létezik ugyan, de nem rendelkezik $\mathcal{A}$-val.

A c entitás létezése során *lehetnek* olyan időszakok, amikor nem rendelkezik $\mathcal{A}$-val. Ugyanakkor *nem kell*, hogy legyenek ilyen időszakok. c történetesen, „a sors kegyéből kifolyólag" rendelkezhet $\mathcal{A}$-val egész létezése során anélkül, hogy ennek szükségszerűen (törvényszerűen) *így kéne* lenni. Ez indokolja az alábbi fogalom értelmességét.

Proprium Egy $\mathcal{P}$ tulajdonság propriuma c-nek t-kor, ha c-t t-ig terjedő történetének minden vagy legtöbb pillanatában jellemzi ugyan, de nem esszenciális tulajdonsága c-nek.

---

however, is purely geometrical; all its properties can be given a priori. Laws like the principle of sufficient reason, etc. are about the net and not about what the net describes. (Tractatus Logico-Philosophicus: 6.35.)

[6] A továbbiakban a „fogalom" és „típus" kifejezéseket — némi pongyolasággal — szinonímaként kezeljük.

Az, hogy a $\mathcal{P}$ tulajdonság propriumként jellemzi-e c-t vagy sem, nem dönthető el c egy adott pillanatnyi temporális szeletének alapján, csak c egész (eddigi) történetét figyelembe véve. A proprium c történetének alapján képzett induktív általánosítás („c-t eddig többnyire jellemezte $\mathcal{P}$"). Ennek következtében adott időpontban c-ből aktuálisan hiányozhat is a $\mathcal{P}$ tulajdonság anélkül, hogy $\mathcal{P}$ megszűnne c propriumának lenni. Ha azonban ez a hiány c-t történetének nagyobb részében jellemezte, $\mathcal{P}$ nem propriuma c-nek.

Az esszenciális tulajdonság és a proprium közötti különbség lényegében a szükségszerű és a valószínű különbsége, és — ennek következtében — az, hogy míg az esszenciális attribútum nem tűr időben kivételeket, a proprium egy bizonyos fokig tolerálja az ilyeneket. A „proprium" tehát az entitást „kitartóan", tendenciaszerűen, de nem szükségszerű erővel jellemző tulajdonságok gyűjtőneve. Azok a vonások, amik az entitást esetleg csak „futólag", átmenetileg, az idő kisebb részében jellemzik, a következő alpont tárgyát képezik.

Fázis. A „fázis" kifejezés a fizikából ismert „fázistér" kifejezés „visszaképzett" formája. A fázis- vagy állapottér az a „szabadsági fokoknak" nevezett dimenziókból álló tér, amelyben egy rendszer összes lehetséges állapotai vannak reprezentálva úgy, hogy minden egyes lehetséges rendszerállapotnak pontosan egy pont felel meg a fázistérben. Egy nem túl erőltetett *analógia* vonható e között a fogalom és Gärdenfors kognitívtér-fogalma között, amennyiben a rendszer szabadsági fokait a kognitív tér dimenzióinak feleltetjük meg. Az analógia a következőképpen fest.

A rendszer szabadsági fokai az entitásra értelmezhető tulajdonságoknak felelnek meg. Ha a rendszer által a létezése során bejárt trajektóriát levetítjük az egyes szabadsági fokokra, akkor azok a vetületek, amik olyanok, hogy a rendszer számára lehetetlen kijutni belőlük, az esszenciális tulajdonságok tartományainak felelnek meg. A rendszer által a létezése során t-ig bejárt trajektória azon vetületei, amelyben a rendszer „a c keletkezésétől t-ig terjedő időintervallum legnagyobb részében" tartózkodik, a propriumoknak feleltethetők meg. Végül, a rendszer olyan állapotai, amikben adott időpontban tartózkodik, az ontológiában annak feleltethetők meg, amit jelen írásban az entitás fázisának nevezünk:

Fázis Az c entitás azon $\mathcal{F}$ tulajdonságait, amelyeket adott időben birtokol, fázisnak nevezzük.

A fenti metafogalmak viszonyait szemlélteti az alábbi táblázat.

	időben stabil	időben instabil
szükségszerű	Esszencia	—
nem szükségszerű	Proprium	Fázis

A fentiekben a „proprium" és „fázis" fogalmakat az *individuális entitásokra* vonatkozóan fogalmaztuk meg. Ennek alapján azonban analóg definíciók fogalmazhatók meg a típusok esetére is. Egy *típus* propriumán például azon tulajdonságok összességét értjük, amelyek a típusba tartozó ténylegesen aktualizált

instanciák többségének (individuális) propriumai. Ehhez hasonlóan a fázis fogalmát is kiterjeszthetnénk típusokra, ám e fogalom használati értéke meglehetősen kicsi, így most eltekintünk tőle.

3.4. Vertikális szerveződés: három szint

A fenti megkülönböztetések alapján az ontológiában lévő fogalmi csomópontok között vertikálisan három szintet különböztethetünk meg (fentről lefelé):

szint	neve	összetevői
I.	Esszenciális fogalmak	esszenciák
II.	Általános fogalmak	esszenciák és propriumok
III.	Egyedi fogalmak	esszenciák, propriumok és fázisok

I. szintű (esszenciális) fogalmak. Ezek a „nyelvi–logikai háló” sajátosságait írják le, ezért a prioriak; jellemzőjük, hogy az általuk kimondott viszonyokat szükségszerűnek ítéljük, azaz a valóság bármilyen alakulása mellett is ragaszkodunk igazságukhoz. Például ilyen az, hogy a (makro-)fizikai entitásoknak van — sok egyéb szükségszerű tartozék mellett — felülete és tömege:

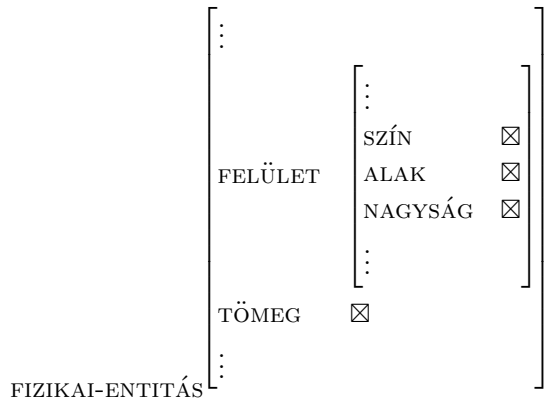
$$\text{FIZIKAI-ENTITÁS}\left[\begin{array}{ll}\vdots & \\ \text{FELÜLET} & \boxtimes \\ \text{TÖMEG} & \boxtimes \\ \vdots & \end{array}\right]$$

Ehhez a következőképpen tudunk interpretációt rendelni: A FIZIKAI-ENTITÁS fogalom bámely instanciájához tartozik a FELÜLET fogalom valamely instanciája, a TÖMEG fogalom valamely instanciája, stb. A dependencia által hozzárendelt típus konkrét értékét ezen a szinten természetesen nem tudjuk specifikálni, hiszen az kontingens.

A fentihez hasonlóan a priori tudjuk, hogy a felületeknek szükségképpen van színe, alakja és kiterjedése (és még esetleg más is lehet a listában):

$$\text{FELÜLET}\left[\begin{array}{ll}\vdots & \\ \text{SZÍN} & \boxtimes \\ \text{ALAK} & \boxtimes \\ \text{NAGYSÁG} & \boxtimes \\ \vdots & \end{array}\right]$$

Azaz a FELÜLET fogalom bármely instanciájához tartozik a SZÍN fogalom egy instanciája, az ALAK fogalom egy instanciája, a NAGYSÁG fogalom egy instanciája, stb. Természetesen az is igaz, hogy egy szükségszerűen kötelező jegy szükségszerűen kötelező jegyei szintén szükségszerűen kötelezőek:



A fentiek I. szintű fogalmak, mert *tisztán* modálisan szükségszerű összefüggéseket tartalmaznak.

II. szintű (általános) fogalmak. Általános fogalom például a „macska” fogalma. Egy általános fogalom deskripciója két típusú információt tartalmaz:

1. esszenciális információk
2. kontingens információk

Az általános fogalom az esszenciális információkat az I. szinten fölötte álló típusoktól mereven megörökli. Például, a MACSKA fogalom minden instanciája fizikai entitás, ezért a MACSKA fogalomban — pontosabban, a MACSKA általános fogalomhoz tartozó csomóponthoz rendelt deskripcióban — benne lesz mindaz, ami a fizikai entitásokat szükségszerűen jellemzi. Ugyanakkor az általános fogalomban lehetnek további esszenciák is — például az, hogy egy macska testhőmérséklete nem lehet 15000 C; ezt anélkül is jól tudjuk, hogy ilyen irányú kísérleteket kéne végeznünk. Ezen a ponton azonban jól látszik, hogy itt a szükségszerűség egy gyengébb — biológiai — fajtájáról van szó, hiszen önmagában logikailag nem ellentmondásos feltételezni, hogy egy macska 15000 fokon is macska maradjon, biológiai ismereteink alapján viszont igen. Azt, hogy mégis az esszenciális információk közé szeretnénk ezt besorolni az indokolja, hogy a hétköznapi (és vélhetőleg a tudományos) gondolkodás is egyaránt *lehetetlennek* ítél egy 15000 Celsius-fokon funkcionáló élőlényt. Az alábbiakban tehát a MACSKA esszenciális jegyének tekintjük, hogy TESTHŐMÉRSÉKLETe a 35.0 és a 42.0 fokos határok közé *kell* hogy essen, mert ha ebből kilép, megszűnik létezni.[7]

A kontingens információk olyan információk, amik az instanciák nagy részét, de esetleg nem mindegyiket jellemzik. Ilyen például az az információ, hogy a MACSKA TESTHŐMÉRSÉKLET jegye a 39.0 ± 0.2 C értéktartományból veszi értékét, mert a macskák többségének testhőmérséklete majdnem minden időpillanatban ebbe az intervallumba esik, azaz propriumról van szó. Egy proprium

[7] Ezt az információt a példa részletesebben kidolgozott változatában a MACSKA az ÉLŐLÉNY típustól örökli.

semmilyen értelemben nem szükségszerű, hiszen pusztán a fogalom tényleges instanciái alapján kialakított generalizáció eredménye, így egy proprium soha nem is mondhat ellent egy esszenciálisnak tekintett megszorításnak.

Az általános fogalmak propriumai az egyes ember számára nem biztos, hogy saját tapasztalaton nyugvó általánosítások, hanem inkább a közösség (szakértőinek) összesített tapasztalatát kodifikálják. A kultúra közvetítésével azonban a különböző létező-típusokhoz rendelt közösségi tapasztalatok beépülnek az egyes ember fogalmi reprezentációiba is kontingens — de nagy valószínűségű — világismeretként.

Mivel az általános fogalmakhoz rendelt propriumok nem a priori szükségszerűségek, ezért *default módon* öröklődnek a csomópontok között. Például, a MACSKA TESTHŐMÉRSÉKLETe ugyan 39.0 ± 0.2 C, de az ANGÓRAMACSKÁé már 39.5 ± 0.1 C.[8]

$$
\text{MACSKA}\left[\begin{array}{ll}
\left(\text{ESSZENCIA}\right) & \left[\begin{array}{ll}
\vdots & \\
\text{FELÜLET} & \left[\begin{array}{ll}
\vdots & \\
\text{SZÍN} & \boxtimes \\
\text{ALAK} & \cdots \\
\text{NAGYSÁG} & \cdots \\
\vdots &
\end{array}\right] \\
\text{TÖMEG} & \leq 20.0 \text{ kg} \\
\text{TESTHŐMÉRSÉKLET} & 35.0 \leq T \leq 42.0 \text{ C} \\
\vdots &
\end{array}\right] \\
\left(\text{PROPRIUM}\right) & \left[\begin{array}{ll}
\vdots & \\
\text{TÖMEG} & 3.5 \leq m \leq 4.5 \text{ kg} \\
\text{TESTHŐMÉRSÉKLET} & 39.0 \pm 0.2 \text{ C} \\
\vdots &
\end{array}\right]
\end{array}\right]
$$

III. szintű (egyedi) fogalmak. Egyedi fogalma egy konkrét individuumnak van, például egy konkrét macskának, mondjuk Félixnek. A FÉLIX egyedi fogalom esszenciális jegyei a felette álló fogalmak esszenciális jegyeinek legspecifikusabbjai lesznek, és azokat nem is tudja felülírni.

A FÉLIXhez rendelt propriumok *kettős rétegzettségűek*: egyik részét a közvetlenül felette lévő *általános fogalomtól* örökli default öröklléssel[9], másrészt neki magának is lehetnek csak rá jellemző *egyedi propriumai*. Például, ha Félix ideje

[8] Az adatok természetesen csak illusztrációs célúak.

[9] ... azaz az így örökölt propriumokat esetleg felülírhatja...

nagy részét a bejárati ajtó előtti lábtörlőn tölti, akkor ez az ő egyedi — de nem a MACSKA általános — fogalmát jellemző proprium.

Végül, Félix különböző sajátos *fázisok*ban is lehet, és e fázisok értékei ellentmondhatnak mind a saját, mind a felette lévő általános fogalom propriumainak. A fázisok ugyanakkor idővel akár Félix egyedi propriumaivá is átalakulhatnak (és ha sok macskánál történik ez meg, akkor a macska általános fogalmához rendelt proprium is megváltozhat). Az alábbi leírás szerint pl. Félix 14:00 órakor — testhőmérséklete alapján látható módon — éppen lázasan feküdt az alomban.

$$
\text{FÉLIX}\left[\begin{array}{ll}
\left(\text{ESSZENCIA}\right) & \left[\begin{array}{ll}
\vdots & \\
\text{FELÜLET} & \left[\begin{array}{ll}
\vdots & \\
\text{ALAK} & \cdots \\
\text{NAGYSÁG} & \cdots \\
\vdots &
\end{array}\right] \\
\text{TÖMEG} & \leq 20.0 \text{ kg} \\
\text{TESTHŐMÉRSÉKLET} & 35.0 \leq T \leq 42.0 \text{ C} \\
\vdots &
\end{array}\right] \\
\left(\text{PROPRIUM}\right) & \left[\begin{array}{ll}
\vdots & \\
\text{SZÍN} & \text{rőt} \\
\text{TÖMEG} & 4.8 \leq m \leq 5.0 \text{ kg} \\
\text{TESTHŐMÉRSÉKLET} & 39.3 \pm 0.2 \text{ C} \\
\text{HELY} & \text{lábtörlő} \\
\vdots &
\end{array}\right] \\
\left(\text{FÁZIS}\right) & \left[\begin{array}{ll}
\vdots & \\
14{:}00 & \left[\begin{array}{ll}
\vdots & \\
\text{TÖMEG} & 4.88 \text{ kg} \\
\text{TESTHŐMÉRSÉKLET} & 40.1 \text{ C} \\
\text{HELY} & \text{alom} \\
\vdots &
\end{array}\right] \\
\vdots &
\end{array}\right]
\end{array}\right]
$$

4. Összefoglalás

A fentiekben — igen vázlatosan — bemutattuk egy készülő általános ontológia keretelméletét. Az ontológia középponti fogalma a típusok közötti dependencia-viszony, illetve az azt reprezentáló dependenciagráf. Az ilyen típuskapcsolatok alkotják az ontológia horizontális szerkezetét. Az ontológia vertikális tagozódásának legfelső, legabsztraktabb szintjén az „erős", logikai–metafizikai jellegű dependendenciák leírása található. Ezek a gráfok kevés, de nagyon általános, kivételt nem ismerő megszorítást fogalmaznak meg a lehetséges létezőkre vonatkozóan. A következő szinten már megjelennek mind a gyengébb („szakmai") modalitások, mind pedig az adott típusra a mi világunkban jellemző propriumok. Végül, az egyedi fogalmak szintjén az addig alulspecifikált értékek is konkretizálódnak. A fenti képet némileg bonyolítja az emberi kognícióval kapcsolatos fogalmak összetettsége, de sajnos helyhiány miatt erre a kérdésre a jelen cikkben már nem tudtunk kitérni.

Hivatkozások

1. Magyar Egységes Ontológia projekt. `http://ontologia.hu`.
2. B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2002.
3. Kit Fine. Part–whole. In Barry Smith and David Woodruff Smith, editors, *The Cambridge Companion to Husserl*. Cambridge University Press, 1995.
4. Franz Baader and Diego Calvanese and Deborah L. McGuinness and Daniele Nardi and Peter F. Patel-Schneider, editor. *The Description Logic Handbook*. Cambridge University Press, 2003.
5. Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. The MIT Press, 2000.
6. Thomas Gruber. Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993. Kluwer Academic Publishers.
7. Edmund Husserl. *Logische Untersuchungen*, volume I.–II. Max Neimeyer: Halle, 1900–1901.
8. C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. The WonderWeb Library of Foundational Ontologies: Preliminary Report. `http://www.loa-cnr.it/Papers/DOLCE2.1-FOL.pdf`, 2003.
9. Peter Simons. Particulars in particular clothing: Three trope theories of substance. In Stephen Laurence and Cynthia Mcdonald, editors, *Contemporary Readings in the Foundations of Metaphysics*, chapter 4, pages 364–385. Blackwell, 1998.
10. Robert Stalnaker. Antiessentialism. *Midwest Studies of Philosophy*, 4:343–355, 1981.

Az ontológiák legfelső generikus szintje, a csúcsfogalmak természetes rendszere és a DOLCE kritikája

Ungváry Rudolf

Országos Széchényi Könyvtár, Könyvtári Intézet
1840 Budavári Palota, F épület 643
ungvary@hungary.com

Kivonat: Az ontológiák fogalmi hierarchiájának elképzelhető a mai fizikai világképen alapuló, ún. természetes rendszere. Csúcsfogalmai és hierarchiájuk lényegében összhangban vannak az eddigi ontológiák csúcsszerkezetével.

1 Bevezetés

Az ontológiák logikai szerkezete fogalmak generikus reláción (logikai leírásokban szokásos jelölése: subclass, is_a, gen) alapuló hierarchiájához kapcsolódik [12][13][14]. Akármilyen is a kapcsolódó logikai szerkezet megoldása, a logikai megoldásnak olyan hierarchia a tárgya, mely alapvetően intuíció segítségével ragadható meg. Más szóval filozófiai szemlélettől (elkötelezettségtől) függően eltérő fogalmi hierarchia választható.

E szemlélettől függ a hierarchia csúcsfogalmainak megválasztása is. Ezek alapvetően meghatározzák, hogy hogyan rendeződnek el legfőbb vonalakban az alárendelt fogalmak, azaz milyen lesz az ontológia hierarchiája. Intuitíve nem csak azt várhatnánk, hogy (a) e generikus hierarchiák részleteikben rendkívül különböznek, hanem azt is, hogy (b) az egyes hierarchiák csúcsfogalmai — tehát az ontológiák kevés számú legáltalánosabb fogalmai — alkotta struktúrák a fogalomtartalmak szempontjából is eltérőek lesznek.

E tanulmány célja, hogy két ontológia csúcsrendszerének példáján bemutassa: az elvárás ellenére az ontológiákhoz kialakított legáltalánosabb fogalmak alkotta struktúrák között — a csúcsfogalmak generikus hierarchiájában — nincsenek igazán áthidalhatatlan értelmezési különbségek. Azaz a második (b) várható következmény valójában nem valósul igazán meg. Következésképp akármekkora — látszólag — az eltérés a generikus hierarchiák részleteiben az egyes ontológiák között, lényegében még ezek rendszere is legáltalánosabb értelemben ugyanazon a szükségszerű fogalmi renden alapszik.

2 Az ontológia

Az ontológia fogalmainak generikus hierarchiája[21] felfogható egy nyelv (tárgynyelv) strukturált (relációkat explicit formában tartalmazó) szótárának: a fogalmak egy-, a relációk kétargumentumú predikátumjelnek (relációjelnek) felelnek meg. Definiálható az ontológia leírására szolgáló metanyelv, amelyben az ontológia relációi, fogalmai konstans jelek, és tartalmazza a következő relációjeleket (nem kimerítő lista):

1. táblázat. Ontológia metanyelvének deklarációja

név	*argumentumszám*	*axióma*
fogalom	egy	a három predikátum terjedelme nem üres, és páronként diszjunkt
reláció	egy	
előfordulás	egy	
fajtája	kettő	$\forall x,y(\text{fajtája}(x,y) \rightarrow ((\text{fogalom}(x) \wedge \text{fogalom}(y)) \vee (\text{reláció}(x) \wedge \text{reláció}(y)))$ a „fajtája” irreflexív, aszimmetrikus és tranzitív (parciális rendezés)
előfordulása	kettő	$\text{előfordulása}(x,y) \rightarrow \text{előfordulás}(x) \wedge \text{fogalom}(y)$
ellentéte	kettő	irreflexív, szimmetrikus
Értelmezve_van	három	$\forall x,y(\text{értelmezve\_van}(x,y,z) \rightarrow \text{reláció}(x) \wedge \text{fogalom}(y) \wedge \text{fogalom}(z))$

A továbbiakban az ontológiában szereplő fogalmakat KISKAPITÁLISSAL, a csúcsfogalmakat **FÉLKÖVÉR KISKAPITÁLISSAL**, a relációneveket normál *dőlt* betűkkel írom. A metarelációk (pl. fajtája) normál betűkkel szerepelnek, ezek nevei mind a meta-, mind a tárgynyelvben előfordulnak. Ha kifejezetten megnevezésről van szó, a nevet ’apsztrofok’ között szerepeltetem.

3 A csúcsfogalmak természetes rendszere

3.1 Rendszerelvek

A Magyar Egységes Ontológia[22] [5] számára a csúcsfogalmak olyan rendszerét alakítottuk ki, melyeket a kialakítás alapjául választott szemlélet alapján „természetesnek” nevezek, mivel nem a filozófia vagy a logika történetileg kialakult általános fogalmain, hanem a mindennapi nyelvhasználatban generikus értelemben legáltalánosabb fogalmakon alapul, különös tekintettel a mai fizikai világkép legál-

[21] A hierarchiát alkotó generikus fogalmi struktúrák önmagukban még nem ontológiák. Csak akkor azok, ha kiegészülnek meghatározott célú logikai következtetőrendszerrel. Ennek ellenére magát az itt tárgyalt csúcshierarchiát önmagában is sokszor felső szintű ontológiának nevezik („upper level ontology”)[1][3]. Az ontológiák generikus hierarchiáját nevezik hibásan taxonómiának is, holott — a taxonómiákkal ellentétben — közvetlenül nem játssza osztályozási rendszer szerepét [11].

[22] A tanulmány a Magyar Egységes Ontológia (MEO) NKFP-2/42/04. sz. projekt keretében készült el.

talánosabb kategóriáira [8][12]. A későbbiekben a DOLCE [1][4] legfelső szintjének négy csúcsfogalmát fogjuk összehasonlítania csúcsfogalmak eme rendszerével.

A fogalmak természetes rendszerének elvi alapjai:

- a generikus reláció és annak polihierarchikus használatának segítségével kifejezett általánosítás: A van B, továbbá A van C. Azaz B fajtája A és C fajtája is A.
- a makrofizikai világkép három kategóriája (anyag, energia, információ);
- az elvont és a konkrét megkülönböztetése;
- a **DOLOG** és a hozzá kapcsolódó szerepfogalmak megkülönböztetése az előző kettőtől,[23]
- **VALAMI**, mint a legáltalánosabb fogalom és ellentétének (**SEMMI**) tételezése

Ebben a természetes rendszerben a legáltalánosabb fogalmak úgy keletkezek, hogy minden fogalom esetében addig tesszem fel a kérdést, hogy **A** van [milyen] **B**, ameddig el nem jutunk a **VALAMI** fogalmáig.

1. SZÉK és KAVICS van FIZIKAI TÁRGY;
- HOMOK és TEJ van ANYAG;
- FIZIKAI TÁRGY és ANYAG van **ANYAGSZERŰ VALAMI**.

(Azaz: egy SZÉK fogalom és egy KAVICS fogalom van két FIZIKAI TÁRGY fogalom.)

2. ÁRAMLÁS és GONDOLKODÁS van FOLYAMAT;
- ERŐ és KÉPESSÉG van HATÁS;
- HATÁS és FOLYAMAT van **ENERGIASZERŰ VALAMI**.

3. SZÍN és HELY van TULAJDONSÁG;
- JEL és FOLYÉKONYSÁG van ÁLLAPOT;
- TULAJDONSÁG és ÁLLAPOT és MINŐSÉG van **INFORMÁCIÓSZERŰ VALAMI**.

A fenti – természetes — általánosítás (és az elvont/konkrét) segítségével minden fogalom alárendelhető a három csúcsfogalomnak, melyek általánosítása a **VALAMI**.

- RELÁCIÓ és SZEREP van ELVONT MOZGÁS, illetve ELVONT ÁLLAPOT (azaz több fölérendelt is megadható);
- ELVONT MOZGÁS van ELVONT ENERGIASZERŰ VALAMI;
- ELVONT ÁLLAPOT van ÁLLAPOT, illetve **ELVONT VALAMI**;
- ELVONT ENERGIASZERŰ VALAMI van **ENERGIASZERŰ VALAMI**, illetve **ELVONT VALAMI**;

Ezek alapján jön létre egyrészt

- az **ANYAGSZERŰ VALAMI**, az **ENERGIASZERŰ VALAMI** és az **INFORMÁCIÓSZERŰ VALAMI** fogalma,
- AZ **ELVONT VALAMI** és a **KONKRÉT VALAMI** fogalma,
- a **DOLOG** és közvetlen alárendeltjeinek (JELENSÉG, TÁRGY, ELŐFORDULÁS, ESEMÉNYSZERŰSÉG) fogalmai (a DOLOG is van VALAMI).

3.2 Minden fogalomnak, amely gondolható

Az így kialakított fogalmai rendszert nevezem a fogalmak természetes rendszerének. Éppen a legátfogóbb fogalom, a **VALAMI**, továbbá az elvont és a konkrét felosztási szempontok következetes alkalmazása, illetve a **DOLOG** és fajtáinak szerepfogalmak-

[23] A „szerep” fogalmát itt az OntoClean értelmében használjuk [6]

ként való felismerése biztosítja, hogy a fogalmak természetes rendszere ne kizárólag a makrofizikai — és egyáltalán: a természettudományos — tapasztalatokon alapuljon. A — természetes — fogalmi rendszerben (és az ezen alapuló ontológiákban) helyet kell tudni találni minden fogalomnak, amely gondolható (pl. ÖRDÖG, KENTAUR ['van' ELVONT LÉNY], FÁBÓL VASKARIKA ['van' ELVONT TÁRGY]), függetlenül attól, hogy természettudományosan megalapozottak-e vagy sem.

E rendszer szerkezeti koherenciáját a generikus reláció — az 1. táblázatban a fajtája — biztosítja. Egy fogalom csak akkor lehet ontológia fogalmi hierarchiájának eleme, ha igaz, hogy e fogalom és fölérendeltje között fennáll a generikus reláció.

Ugyanakkor nem tehető, hogy bármilyen fogalmat kizárjunk, ha egyébként gondolható és megnevezhető. Ha tehát van olyan fogalom, hogy MINŐSÉG, van olyan, hogy TULAJDONSÁG, és van olyan, hogy JELLEMZŐ, akkor ezek a rendszer tárgyfogalmai, és mindegyiknek meg kell határozni a fölérendeltjét (és nem utolsó sorban ezáltal az egymáshoz való viszonyát is expliciten meg kell adni).

Az ontológia generikus hierarchiájában a konkrét, természetes nyelven megnevezett fogalmakat (pl. MINŐSÉG, TULAJDONSÁG, ANYAG, MOZGÁS, **DOLOG**) vagy fogalmak leírását (pl. INFORMÁCIÓSZERŰ VALAMI, BIOLÓGIAI ÁLLAPOT, KONNATÍV PROPOZÍCIONÁLIS TUDATI ÁLLAPOT) kell megadni, nem pedig metaforákat (metaforaként használandó megnevezéseket).

Mindebből következik, hogy a **SEMMI** fogalmának is helye van a rendszerben. Ez a **VALAMI** fogalmának ellentéte, szintén legáltalánosabb fogalom.

3.3 Fogalom és fény

A természetes fogalmai rendszerben az olyan fizikai kategória is, mint az ANYAG, az ENERGIA és az INFORMÁCIÓ is megfelelő fölérendelttel rendelkezik, ugyanis **ANYAGSZERŰ VALAMI**, **ENERGIASZERŰ VALAMI**, illetve **INFORMÁCIÓSZERŰ VALAMI**. Ezek a fogalmak rigidek[24]: minden, ami terjedelmükbe tartozik, létezése első pillanatától az utolsói a terjedelmébe tartozik (az ÉLŐLÉNY például mindig **ANYAGSZERŰ VALAMI**, FUTÁS mindig **ENERGIASZERŰ VALAMI**, a TULAJDONSÁG mindig **INFORMÁCIÓSZERŰ VALAMI**)-

Noha a kvantumlogikailag leírható mikrofizikában az anyag, az energia és az információ jelenségei egyre közelebb kerülnek egymáshoz (a fény jelenségében pedig egybeesnek), a fogalmi rendszer csúcsszerkezetét ez nem befolyásolja. Egyrészt fogalmaink alapja a makrofizikai (érzéki) tapasztalat, melyet a kvantumfizikai felismerések tovább finomítanak. Másrészt éppen a legátfogóbb fogalom, a **VALAMI** az, amelyben ugyancsak „egybeesnek" a fent felsorolt természetes fogalmi kategóriák.

Az olyan, rendkívül általánosnak festő fogalom, mint a **DOLOG** (szinonimája az **ENTITÁS**), ugyancsak speciálisabb tartalmú, mint a **VALAMI**. A **DOLOG** — szemben a fenti három rigid csúcsfogalommal — nem rigid szerepfogalom: valami akkor dolog, ha abból és csak abból a szempontból gondoljuk, hogy létezik. Ennél fogva az, ami a terjedelmébe tartozik, nem tartozik mindig (rigiden) a terjedelmébe. Például egy ember létezése első pillanatától az utolsóig ember, de csak bizonyos esetekben dolog.

A **VALAMI** fogalmáról (és szükségképpen összes fajtáiról is) tárgynyelven pusztán csak az állítható, hogy van, metanyelven pedig az, hogy fogalom. A **VALAMI** fogalma

[24] A rigiditás fogalmát az OntoClean módszertan szerint értelmezem [6].

és vele minden fogalom (melyek a fajtái) tekinthető akár afféle „elvont, tudaton belül létező fénynek” is, melyben minden kategória „egybeesik” [9].

3.4 A természetes csúcsfogalmak logikai deklarációi

A csúcsfogalmakat kétféle formában írjuk le: alárendeltjeikkel, ill. a fogalom tartalma szerint (mindegyik az esetben a rendszeren belüli elemeket használunk). Nem mindig volt megadható kétféle deklaráció.

VALAMI, SEMMI

metanyelven:

fogalom(VALAMI),

¬ VALAMI=SEMMI

∀x(előfordulás(x) → előfordulása(x,VALAMI) ∧ ¬előfordulása(x,SEMMI)),

„Minden a **VALAMI** előfordulása, és a **SEMMI**nek nincs előfordulása”.

Minden a **VALAMI** terjedelmébe tartozik, ami van, létezik, de ez a rendszeren belüli elemekkel nem deklarálható, mert a „van”, „létezik” reláció nem eleme.

∃ yfajtája(y,VALAMI),

∀y(fogalom(y) ∧ ¬y= SEMMI ∧ ¬y= VALAMI → fajtája(y, VALAMI))

„A **SEMMI** és a **VALAMI** kivételével minden fogalom a **VALAMI** fajtája”.

tárgynyelven:

∃ xVALAMI(x) ∧ ∀XVALAMI(x)

„Minden a **VALAMI** előfordulása, és van előfordulása”.

¬ ∃ xSEMMI(x)

„A **SEMMI**nek nincs előfordulása”

ANYAGSZERŰ VALAMI

∀x(ANYAGSZERŰ_VALAMI(x) ↔ KONKRÉT_ANYAGSZERŰ_VALAMI(x) ∨ ELVONT ANYAGSZERŰ_VALAMI(x))

∀x(ANYAGSZERŰ_VALAMI(x) ↔ ∃ y(*tulajdonsága*(y,x) ∧ (ALAKZAT(y) ∨ AMORF(y)))

„Az anyagszerű valami vagy konkrét vagy elvont anyagszerű.”

„Az anyagszerű valami vagy amorf, vagy van alakja”.

ENERGIASZERŰ VALAMI

Míg az **ANYAGSZERŰ_VALAMI** fogalmát expliciten[25] egyszerűen tudtuk definiálni, az **ENERGIASZERŰ_VALAMI** esetében a helyzet bonyolultabb. Visszavezetem közvetlen fajtáinak definíciójára, azokat azonban csak impliciten (egymástól függően) tudjuk jellemezni.

∀x(ENERGIASZERŰ_VALAMI(x) ↔ MOZGÁS(x) ∨ HATÁS(x) ∨ ENERGIA(x))

∀x(MOZGÁS(x) ↔ ENERGIASZERŰ_VALAMI(x) ∧ ∃ y*oka*(y,x) ∧ ∀y(*oka*(y,x) → HATÁS(y)))

„Az energiaszerű valami vagy mozgás vagy hatás vagy energia.”

„A mozgás olyan energiaszerű valami, amelynek van oka, és ez csak hatás lehet”

[25] Explicit definíció azt jelent, hogy az A predikátumhoz van egy, a ∀x(A(x) ↔ φ(x)) sémájú formula, ahol φ(x)-ben nem szerepel A.

$\forall x(\text{HATÁS}(x) \leftrightarrow \text{ENERGIASZERŰ\_VALAMI}(x) \wedge \exists y \textit{hordozója}(y,x) \wedge$
$\forall y(\textit{hordozója}(y,x) \rightarrow \text{ENERGIA}(y)) \wedge \exists y(\textit{oka}(x,y) \wedge \text{MOZGÁS}(y)))$

„A hatás olyan energiaszerű valami, amelynek van hordozója, és a hordozója csak energia lehet, továbbá mozgást okoz”.

$\forall x(\text{ENERGIA}(x) \leftrightarrow \text{ENERGIASZERŰ\_VALAMI}(x) \wedge \exists y(\text{HATÁS}(y) \wedge \textit{hordozója}(x,y)))$.

„Az energia olyan energiaszerű valami, amely hatást hordoz”.

INFORMÁCIÓSZERŰ VALAMI

$\forall x(\text{INFORMÁCIÓSZERŰ\_VALAMI}(x) \leftrightarrow$
$\text{SZUBSZTANCIÁLIS\_INFORMÁCIÓSZERŰ\_VALAMI}(x) \vee$
$\text{AKCIDENTÁLIS\_INFORMÁCIÓSZERŰ\_VALAMI}(x) \vee$
$\text{ÉRTÉKELT\_INFORMÁCIÓSZERŰ\_VALAMI}(x))$
$\forall x(\text{INFORMÁCIÓSZERŰ\_VALAMI}(x) \leftrightarrow$
$\text{VALAMI}(x) \wedge$
$\exists y(\textit{tulajdonsága}(y,x) \wedge (\text{BELSŐ}(y) \vee \text{KÜLSŐ}(y) \vee \text{MINŐSÉG}(y) \vee \text{MENNYISÉG}(y)))$[26]

„Az információszerű valami vagy szubsztanciális, vagy akcidentális vagy értékelt információ.”

„Az információszerű valami vagy a valami belső vagy külső vagy minőségi vagy mennyiségi tulajdonsága”.

DOLOG

$\forall x(\text{DOLOG}(x) \leftrightarrow \exists y(\text{SZEMÉLY}(y) \wedge \textit{ismeri}(y,x))$

„A dolog olyan valami, melynek létezéséről tudomásunk van”. (Más megfogalmazásban: ”a dolog olyan valami, amelynek létezéséről tudunk”).

4 A DOLCE csúcsfogalmai

4.1 A csúcsfogalom (avagy a gyökér): thing, dolog, valami

A DOLCE csúcsfogalmainak kialakítására a hagyományos filozófiai elkötelezettség, és az extenzionális megalapozásra való törekvés jellemző. Ez utóbbi következtében egyrészt szinonimnak tekintenek számos olyan megnevezést, melyek látszólag azonos terjedelemmel rendelkeznek (például a TULAJDONSÁG, JELLEMZŐ, MINŐSÉG fogalmi hármasból csak a **MINŐSÉG** fogalmát használják fel[27]). Az információkereső nyelvek terminológiája alapján azt mondhatjuk, hogy ez utóbbit tekintik deszkriptornak, mely a másik kettő helyett is használandó. A köznapi gondolkodás felől nézve viszont a **MINŐSÉG** a metafora szerepét játssza a rendszerben. A PIROS a SZÍNES, tranzitíve a TULAJDONSÁG egyik fajtája. Minőség a piros csak akkor lehet, ha egy értékelési rendszerben (például a baloldali politikai nézetek rendszerében) a piros nagy megbecsülésnek örvend. Ekkor a PIROS a MINŐSÉG egyik fajtája lesz, de nem rigiden. A MINŐSÉG ebben az esetben a szerepfogalom helyzetében van.

---

[26] A BELSŐ, KÜLSŐ, MINŐSÉG, MENNYISÉG így szereplő kifejezések az ontológiában.

[27] A TULAJDONSÁG fogalmat univerzálénak tekintik, szemben a **QUALITY** fogalmával, mely felfogásuk szerint partikuláré [1]. Ez valójában önkényes döntés: a fogalmak természetes rendszerében például mind a TULAJDONSÁG, mind a MINŐSÉG univerzálénak számítanak.

A DOLCE legfelső szintű fogalma (a monohierarchikus[28] gráf gyökere) az ontológiák angol nyelven kifejezett generikus szerkezetében a **THING**. A **THING** itt valójában **SOMETHING**: jelentése nem csak dolog (azaz entitás), hanem akármi, minden, azaz nem csak az, amiről tudunk, hanem az is, amiről nem tudunk, de fogalmát gondolhatjuk (pl. KENTAUR). Manapság ontológiákat olyan ismeretbázisok céljaira igyekszenek felhasználni, melyekben elsősorban konkrét tárgyakra vonatkozó információkat kezelnek (nem pedig mondjuk olyan vallási információkat, melyek pl. a szentlélekre, varázslatra vonatkoznak). Ezért a számítástechnikus rendszertervezők ösztönösen a 'somtehing'-nél speciálisabb jelentésű 'thing' kifejezést választják.

Ráadásul az angol nyelvben a **SOMETHING** formálisan a **THING** fajtája, ahogy az a magyar tautológia is, hogy a **BÁRMILYEN VALAMI** a **VALAMI** fajtája. Ezért angolul kézenfekvő a **THING** csúcsfogalmi használata. Ez nem változtat azonban azon, hogy ebben a szerepében a **THING** fogalmának terjedelmébe bármi tartozhat, függetlenül attól, hogy ismerjük-e vagy sem, és ezért magyar ontológiákban csúcsfogalomként nem a **DOLOG**, hanem a **VALAMI** használandó.

THING
metanyelven:
fogalom(THING) $\forall$x(előfordulás(x) $\rightarrow$ előfordulása(x,VALAMI)
„Minden a **THING** előfordulása".
$\exists$ yfajtája(y,THING),
$\forall$y(fogalom(y) $\wedge$ $\neg$y= VALAMI $\rightarrow$ fajtája(y, THING))
„A **THING** kivételével minden fogalom a **THING** fajtája".

tárgynyelven:
$\exists$ XTHING(x) $\wedge$ $\forall$XTHING(x)
„Minden a **THING** előfordulása, és van előfordulása".

4.2 A THING közvetlen fajtái[29]

ENDURANT
A fogalom tartalma: térbelileg végesen létező [wholly present at any time]; időben készen, egészként létezés [enduring entity, sein in der Zeit]; passzívan/adottan nem maga a hatás, legfeljebb a hatásban részt vevő [aktor, participation].

Kváziszinimája[30]: **CONTINUANT**
A fogalom tartalma: maradandóan, térben folytonosan fennálló.

Fajtái: fizikai dolog (anyag, anyagszerűen megjelenő dolog, fizikai tárgy) [PHYSICAL ENDURANT], nem fizikai dolog (mentális objektum, szociális objektum) [NON-PHYSICAL ENDURANT], mesterségesen létrehozott összesség [ARBITARY SUM].

---

[28] Szemben a fogalmak természetes rendszerén alapuló ontológiával, a DOLCE generikus szerkezete monochierarchikus

[29] A fogalmak tartalmi elemzését a [3] alapján végeztem el.

[30] A DOLCE rendszerében ugyan teljesen ugyanazt értik az endurant és a continuant kifejezéseken, valójában azonban ezek mégsem tökéletes szinonimák. Az ilyen eseteket nevezik az információkereső nyelvek világában kváziszinonimáknak

PERDURANT

A fogalom tartalma: időben ˙(tova)terjedve létező [extend in time by different part]; időben történés [in time present, geschehen in der Zeit]; maga az aktivitás, hatás.

Kváziszinimája: **OCCURENCE**

A fogalom tartalma: esemény, eseményszerűség.

Fajtái: eset (a DOLCEban speciális a jelentése: minden **PERDURANT**, aminek logikus végpontja van) [EVENT], mozgás/változás/folyamat (minden **PERDURANT**, ami tart, folyik, fennáll) [STATIVE PERDURANT].

QUALITY

A fogalom tartalma: érzékelhető/észlelhető [perceive] vagy mérhető [measure], skálával rendelkező vagy más egyéb fogalmi térben elhelyezhető minőség. A minőség a DOLCE szerint partikuláré, a dolgokkal inherens, csak azokkal együtt létezik [qualities *inhere* to entities], nem azonos a tulajdonsággal, mely univerzálé.[1][3][4]

Fajtái: időleges minőség [TEMPORARY QUALITY], fizikai minőség [PHYSICAL QUALITY], elvont minőség [ABSTRACT QUALITY].

ABSTRACT

A fogalom tartalma: elvont.

Fajtái: tény (valójában: elvont tény) [FACT], halmaz, elvont összesség [SET], állapot-tartomány (valójában: idő, hely, állapot értéktartományát kifejező fogalmak) [QUALITY REGION].

Talán az ABSTRACT kivételében valójában nem csak az előző fejezetben már tárgyalt minőség, hanem a másik három csúcsfogalom használata is — gondolati szempontból — metaforikus, a rendszeralkotók szándéka felől nézve pedig speciális, szűkített. Ráadásul az **ENDURANT** és **PERDURANT** egyfajta tulajdonságot vagy állapotot jelent, a deklarációk szerint viszont fizikai tárgyak, ill. folyamatok a fajtáik. Ez olyan, mintha a KUTYA az ÁLLATI vagy az ÁLLATSÁG fajtája lenne, ami abszurd és alapvető ellentmondása a DOLCE fogalmi szerkezetének. Ezt nem teheti jóvá, hogy a rendszeren belül hogyan határozzák meg ezt a két fogalmat: tény, hogy az ontológia fogalmi rendszerének két legfontosabb csúcsfogalmát nem a generikus reláció alapján határozzák meg (mely relációt ugyanakkor az ontológia alaprelációjaként definiálják), hanem predikatív módon. Ezen az alapon azonban pl. a **THING** helyett akár az ESZME, akár az ANYAG lehetne a csúcsfogalom, filozófiai elkötelezettségtől függően. A generikus relációhoz való ragaszkodás azonban éppen azt biztosítja, hogy a filozófiai elkötelezettség ne legyen ennyire egyoldalúan idealista vagy materialista.

Ezzel szemben a természetes rendszer csúcsfogalmai (**ANYAGSZERŰ, ENERGIASZERŰ, INFORMÁCIÓSZERŰ VALAMI**) nem metaforák, hanem olyan megnevezések, melyek a „-szerű” toldalék segítségével nem metaforikus áttétellel, hanem explicite valamiféle, a Wittgensteini értelemben vett családhasonlóságot[31] fejeznek ki. Mintegy generikusan összefogják mindazokat a fogalmakat, melyeket az adott „-szerűség” tart össze. A **VALAMI** pedig legalábbis az idealista és materialista elkötelezettségtől független legfelső csúcsfogalom.

[31] Family resemblance, cluster definition [2][15, §66–67, 69, 76–78]

4.3 A DOLCE csúcsfogalmainak logikai deklarációi

A csúcsfogalmakat kétféle formában írjuk le: alárendeltjeikkel (ebben az esetben a DOLCE rendszerén belül maradtunk), ill. a természetes fogalmak deklarációi szerint (ebben az esetben a DOLCE rendszerén kívüli elemeket is használunk).[32]

ENDURANT

$\forall x(\text{ENDURANT}(x) \leftrightarrow \text{PHYSICAL\_ENDURANT}(x) \vee \text{NON-PHISICAL\_ENDURANT}(x))$

$\forall x(\text{PHYSICAL\_ENDURANT}(x) \leftrightarrow \exists y(\textit{tulajdonsága}(y,x) \wedge (\underline{\textit{FORM}}(y) \vee \underline{\textit{AMORPH}}(y) \vee \text{FEATURE}(y)))$

„Az **ENDURANT** vagy fizikai vagy nem fizikai."

„A PHISICAL ENDURANT vagy amorf, vagy van formája, vagy feature".

A $\underline{\textit{FORM}}$ és az $\underline{\textit{AMORPH}}$ rendszeren kívüli elem, melyet más írással jelöltünk.

Csak a példa kedvéért a DOLCE szerinti leírás:

$\forall x(\text{ENDURANT}(x) \leftrightarrow \exists y,t(\textit{participates}(x,y,t))$. Azaz: „Van egy idő, amikor valaminek a résztvevője."

A természetes rendszerben az idő mind az **ANYAGSZERŰ VALAMI** , mind az **ENERGIASZERŰ VALAMI** tulajdonsága:

$\forall x((\text{ANYAGSZERŰ\_VALAMI}(x) \wedge (\text{ANYAGSZERŰ\_VALAMI}(x)) \leftrightarrow$
$\exists y,t(\textit{tulajdonsága}(x,y,t))$

PERDURANT

$\forall x(\text{PERDURANT}(x) \leftrightarrow (\text{EVENT}\ (x) \vee\wedge\ \text{STATE}\ (x) \vee \text{PROCESS}\ (x))$

„A **PERDURANT** vagy EVENT, vagy STATE vagy PROCESS".

$\forall x(\text{PERDURANT}(x) \leftrightarrow \text{THING}(x) \wedge \exists y\textit{oka}(y,x) \wedge \forall y(\textit{oka}(y,x) \rightarrow \underline{\textit{HATÁS}}(y))$

„A **PERDURANT** olyan ***THING***, amelynek van oka, és az oka csak hatás lehet."

A $\underline{\textit{HATÁS}}$ rendszeren kívüli elem, melyet más írással jelöltünk.

QUALITY

$\forall x(\text{QUALITY}(x) \leftrightarrow \text{TEMPORAL\_QUALITY}(x) \vee \text{PHYSICAL\_QUALITY}(x) \vee \text{ABSTRACT\_QUALITY}\ (x))$

$\forall x(\text{QUALITY}(x) \leftrightarrow \text{THING}(x)\ \exists y(\textit{tulajdonsága}(y,x) \wedge (\text{TIME}(y) \vee \text{SPCE}(y) \vee \text{ABSTRACT}(y) \wedge \exists y(\underline{\textit{hordozója}}(y,x) \wedge \forall y(\underline{\textit{hordozója}}(y,x) \rightarrow \text{PHYSICAL ENDURANT})$

„A **QUALITY** vagy temporális vagy fizikai vagy elvont."

„A **QUALITY** vagy a **THING** időbeli vagy térbeli vagy elvont tulajdonsága, vagy hordozója van és a PHYSICAL ENDURANT".

A $\underline{\textit{hordozója}}$ rendszeren kívüli reláció, melyet más írással jelöltünk.

A DOLCE a szerepfogalmakat (DOLOG, ESET, ESEMÉNY, TÖRTÉNÉS, ELŐFORDULÁS, TÁRGY, JELENSÉG, ALANY) nem tartalmazza. A DOLCE EVENT fogalma szűkebb értelmű, a logikus végponttal rendelkező eseményeket, folyamatokat, tevékenységeket jelenti. A DOLCE **ABSTRACT** és a természetes rendszer **ELVONT VALAMI** fogalmát itt nem vizsgáltuk.

[32] A deklarációkat nem a DOLCE kialakítói által közzétett formában [4] használjuk fel, mert ez megnehezítette volna az összehasonlítás.

5 A csúcsfogalmak két rendszerének összehasonlítása

5.1 Elemzés

A **VALAMI** és a **THING** azonossága a 4.1 fejezet logikai deklarációja alapján triviális.

Az alattuk levő hierarchiaszint csúcsfogalmainak összehasonlításához a csúcsfogalmaknak a fenti logikai leírásokból következő fajtáit vesszem alapul.

Az összehasonlítást a 2. táblázat tartalmazza.

Az **ANYAGSZERŰ VALAMI** és az **ENDURANT** lényegében csak a FEATURE egyes fajtáiban tér el egymástól. Ezek egy része — pl. nyílás, szakadék — a természetes rendszerben ugyancsak anyagszerű valami, másik része — pl. határ, felület — viszont információszerű valami.

Az **ENERGIASZERŰ VALAMI** és a **PERDURANT** lényegében csak a state (pl. ülve levés, nyitva levés, boldognak levés, pirosnak levés) esetében térnek el egymástól. A természetes rendszerben ezek legnagyobb része információszerű valami.

A legnagyobb eltérés az **INFORMÁCIÓSZERŰ VALAMI** és a **QUALITY** esetében áll fenn. A fogalomtartalmak alapján azonban az eltérés látszólagos, mivel egyrészt arról van szó, hogy ugyanannak a fogalomnak mások a felosztási szempontjai a két rendszerben, másrészt a DOLCE kialakítói ebben a fogalmi tartományban hagyják a leginkább figyelmen kívül a természetes nyelven megnevezett fogalmakat: az ÁLLAPOT[33], MINŐSÉG, JELLEG, TULAJDONSÁG, JELLEMZŐ helyett a DOLCE csak a QUALITY fogalmát tartalmazza.

2. táblázat. Az ismertetőjegyek és fajták összehasonlítása

ANYAGSZERŰ VALAMI	**ENDURANT**
AMORF ANYAG	AMOUNT OF MATTER
TÁRGY	PHYSICAL OBJEKT
–	FEATURE
ENERGIASZERŰ VALAMI	**PERDURANT**
MOZGÁS	PROCESS
	STATE
HATÁS	–
ENERGIA	–
INFORMÁCIÓSZERŰ VALAMI	**QUALITY**
SZUBSZTANCIÁLIS INFORMÁCIÓSZERŰ V	
AKCIDENTÁLIS INFORMÁCIÓSZERŰ V.	
ÉRTÉKELT INFORMÁCIÓSZERŰVALAMI	
	TEMPORAL QUALITY
	PHYSICAL QUALITY
	ABSTRACT QUALITY

Valójában a DOLCE szinte minden **QUALITY** fogalma fajtája a természetes rendszer információszerű valami fogalmának. Ezt mutatják a folytonos egyirányú gráfélekkel jelölt fajtája relációk.

[33] A természetes rendszer ÁLLAPOT fogalmának a DOLCE QUALITY REGION fogalma felel meg, amely fajtája a QUALITY fogalmának.

Fordítva ez nem igaz: az **INFORMÁCIÓSZERŰ VALAMI** fogalmának fajtái közül számos a DOLCE rendszerében a FEATURE, illetve a STATE fogalmához, és rajtuk keresztül az **ENDURANT** és a **PERDURANT** fogalmához kapcsolódik. Azaz a FEATURE fajtáinak egy része az **INFORMÁCIÓSZEREŰ VALAMI**, másik része az **ANYAGSZERŰ VALAMI** fogalmának fajtája, a STATE fajtáinak kisebb része pedig az **ENERGIASZERŰ VALAMI**, nagyobb része az **INFORMÁCIÓSZERŰ VALAMI** fajtája. Ezt mutatják a szaggatott gráfélek, a vastagabb éllel jelölve a több kapcsolódó fajfogalmat.

A DOLCE négy csúcsfogalmát elemezve kimutatható, hogy a két fogalmi rendszer között az elkötelezettségek rendkívül eltérő volta ellenére mélyreható megfelelés áll fenn.

5.2 Összegezés

Noha első látásra kétségtelen eltérések tapasztalhatók a két rendszer között, mégis: az eltérések mértéke nincs arányban azzal, hogy két gyökeresen különböző szemlélet — és nem utolsó sorban rendeltetés — alapján kialakított fogalmi rendszerről van szó.

A két rendszer a legfelső fogalmi szinteken valójában nagyon jól összehasonlítható és az alapvető fogalmi hasonlóságok szembeötlők.

Mindez arra utal, hogy a fogalmi rendszer, s vele a fogalomalkotás, a gondolkodás mélyén létezik egy közös, a formalizálás által még el nem ért, sajátos mélyszerkezet.

Mivel az 'endurant' és a 'perdurant' kifejezések magyarra jószerint lefordíthatatlanok, kézenfekvő helyettük az 'anyagszerű valami' és az 'energiaszerű valami' kifejezéseket használni, adott esetben azzal a megkötéssel, hogy a DOLCE szerinti **ENDURANT**, ill. **PERDURANT** értelmében használjuk.

Köszönetnyilvánítás

E helyen köszönöm Szőts Miklósnak azt a segítséget, melyet a vele folytatott beszélgetésekből nyertem és a logikai leírások dolgában adott. A tanulmány a Magyar Egységes ontológia (MEO) NKFP-2/42/04. sz. projekt keretében késült el.

Bilbiográfia

1. DOLCE. Descriptive Ontology for Linguistic and Cognitive Engineering. <http://www.loa-cnr.it/DOLCE.html>
2. Gabriel, G.: Familienähnlichkeit. In: Mittelstraß (Hrsg.): Enzyklopädie Philosophie und Wissenschaftstheorie, 2. Aufl. (2005), 473 f
3. Gangemi, A. [et al.]: Sweetening Ontologies with DOLCE. <http://www.loa-cnr.it/Papers/DOLCE-EKAW.pdf>
4. Masolo, C [et al]: WonderWeb Deliverable D18.Ontology library (final) <http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf>
5. Szakadát I.: MEO. Magyar Egységes Ontológia. NKFP-2/042/04.sz. projekt In: W3C Szemantikus web. Műhelykonferencia. <http://www.w3c.hu/rendezvenyek/2006/szemweb/eak/bmemokk_syi.pdf>

6. Szőts M., Lévay Á.: Szerepfogalmak az ontológiákban — az OntoClean metodológia továbbfejlesztése. In: Magyar Számítógépes Nyelvészeti Konferencia. 2005. Szeged (2005) 56–67
7. Ungváry R.: Application of the thesaurus method to the communication of knowledge. In: International Classification (1983) No. 2. 63–68
8. Ungváry R.: Ein natürliches System der Gegenstände. – Anwendung der Klassifikation. Proc. der 8. Jahrestagung der Ges. für Klassifikation, Hofgeismar, 10–13. April 1984. Frankfurt/Main : Indeks Verlag (1985) – (Studien zur Klassifikation ; Bd. 15) 19–41
9. Ungváry R.: Über den Begriff des Bildes. In: Photogeschichte. Beiträge zur Geschichte und Ästhetik der Fotografie (1987) 26. Heft 57–63
10. Ungváry R.: A számokról**.** In: Café Bábel (2000) 1. 3–15
11. Ungváry R.: Tezaurusz és ontológia, avagy a fogalmi ismertetőjegyek generikus öröklődésének formalizálása. In: Tudományos és Műszaki Tájékoztatás (2004) 5. sz. 175–191
12. Ungváry R.: A kategóriák rendszere (2004) <http://ontologia.hu/document/paper/>
Ungváry R.; Radnai T.: Thesaurus in user interface. Optimum presentation of thesauri. In: IEEE 3rd International Conference on Computational Cybernetics, 2005. april 13–16. Proc. Mauritius (2005) 175–180
13. Ungváry R.: A tezaurusz mint „kisvilág”. 2006. <http://ontologia.hu/document/paper/>
14. Wittgenstein, L.: Philosophische Untersuchungen. Kritisch-genetische Edition. Schulte, J. (Hrsg.) Frankfurt am Main (2001)

Igei wordnet és igei eseményszerkezet ábrázolása

Kuti Judit[1], Varasdi Károly[1], Cziczelszki Judit[1], Gyarmati Ágnes[1],
Nagy Anikó[1], Tóth Marianna[1], Vajda Péter[1]

[1] Magyar Tudományos Akadémia, Nyelvtudományi Intézet
1068 Budapest, Benczúr u. 33.
{kutij,varasdi,judit,aagnes,nagya,masa,vajda}@nytud.hu

Kivonat: A magyar igei wordnet készítése során az igei jelentések viszonyainak ábrázolásához szükséges kibővíteni a Princeton WordNet által főleg a főnevekre kidolgozott relációstruktúrát. Az igéket eseményszerűségekként vizsgáltuk, és számba vettük a magyarban az igei jelentések meghatározásához fontos aspektuális információkat. Bemutatjuk Moens és Steedman eseménynukleusz fogalmát, melyet felhasználva aspektuális szempontból csoportosítottuk a magyar igei synsetek egy részét, így pszicholingvisztikailag releváns információkat is ábrázolni tudtunk. Az aspektuális információk jelölésére a wordnet struktúrájába illő új relációkat vettünk fel. Cikkünkben megemlítjük a magyar igei wordnet számítógépes nyelvészeti alkalmazási lehetőségeit is.

1 Bevezetés

Tanulmányunkban a Magyar WordNet (továbbiakban "HuWN")[34] igei részének készítése során felmerült néhány olyan nyelv- és szófajspecifikus kérdésre, problémára térünk ki, melyekre a munka alapjául szolgáló wordnetek[35] nem ajánlottak kielégítő megoldást, és bemutatjuk, hogy az ezekre adott megoldásaink mily módon viszonyulnak a szabvánnyá vált wordnet alap(ok)hoz. Mivel a wordnet műfaja a főnévi szófaj jellegzetesen hierarchikus viszonyaira épül, a már elkészült wordnetekben egyrészt mennyiségileg a főnévi rész dominál, másrészt annak relációtípusai adtak mintát az igei relációk felvételéhez is. Egy, a főnévi hierarchiát az igei szófajra leképezni próbáló jelentésábrázolás azonban nem bizonyulhat elegendő keretnek az igék által kifejezett jelentések szemantikai viszonyainak meghatározásához, hiszen ehhez elengedhetetlen az igék eseményszerkezetének megvizsgálása. Tanulmányunkban tehát azt mutatjuk be, hogy a magyar igei wordnetben hogyan valósítjuk meg bizo-

[34] Az adatbázis a 2005 tavaszán indult Magyar Ontológia Építése projekt keretén belül készül, mely a Szegedi Tudományegyetem, a MorphoLogic Kft. és a Nyelvtudományi Intézet közös projektuma (GVOP – 2004 – 3.1.1.). A Nyelvtudományi Intézet a Magyar Wordnet igei részének elkészítését vállalta magára. Ezúton is szeretnénk köszönetet mondani a projekt támogatásáért.

[35] Princeton Wordnet (PWN) [3], BalkaNet [7], EuroWordNet [10]

nyos, az ige eseményszerkezetéből adódó, szemantikai relációkat meghatározó információk tárolását, ábrázolását a wordnet műfaja által lehetővé tett relációs keretek között.

Eddigi tapasztalataink szerint igei jelentések meghatározásához, elkülönítéséhez a magyarban szükséges figyelembe venni aspektuális tulajdonságokat is. Ezért a következő részben néhány, az ige eseményszerkezetével kapcsolatos és aspektuális szempontból releváns megállapítás után felvázoljuk Moens és Steedman *nukleusz* nevű eseménystruktúráját. A továbbiakban megmutatjuk, hogy a nukleusz fogalmának felhasználásával egy olyan eszköztár kerül a kezünkbe, amelynek segítségével

- könnyebben helyezhetők el a magyarban lexikalizált jelentések a wordnet hálójában,
- olyan, pszicholingvisztikailag releváns kapcsolatokat tudunk ábrázolni jelentések között, amelyek eddig hiányoztak a magyar WordNet alapját képező wordnetekből, és
- a HuWN esetleges későbbi számítógépes nyelvészeti felhasználása során is hasznosítható információk kerülnek kódolásra.

2 Eseményszerűségek[36] és aspektuális jellemzőik

2.1 Logikai következtetések igei jelentések között

Mivel a magyarban bizonyos igekötők gyakran morfológiailag jelölik az ige által kifejezett eseményszerűség aspektusát, ám a HuWN alapjául szolgáló Princeton WordNet az angol nyelv tipológiai különbözősége miatt nem tartalmaz aspektuális információkat, illetve, mint említettük, az alapvető igei reláció, a hipo-hipernima reláció is főnévi mintára készült, szükséges megvizsgálnunk, hogy a wordnet relációinak alapját képező logikai implikációs viszony milyen módon áll fenn igék között. A wordnet módszertana szerint ugyanis a vizsgált morfémák közötti szemantikai viszonyokat bizonyos logikai következtetések elvégezhetőségén keresztül kell meghatározni. Míg azonban például a főnevek esetében a *"minden X-re igaz, hogy: X (egy) N1, tehát X (egy) N2"* séma igazolásán keresztül az *N1* < *N2* (N1 hiponimája N2-nek) összefüggés belátható, az igei morfémák ilyen egyszerű rendezésére nincs mód. A logikai kapcsolatok ugyanis csak teljes (potenciálisan igazságértékkel bíró) kijelentések (illetve az azokat kifejező mondatok) között állapíthatók meg, ám a mondatok logikai szerkezetét éppen az igék alakítják ki módosítóikkal, vonzataikkal együtt.[37] Az ige-vonzat viszony azonban erősen aszimmetrikus: a mondat logikai potenciáljá-

[36] Az *eseményszerűség* terminust az Emmon Bach által bevezetett *eventuality* terminus megfelelőjeként használjuk.

[37] Amikor igék aspektuális tulajdonságairól beszélünk, valójában igei frázisokat kéne említenünk, hiszen az igék önmagukban alulspecifikáltak erre az információra nézve (ld. pl. [9].

ért az ige a felelős; a vonzat ennek csupán többé-kevésbé passzív szereplője.[38] Ehhez csatlakozik az a nehézség is, hogy az adott mondattal elvégezhető következtetések nagyban függnek mind az igeidőtől, mind pedig az aspektustól.

Az utóbbi jelenséget például a következő példázza: Míg abból a mondatból, hogy *Mari éppen ment át az utca túloldalára, amikor megpillantotta Bélát* nem következik, hogy Mari át is ment az utca túloldalára (vissza is fordulhatott, hogy Bélát üdvözölje), addig abból a mondatból, hogy *Mari átment az utca túloldalára, amikor megpillantotta Bélát* egyértelműen következik, hogy Mari befejezte a megkezdett cselekvését (sőt, még az a pragmatikai implikatúra is jelen van, hogy Mari pontosan azért hajtotta végre a cselekvést, mert Bélát megpillantotta). A két főmondat pusztán aspektuális értékében[39] különbözik: az első esetben progresszív (imperfektív, folyamatos), a második esetben perfektív (befejezett) szemléletű mondattal állunk szemben, és a mondatok eltérő logikai tulajdonságáért éppen ez a felelős.[40] Nyilvánvaló tehát, hogy az a kérdés, hogy az "át+megy" igekötő-ige pár milyen következtetésekben vehet részt, nem választható el attól, hogy milyen aspektuális értékkel szerepel a vizsgált mondatokban. Az (igekötős) ige aspektuális lehetőségei tehát – és vele együtt azon következtetések összessége, amelyben az igével képzett mondat részt vehet – az ige által kifejezett (néhány elemi típusba sorolható) eseménystruktúra függvénye, melyekről alább részletesen is lesz szó. Ezeket az (eseményontológiai) típusokat a nyelvelmélet számára is használható módon először Vendler Zénó [8] határozta meg, majd az ő osztályozását fejlesztette tovább Emmon Bach (ld. [1]), illetve a számítógépes nyelvészet területén Marc Moens és Mark Steedman [6].

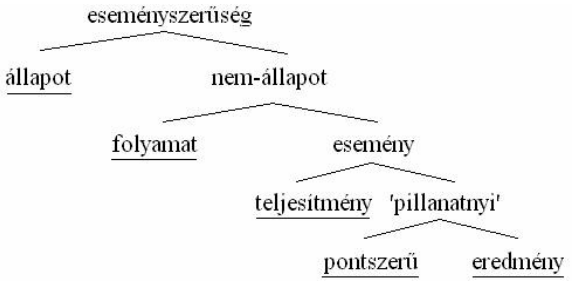
2.2 A Vendler- és Bach-féle aspektuális osztályok

Az eseményszerűségek vendleri felosztása négy aspektuális osztályt különböztet meg arra vonatkozólag, hogy az ige által kifejezett esemény milyen belső időszerkezetet fejez ki. A vendleri osztályozás szerinti négy fő eseménytípus vonzatokkal és kontextussal kiegészülve különböző aspektusokat vehet fel: a folyamatok (pl. úszik) tipikusan a progresszív aspektust, a teljesítmények (pl. kimegy (a szobából)) a progresszív és a perfektív aspektus, az eredmények (pl. kidurran), pedig tipikusan a perfektív aspektust. Az állapotok osztálya sem a progresszív, sem a perfektív aspektust nem tudja felvenni. A Bach által továbbfejlesztett és kiegészített felosztás egy bináris rendszerben elhelyezve az 1. ábrán látható módon jeleníti meg az aspektológiai kategóriákat, kiemelve olyan pontszerű események meglétét, amelyek elkülönülnek az eredményektől (pl. *kattan*):

[38] A mondataspektus kialakításában tárgyas ige esetén az ige tárgya is részt vesz. Azonban a tárgy hatása az aspektusra meglehetősen jól megjósolható az ige eseményszerkezetéből és a tárgy sajátosságaiból kiindulva, így azzal a wordnet keretei között külön nem kell foglalkozni.

[39] Az aspektus fogalmát Kiefer (2000)-rel (ld. [4]) összhangban értelmezzük.

[40] A fenti jelenséget a szakirodalom egyébként az "imperfektív paradoxon" néven ismeri, ld. [2].

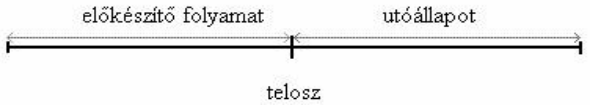


1. Ábra Eseményszerűségek Bach-féle aspektuális felosztása

A négy vendleri alosztályt az a tulajdonság is jellemzi, hogy időtartományuk osztható-e avagy sem – azaz, hogy az adott ige által denotált eseményszerűség az időtartomány legtöbb résztartományára is érvényes-e. Ennek értelmében a *folyamatok* (activity), *állapotok* (state), *teljesítmények* (accomplishment) és *eredmények* (achievement) közül az első kettő homogén eseményszerűségnek tekinthető, hiszen olyan predikátumok fejezik ki őket, melyeknek tetszőleges részei maguk is ugyanazon predikátum által leírhatók. A *teljesítmények* és *eredmények* ellenben e tekintetben heterogén eseményszerűségek: különböző jellegű eseménykomponensek koherens egységei. A *pontszerű* eseményeket szintén nem komplex eseményszerűségeknek tekintjük.. A magyar igei wordnet elkészítésének szempontjából a komplex eseményszerűségeknek – *eredmények*, *teljesítmények* – az aspektuális jellemzőihez hű ábrázolása érdekes. Ezeknek komplexitását a Moens és Steedman által bevezetett nukleusz struktúra segítségül hívásával értelmezhetjük.

2.3 Moens és Steedman eseménynukleusza

Moens és Steedman [6] az eseményszerűségeknek egy, a vendleri [8] felosztáson alapuló, de annál finomabb rendszerét vezeti be, melynek központi fogalma az *(esemény)-nukleusz,* más néven *triád.* A *triád* név utal arra, hogy egy idealizált eseményszerűség potenciálisan három összetartozó komponensből áll: *előkészítő folyamat*, *telosz/sikerpont*[41], és *utóállapot*.



2. Ábra A Moens és Steedman-féle eseménynukleusz / triád

Az eseménytriádot rendezett hármasként is lehet ábrázolni:

<a, b, c>

ahol a = ELŐKÉSZÍTŐ FOLYAMAT, b = TELOSZ, c = UTÓÁLLAPOT. Moens és Steedman ezt az esemény-alapegységet a lexikalizált igei szófajjal megjelenített ver-

[41] Moens és Steedman a *telosz* kifejezés helyett gyakrabban használja a *kulmináció* (*culmination*) kifejezést. Mi azonban az előbbi terminus mellett maradunk, annak érdekében, hogy a *sikerpontot* megkülönböztessük a Moens és Steedman által szintén kulminációnak nevezett vendleri *eredményektől* (ld. 2. Táblázatot).

bális nyelvi szint fölött helyezi el. Az esemény-nukleusz egyes komponenseit tehát nem egyes nyelvekben lexikalizált elemek, hanem metanyelvi elemek töltik ki[42].

metanyelvi szint	<a, b, c>
nyelvi szint	lexikalizált nyelvi elemek, (synsetek)

3. Ábra Az eseménynukleusz metanyelvi és nyelvi szinten való megjelenése

A három nukleusz-eseménykomponens[43] egységként való kezelésének létjogosultságára a következő magyarázat adható: Lévén, hogy aspektológiai szempontból vizsgáljuk az eseményszerűségeket, releváns információnak kell tekintenünk, hogy ha egy-egy tetszőleges lexikalizálódott nyelvi kifejezést aspektológiai jellemzőkre érzékeny nyelvi tesztekkel[44] (a magyarban a progresszív és a perfektív aspektus tesztjeivel) vizsgálunk, legfeljebb a fent leírt három komponens együttes megléte mutatható ki. Egy adott eseményszerűséget leíró igéről aspektuális szempontból tehát az mond el valamit, hogy ezen fölöttes esemény-komponensek közül melyeket tölti ki *konceptuálisan*. A *„kimegy a szobából”* igei frázissal megnevezett eseményszerűség példáján bemutatva:

A progresszivizálhatóság az első komponens meglétét teszteli. Egy kifejezés ugyanis akkor és csak akkor elfogadható progresszívben, ha a hozzárendelt triád első komponense konceptuálisan ki van töltve. A triád első komponensének konceptualizáltságát tesztelhetjük a következő mondattal:

János éppen ment ki az épületből, amikor találkoztam vele.

A perfektiválhatóság a harmadik komponens meglétét teszteli, amely gyakorlatilag együtt jár a második komponens meglétével (ld. [6]).A magyar nyelv sajátosságai miatt a magyar mondat angolra való fordításával, és Present Perfect igeidőbe való áttételével tesztelhetjük legkönnyebben, hogy egy magyar igei jelentés mögött konceptualizálódott-e a triád második, ill. harmadik komponense.[45]

By the time Sue arrived, John has gone out of the building.

A két teszt eredményeképp megállapíthatjuk, hogy a *kimegy az épületből* frázis tehát mindhárom triád-komponenst konceptualizálja:

kimegy az épületből <MEGY A KAPU FELÉ, ÁTLÉP A KAPUN, KINT VAN AZ ÉPÜLETBŐL>

[42] Mivel ezekre csak nyelvi elemekkel tudunk utalni, jelölési konvencióként a kis kapitálisokkal való írásmódot fogjuk alkalmazni, hogy elkülönüljenek a kurzívval szedett, nyelvi elemektől.

[43] Az eseménykomponenseket most nyelvi lexikalizáltságuktól függetlenül tekintjük.

[44] A perfektivizálás és progresszivizálás lehetséges tesztjeről ld pl. [4]. Az itt alkalmazott két teszt részben Kiefer (2000)-n, részben saját elgondoláson alapul.

[45] Az angol fordítás ellenére a magyar anyanyelvünk elég erős befolyásoló tényező ahhoz, hogy paradox módon az angol nyelvi mondat helyességét is magyar nyelvi intuíciónk alapján ítéljük meg, s így, bár angol nyelvű teszttel, de a magyar jelentésről nyerjünk információt.

Moens és Steedman a triád komponenseinek a meglétével ill. hiányával pontosítja a vendleri ill. bachi rendszerben már megnevezett kategóriákat. Hogy lássuk hogyan viszonyul a triádok által kialakított osztályozás a Vendler-féle osztályokhoz, figyeljük meg az 1. táblázatot. Ebben a Moens és Steedman rendszerében használt szempontok (atomi ill. kiterjedt és utóállapottal rendelkező, ill. anélküli eseményszerűségek) szerinti csoportosítás látható, mely egyben explicit módon utal a vendleri és Bach-féle aspektuális osztályokkal való megfelelésre (ahol az újabb terminológia eltér a hagyományostól, ott zárójelben megadjuk a régebbit).

1. Táblázat:
Eseményszerűségek felosztása Moens és Steedman rendszerében

	"nem-állapot"		**állapot**
	atomi	kiterjedt	
+utóállapot	**kulmináció** (= EREDMÉNY) *felismer, megpillant, megnyeri a versenyt*	**kulminált folyamat** (= TELJESÍTMÉNY) *épít egy házat, felmegy a hegyre*	*ismer, szeret, tud, hasonlít*
– utóállapot	**pontszerű esemény** *vakkant, dörren, koppan, megbotlik*	**folyamat** *fut, úszik, sétál, zongorázik*	

A nukleusz eseménykomponenseinek összetartozása több, mint puszta időbeli egymásutániság; az okozáshoz, és a lehetővé tételhez hasonlatos, de azokkal nem megegyező viszony, melyet Moens és Steedman *contingency*-nek nevez. A három nukleusz-komponens kölcsönös dependenciája miatt egyiket sem lehet saját jogán előkészítő *folyamatnak*, *sikerpontnak* illetve *utóállapotnak* tekinteni. Az olyan eseményszerűség, amely a fenti tesztek alapján látszólag rendelkezik *előkészítő folyamattal*, ám nincs sem *sikerpontja*, sem *utóállapota*, (jelölése lehetne <a, ∅, ∅>) nem tekinthető *előkészítő* folyamatnak, hiszen nem készít elő semmit. Az ilyen eseményszerűség a vendleri *folyamat* kategóriának feleltethető meg. Az olyan eseményszerűség, amely a fenti tesztek alapján látszólag rendelkezik *utóállapottal*, ám nem rendelkezik *sikerponttal* (jelölése lehetne <∅, ∅, c), értelemszerűen nem lehet *utó*állapot, hanem a vendleri értelemben vett *állapotnak* tekintendő. Az olyan eseményszerűség pedig, mely *előkészítő folyamat* és *utóállapot* nélküli *sikerpontnak* tűnik, szintén nem nevezhető annak, hanem a bachi értelemben vett *pontszerű esemény*nek kell tekintenünk (jelölése lehetne <∅, b, ∅>). Összefoglalva tehát elmondhatjuk, hogy a triád kitöltött *utóállapota* implikálja egy *sikerpont* meglétét, a kitöltött sikerpont pedig implikálja egy *előkészítő folyamat* meglétét.

A nukleusz egyes komponenseinek konceptuális kitöltöttsége szerint tehát potenciálisan 2^3 számú különböző aspektuális kategória különböztethető meg. Ezekből az imént említett hármat (folyamat, pontszerű esemény, állapot), mint nem komplex eseményszerűséget, Moens és Steedman nem tárgyalja részletesen, mi azonban foglalkozunk velük a HuWN-ben. A lehetséges kombinációs lehetőségek közül további hármat világismereti okokból kizárhatunk: A mindhárom komponens kitöltetlenségét jelentő kombináció sem nyelvileg, sem konceptuálisan nem lehet releváns számunkra; a kizárólag előkészítő folyamattal és telosszal rendelkező eseményszerűség (jelö-

lése lehetne <a, b, ∅>), csakúgy mint az *előkészítő folyamattal* és *utóállapottal* rendelkező eseményszerűség (jelölése lehetne <a, ∅, c>), a *telosz* és az *utóállapot* összetartozása miatt nem lexikalizálódhat.

3 Az eseménynukleusz fogalmának felhasználása a HuWN-ben

dv) Moens és Steedman nukleusz-strukúrájának segítségével tehát nyelvi lexikalizált kifejezések mögött lévő metanyelvi elemek konceptuális meglétét illetve hiányát állapíthatjuk meg. Az az információ, hogy egy ige egy idealizált komplex esemény-egységhez képest hány elemet konceptualizál, nem más, mint az eseményszerűség *telikusságára*, ill. *atelikusságára* vonatkozó információ. Amennyiben egy ige által leírt esemény-nukleusz harmadik komponense ki van töltve[46], *telikus* eseményszerűségről, amennyiben nincs, *atelikus* eseményszerűségről beszélünk.

3.1 A telikusság jelölése a HuWN-ben

12. Ha a telikusság szempontjából próbáljuk meg áttekinteni a nukleusz egyes komponenseinek kitöltöttsége szerinti öt, nyelvileg is realizált lehetséges mintázatot, a 2.3-ban bevezetett rendezett hármasokat idéző ábrázolásmód tűnik alkalmasnak:

2. Táblázat:
Komplex eseményszerűségek telikussága Moens és Steedman nukleusz-struktúrájának segítségével ábrázolva

a triád komponenseinek kitöltöttsége	*a triádot lexikalizáló frázis konceptualizált komponenseinek metanyelvi megnevezése*	*az igei frázis telikussága*
<a, b, c>	*kimegy az épületből:* <MEGY A KAPU FELÉ, ÁTLÉP A KAPUN, KINT VAN>	+ utóállapot → telikus
<∅, b, c >	*kidurran:* <∅, A KIDURRANÁS PILLANATA, KIDURRANT ÁLLAPOT>	+ utóállapot → telikus

13. Az egyszerű eseményszerűségek (*folyamatok*, *állapotok*, *pontszerű események*) közül a *folyamatokat* és az *állapotokat* atelikusnak szokás tekinteni, a *pontszerű események* pedig kontextus nélkül, önmagukban alulspecifikáltak erre az információra nézve.

A HuWN készítésekor arra az igei jelentések rendszerezése során visszatérő kérdésre, hogy kódoljuk-e, és ha igen, milyen formában, a csupán aspektuálisan eltérő, a wordnet szinonima fogalma alapján azonban szinonimnak értékelendő jelentéseket, az eseménynukleusz fogalmának felhasználásával igennel válaszolhatunk. Amennyiben egy-egy, a wordnetben synsetként szereplő jelentést minimális propozícióvá alakítunk, meg tudjuk mondani róla, hogy az általa lexikalizált nukleusz utóállapota

[46] Amint az előző pontban említettük, a harmadik eseménykomponens kitöltöttsége komplex eseményeknél magával vonzza a második komponens kitöltöttségét is.

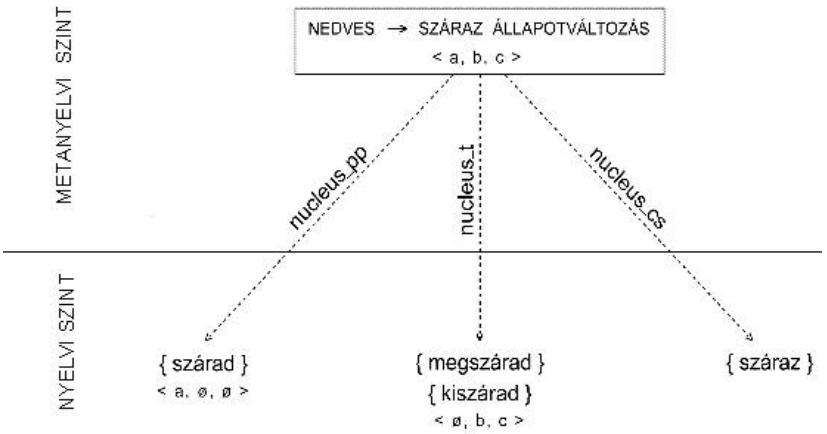
ki van-e töltve, vagy sem.[47] Azáltal, hogy egy jelentésről kódoljuk, hogy van-e utóállapota (és ezáltal telosza is) vagy sem, és hozzárendelünk egyet a triád-komponensek konceptualizáltságának megfelelő öt minta közül, egyértelműen leolvasható lesz az ige által kifejezett eseményszerűség telikus ill. atelikus volta. Ezt az információt a wordnetben a synsetekben tároljuk, oly módon, ahogyan maga a vonzatkeret-információ is tárolásra kerül: megjelöljük, hogy a triád három komponense közül melyek konceptualizálódtak a magyarban.

A három, a 2.2 és 2.3. alatt említett egyszerű eseményszerűség esetében is, ahogy fentebb utaltunk rá, az átláthatóság és egységesség kedvéért megtartjuk az aspektuális információ rendezett hármasként való jelölési konvencióját. Ennek megfelelően pl. a {fut} synset esetében a rendezett hármasként való jelölésből (<a, □, □>), ahol '□' a nem konceptualizálódott triád-elemekre utal, leolvasható az az információ, hogy az eseményszerűség atelikus, de leolvasható az is, hogy az eseményszerűség egy vendleri folyamat, amire az önmagában betöltött előkészítő folyamat pozíció utal.

3.2 Komplex eseményszerűségek az igei HuWN-ben

Azon túl, hogy egyes igei synsetekben az adott jelentésre vonatkozó minimális aspektuális információt tárolunk, a wordnet relációs struktúrája és az eseménynukleusz egységként való felfogása megengedi, hogy olyan komplex eseményszerűségek esetében, amelyeknek bizonyos triád-komponensei nemcsak koncepuálisan vannak kitöltve, de nyelvileg is lexikalizálódtak, megjelenítsük az egyes eseménykomponensek összetartozására vonatkozó információt is. Ennek megfelelően annak analógiájára, ahogyan PWN mellékneveinek a főnévi hierarchiától független strukturálása vált elfogadottá, az igei wordnet készítésekor is példaként lehet venni a melléknévi synsetek jelentős részének szófajspecifikus (bokros / clusteres) elrendezését. Az imént felvázolt triád ugyanis leképezhető a wordnet rendszerére relációs formában. Ez annyit jelent, hogy a Moens és Steedman-féle eseménytriád által leírt metanyelvi szintnek, ahol tudjuk, megfeleltethetjük a nyelvi szinten lexikalizálódott elemeket – amelyeket a wordnetben synsetek jelölnek. A két szint összekapcsolását a 4. ábra mutatja.

[47] A wordnet synseteiben előforduló jelentések minimális propozíciókká alakíthatóságát az biztosítja, hogy egy-egy igéhez hozzárendeljük az adott jelentésben előforduló összes vonzatkeretét. Sokszor egy-egy vonzatkeret-bejegyzés több összevont vonzatkeretet takar, opcionális argumentumokkal. Ilyenkor az igéket a minimálisan kötelező argumentumszámukkal értelmezzük. Pl. az opcionális tárggyal felvett *eszik* vonzatkeret esetében a tárgy nélkül képzett minimális predikátumról mondhatjuk ki atelikus voltát.



4. Ábra Moens és Steedman eseménynukleuszának leképezése a wordnet synseteire

A HuWN-ben a synsetek strukturálhatóságának céljából bevezetett mesterséges csomópontok (ld. [5]) alkalmasak arra, hogy egy-egy, a metanyelvi szinten elhelyezett eseménynukleuszt megnevezzünk általuk – a fenti példában a nedves és száraz állapotváltozást denotáló komplex eseményszerűséget.[48] A wordnetek relációs felépítése lehetővé teszi, hogy annak megfelelően, hogy egy lexikalizálódott igei kifejezés a triád mely komponensét lexikalizálja, kapcsoljuk külön-külön névvel a triád egységét megjelenítő mesterséges csomóponthoz. A triád komponenseinek eredeti, angol nevét alapul véve a három felvehető reláció neve *nucleus_pp* (a *preparatory process* kifejezésből adódóan), *nucleus_t* (a *telosz* kifejezésből adódóan), és nucleus_cs (a *consequent state* kifejezésből adódóan). A három reláció közül a *nucleus_pp* értelemszerűen az előkészítő folyamatot lexikalizáló synset felé mutat, a *nucleus_t* a sikerpontot lexikalizáló synset felé, míg a *nucleus_cs* az utóállapotot lexikalizáló synset felé.

Ezáltal olyan, az angolban egyetlen igeként lexikalizálódott jelentések különíthetők el, amelyek a magyarban gyakran egy igekötős, aspektuális többletinformációt hordozó, és egy igekötő nélküli, aspektuálisan alulspecifikáltabb igeként jelennek meg. A fenti példában mind a {szárad}, mind a {megszárad} synset az angol {dry:2} synsetnek feleltethető meg. A nukleusz struktúra wordnetbe való leképezése nélkül a {megszárad} synset az eddigi, rendelkezésünkre álló relációk felhasználásával kizárólag a {szárad} synset hiponimájaként lenne felvehető. Ez a tárolási mód azonban összemosná az említett két jelentés között fennálló implikációs viszonyt az argumentumokra vonatkozó megszorítás különbségeire építő hiponima- hipernima viszonnyal (amely pl. a {hervad, fonnyad} és a {rohad} synset között áll fenn, mivel előbbinek csak növényi alanya lehet, utóbbinak pedig nemcsak az). A nukleusz struktúra wordnetbe való leképezésével a két igei synset között nem szükséges explicit módon egy külön relációt felvennünk: az őket összefogó mesterséges csomóponton keresztül, a *nucleus_pp* és a *nucleus_t* relációk mentén egyértelműen meghatározható, hogy a {szárad} synset a {megszárad} synset előkészítő folyamata, így per definitionem a

[48] A mesterséges csomópontokat a wordnetben a kapitálisokkal való írásmóddal különböztetjük meg a természetes nyelvi synsetektől.

{megszárad} implikálja a {szárad}-ot, de nem fordítva.[49] Az egy triádon belül lexikalizált igék esetében tehát az igekötős és igekötő nélküli alak elhelyezése a wordnetben sok esetben kézenfekvőbb, mint az eddig rendelkezésünkre álló relációk segítségével. Ezenkívül a fenti példában a triád harmadik komponensét lexikalizáló melléknévi synsethez ({száraz}) mutató *nucleus_cs* relációnak köszönhetően egy olyan, pszicholingvisztikailag releváns információ lesz leolvasható a {szárad}, {megszárad}, ill. a {száraz} synsetek között, mely az angol wordnetet alapul véve, a triád beépítése nélkül elveszne.

3.2.1 Jellegzetes triádok az igei HuWN-ben

Lévén, hogy a Moens és Steedman féle eseménynukleusz mindhárom komponensének kitöltött állapota egy *előkészítő folyamattal*, egy *sikerponttal* és egy *utóállapottal* rendelkező eseményszerűséget feltételez, abból a feltevésből indultunk ki, hogy a triád-struktúra adaptálása a wordnetbe a valamilyen változást (leginkább talán állapotváltozást) jelentő igék körében nagy százalékkal bizonyul majd hasznosnak. Ennek megfelelően a {változik:1} ill. {változtat:1} csomópontokat – melyek a PWN csomópontjainak önmagukban is egy negyedét kiteszik – választottuk ki arra a célra, hogy megvizsgáljuk a triád-struktúra gyakorlati beépítésének előnyeit a HuWN-be.

Az a sejtésünk, hogy a {változik:1} ill. {változtat:1} synsetek hiponimáiból álló részfákban számos olyan csomópontot találhatunk, amelyek alkalmasak a triádban való ábrázolásra, beigazolódott, amikor egy-egy esemény-nukleusz komponenseinek kitöltöttségét kódoltuk. Eredményeinket a 3. táblázatban bemutatott adatok illusztrálják.

3. Táblázat:
A felvett nukleuszok aránya a {változik:1} ill. {változtat:1} csomópontok alatt

nukleusz-típus	{változik}	{változtat}
<*, b, c>	24%	14%
<a, b, c>	30%	42%
Összesen	54%	56%

A {változik} ill. {változtat} synsetek közvetlen első ill. második szintű hiponimái közül összesen 150-et megvizsgálva tehát arra jutottunk, hogy triádok felvétele az esetek több mint felében tette könnyebbé a magyarban lexikalizálódott jelentések ábrázolását a wordnetben.

A {változik} és {változtat} csomópontok (ill. megfelelő hiponimái) közötti konceptuális hasonlóság miatt felmerült, hogy az ezekhez tartozó eseménynukleuszok párhuzamos módon épülnek fel. Közelebbről megvizsgálva azonban a triádok egyes pontjai közti viszonyok különböznek, amit a kauzatív-inchoatív alternációval magyarázhatunk. A különbség a következő: a <SZÁRAD, MEGSZÁRAD, SZÁRAZ> triád második és harmadik eleme között *következmény*-reláció áll fenn, amennyiben a *megszárad* igével kifejezett esemény következménye a *megszárad* alanyának száraz volta. A

[49] Ld. a 2.3 alatti leírást a Moens és Steedman által *contingency*-nek nevezett összetartozásról a triad elemei között.

látszólag párhuzamosan felépülő <SZÁRÍT, MEGSZÁRÍT, TÚLVAN A MEGSZÁRÍTÁS FOLYAMATÁN> triád második és harmadik komponense között más a viszony: az eseményszerűség utóállapota a *megszárít* ige *tárgyára* vonatkozik, azaz a {megszárít:1} synset *okozás* (*causes*) relációban áll a {száraz} melléknévi synsettel.

Bár az angol oldalon a {dry:1} (kauzatív) és {dry:2} (inchoatív) synsetek között fel van véve a *causes*-reláció[50], sem a {dry:1}, sem a {dry:2} synset felől nem mutat semmilyen reláció a melléknévi {dry:1} felé. Ha a magyar wordnetben a két triád viszonyának jelölése az angol mintára történne, mind az előkészítő folyamatot lexikalizáló kauzatív / inchoatív igepár ({szárít} - {szárad}), mind a sikerpontot lexikalizáló kauzatív/ inchoatív igepár ({megszárít} – {megszárad}) között fel kellene venni egy-egy *causes*-relációt, ami azonban szükségtelenül megduplázná a felveendő relációk számát, s a kauzatív igék és a melléknévi synsettel lexikalizált okozat között még mindig csak közvetetten lenne a kapcsolat leolvasható. Ezért az egyes triádokat jelölő mesterséges csomópontokat kötjük össze causes-relációval, azaz: MEGSZÁRÍT --causes--> MEGSZÁRAD, aminek köszönhetően éppúgy kiszámítható a triád élei mentén a kauzatív triád és az inchoatív triád igéi között fennálló okozás-reláció, mint a kauzatív triád igéi és az inchoatív triád utóállapotát lexikalizáló melléknévi synset között fennálló okozás-reláció.

4 Alkalmazási lehetőségek

Amellett, hogy a nyelvi elemek idioszinkratikus tulajdonságainak egységes ábrázolása jól beleillik a wordnetek fő feladatai közé, a HuWN ilyen típusú bővítése gyakorlati haszonnal is jár. Mint láttuk, az igék telikus-atelikus ill. befejezett-folyamatos volta könnyen meghatározható a triád komponenseinek segítségével. A HuWN ezen információinak felhasználása javíthat egy (magyar-angol) gépi fordító működésén, például az igeidők pontosabb megfeleltetése terén. Különösen hasznos lehet ez, ha a magyarban morfológiailag meglévő két igeidőt vetjük össze az angol igeidőkkel. A magyarra jellemző, hogy a morfológiailag jelen időben megjelenő telikus igealakok jövő idejű referenciával bírnak. A *Felhívom Pétert* mondatnak angolul nem *I call Peter*, hanem *I will call Peter* lenne a helyes fordítása. Hasonlóan, a folyamatos aspektussal bíró múlt idejű alakok nem egyszerű, hanem folyamatos múlttal fordítandók angolra: *Péter az udvaron játszott* megfelelője *Peter played in the court* helyett *Peter was playing in the court*. Az aspektuális információk mondatgenerálásnál is felhasználhatók – legyen szó akár angol-magyar fordításról, akár más, generálást igénylő feladatról.

Az igék ezen tulajdonságai a gépi mondatmegértésben is segíthetnek. Egy számítógép belső reprezentációjának fontos eleme az igék ezen idioszinkratikus tulajdonságainak ismerete. Enélkül, pusztán az esetleg a mondatban meglévő időhatározókat figyelembe véve nem lehet pontos egy történet időbeli struktúrájának ábrázolása.

[50] Sajnos a {change:2} és {change:1} csomópontok közötti ilyen kapcsolatmegadás nem konzekvens a PWN-ben.

Bibliográfia

1. Bach, E.: The Algebra of Events. Linguistics and Philosophy 9 (1986) 5-16
2. Dowty, D.: Word Meaning and Montague Grammar. D. Reidel, Dordrecht, W. Germany (1979)
3. Fellbaum, C.: WordNet An Electronic Lexical Database. MIT Press (1998)
4. Kiefer, F.: Jelentéselmélet. Corvina, Budapest (2000)
5. Kuti, J., Vajda, P., Varasdi K. Javaslat a magyar igei WordNet kialakítására. In III. Magyar Számítógépes Nyelvészeti Konferencia - MSZNY 2005, Szeged, 2005, 79-88.
6. Moens, M., Steedman, M.: Temporal ontology and temporal reference, Computational Linguistics Volume 14 Issue 2 (1988) 15-28
7. Tufis, D. et al. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In: Romanian Journal of Information Science & Technology, 7(1-2) (2004) 1-35.
8. Vendler, Z.: Verbs and Times. Philosophical Review 66 (1957) 143-160
9. Verkuyl, H. J.: On the compositional nature of the aspects: Foundations of Language. Supplementary Series 15. Reidel, Dordrecht (1972)
10. Vossen, P.: EuroWordNet General Document. Technical Report EuroWordNet (LE2-4003, LE4-8328) (2005)

Főnevek a Magyar WordNetben

Hatvani Csaba[1], Kocsor András[1], Miháltz Márton[2],
Szarvas György[1], Szécsi Katalin[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2.
{hacso, kocsor, szarvas}@inf.u-szeged.hu
[2] MorphoLogic Kft., 1126 Budapest, Orbánhegyi út 5.
{mihaltz, szecsi}@morphologic.hu

Kivonat: A Magyar WordNet, a többnyelvű BalkaNet/EuroWordNet rendszerekhez kapcsolódó magyar nyelvű wordnet-ontológia fejlesztése három intézmény részvételével 2005-ben indult egy hároméves pályázati projekt keretében (GVOP-AKF-2004-3.1.1). A tanulmány a Magyar WordNet ontológia teljes főnévi részének felépítését mutatja be. Részletesen leírjuk azokat a módszertani elveket és bővítési lépéseket, melyeknek segítségével kialakítottuk a jelenleg mintegy 20.000 főnévi csomópontból (synsetből) álló lexikális fogalmi hálózatot. Ismertetjük továbbá az általunk követett bővítési módszertan minőségi vizsgálatának módszereit és eredményeit is.

1. Bevezetés

A Magyar WordNet (HuWN) ontológia ([1], [4]) létrehozásával a magyar nyelv bekapcsolódott a Princeton WordNet (PWN) ([3]) architektúrájára épülő, EuroWordNet (EWN) ([6]) és a BalkaNet (BN) ([2], [5]) többnyelvű ontológiarendszerekbe. A napjainkra a legtöbb nagy európai nyelvet tömörítő kezdeményezéshez való csatlakozással számos nyelvtechnológiai probléma, mint például a gépi fordítás előtt is új távlatok nyílhatnak meg. Az alábbiakban bemutatjuk a Magyar WordNet főnévi állományának főbb jellemzőit, az ontológia kialakítása során alkalmazott módszertani elveket, megoldásokat, valamint egy vizsgálat eredményeit, mellyel bővítési módszertanunk minőségét vizsgáltuk.

2. Módszertani elvek

A legfőbb célunk a Magyar Wordnet ontológia kialakítása során olyan felső szintű, általános nyelvi tudást megjelenítő fogalmak felvétele volt, melyekhez a későbbiek során könnyen kapcsolhatók kisebb, domain-specifikus fogalomi hálók (mint pl. a későbbiekben létrehozandó gazdasági szókincset leíró ontológia).

A *fogalmi sűrűség elvének* nevezett, gyakorlati szempontból lényegesnek tartott elv alatt azt a törekvést értjük, hogy a létrehozott Magyar WordNet ontológiában az összes olyan fogalom szerepeljen, ami egy másik, az ontológiába felvett csomópont által szimbolizált jelentést magában foglal, annál általánosabb. A fogalmi sűrűség kritérium teljesíthető, ha minden bővítési szakasz után képezzük a főnévi hálózatnak az angol wordnet hipernima-relációi szerinti lezártját, és az esetlegesen hiányzó synsetekkel bővítjük azt.

3. A főnévi hálózat bővítése

A munka korábbi szakaszában elkészítettük a BalkaNet közös fogalmi készletének, a BalkaNet Concept Set (BCS) synsetjeinek magyar reprezentációját. A BCS 8516 synsetje (köztük 5896 főnévi) tartalmazza a EWN projekt 8 nyelvében, valamint a BN további 5 nyelvében legfontosabbnak tartott, az ontológiai hierarchia szempontjából alapvetőnek számító fogalmakat, melyeket minden nyelven implementáltak, biztosítva ezzel a nyelvek közötti minimális átjárhatóságot A munkáról bővebben lásd [4].

Ezt a főnévi magontológiát a bővítettük ki 19.500 tételesre, az alábbiakban részletesen ismertetjük ennek menetét.

3.2 Lokális alapfogalmak

A EWN és BN projektekben alkalmazott metodológiát követve először elkészítettük a lokális főnévi alapfogalmakat (Local Base Concepts, LBC), vagyis a magontológiába tartozó, de a közös halmazban (BCS) már nem szereplő fogalmakat reprezentáló synseteket. Ehhez korpuszstatisztikai módszereket alkalmaztunk: a Magyar Nemzeti Szövegtár főnévi gyakorisági listáját, illetve az Értelmező Kéziszótár (EKSZ) egy elektronikus változatában a főnévi definíciók szemantikai elemzéseit. A leggyakoribb MNSZ-ben szereplő, valamint az EKSZ definíciókban leggyakrabban genus proximum-ként szereplő főneveknek heurisztikusan megállapítottuk a leggyakoribb jelentéseit. A magyarra lefordított BCS-ben (BCSHu) felvett EKSZ azonosítók segítségével meghatároztuk ezek közül azokat a fogalmakat, amelyekhez még nem létezett synset a BCSHu-ban. Ezek alapján 250 szójelentéshez vettünk fel új synseteket, illetve EKSZ hivatkozásokat létező, megegyező jelentésű synsetekhez. A magyar főnévi mag-ontológia ezek után nagy valószínűséggel tartalmazza a BalkaNet/EuroWordnet alapfogalmain túl a magyar nyelvben legfontosabb kiinduló jelentéseket is.

3.2 Koncentrikus bővítés

A BCS és a magyar nyelvre fontosnak ítélt LBC-k elkészítése után a főnévi fogalmi háló bővítése során az angol nyelvre meglévő fogalmi hálót tekintettük kiindulási alapnak. Célszerű választás volt minden bővítési szakaszban a már elkészült magyar hálózat angol nyelvű képéből közvetlenül elérhető csomópontok közül válogatni. Így egyrészt az angol oldalról automatikusan teljesült a fogalmi sűrűség elve, másrészt –

lévén a magontológiából indultunk ki – többnyire általánosabb, a hipernima-hierarchiában magasabb szinten levő fogalmak kerültek a jelöltek közé.

Mivel a felsőbb szintű, absztraktabb fogalmaknak tipikusan egynél több hiponimájuk van, a fejlesztés során végig 30-40 ezer közvetlenül elérhető angol fogalom alkotta a jelöltek halmazát. Egy munkafázisban általában néhány ezer csomóponttal bővítettük az ontológiát, így a jelöltek közül szükségessé vált a céljainknak legmegfelelőbbek kiválasztása. A rangsoroláshoz négy, egymással nem feltétlenül összhangban lévő szempontot használtunk:

Fordíthatóság: A fogalomjelölt előkészíthető volt a korábban kidolgozott automatikus fordítási heurisztikákkal ([1], [4]). Ebben az esetben a synset létrehozása magyar nyelven egyszerűbben és gyorsabban elvégezhető volt, hiszen egy vagy több literál azonnal az annotátor rendelkezésére állt magyarul is.
Gyakoriság: A fogalomjelölt literáljai angol nyelvű korpuszokban (British National Corpus, American National Corpus First Release, SemCor) gyakran fordultak elő. Ez legtöbbször azt jelzi, hogy az adott szó a kommunikációban gyakran előkerülő fogalmat takar, azaz a felvétele a Magyar WordNetbe indokolt.
Nyelvek közötti átfedés: A megfelelő synset az angolon kívül sok más nyelvű wordnetben is előfordul. Ilyen synsetek felvételével egyrészt maximalizálhatjuk a Magyar Wordnet más nyelvekkel való átfedését, ami pl. fordítási feladatokhoz előnyös lehet, másrészt olyan fogalmakat veszünk fel, melyeket több más kutatócsoport is fontosnak ítélt, hiszen felvette az adott nyelv ontológiájába.
Relációk száma: A bővítés első szakaszában figyelembe vettük, hogy az adott synset hány új fogalmat tesz elérhetővé. A sok hiponimával rendelkező gyűjtőfogalmak felvétele célszerű volt, mert növelte a későbbi bővítési fázisokhoz a beválasztható synsetek számát.

Minden lépésben a gyakoriság és a nyelvek közötti átfedés (továbbá első körben a relációk száma) alapján rangsorolt fogalmakat választottunk ki a magyar ontológia bővítésére úgy, hogy az automatikus fordítással rendelkező synsetekből 3-4-szer többet vettünk fel, mint a fordítással nem rendelkező jelöltekből. Az úgynevezett koncentrikus bővítés során az első fázisban 2705, a második szakaszban 4385, majd végül további 800 fogalmat dolgoztunk ki.

3.3 Teljes részfák felvétele

A főnévi synset-állomány iteratív koncentrikusan kifelé terjeszkedő bővítése mellett kiválasztottunk néhány speciális területet, ahol minden PWN-ben ismert fogalmat lefordítottunk, vagyis az adott fogalmi körhöz tartozó teljes hipernima-részfákat átemeltünk. Ezzel az ontológia általános enciklopédikus tudását igyekeztünk az adott területeken teljessé tenni.

A következő fogalmi körökre alkalmaztuk ezt az eljárást:

- földrajzi nevek (országok, fővárosok, nagyvárosok, országon belüli (tag)államok (pl. USA államok), földrajzi területek (geopolitikai régiók), egyéb régiók, földrészek, "víznevek" (tavak, folyók, tengerek, öblök, óceánok, vízesések), hegycsúcsok, szigetek);
- emberi nyelvek (és nyelvcsaládok);

- embercsoportok (népek, ill. egy-egy régió lakosai);
- a világ országainak pénzegységei.

Összesen 3,200 synsetet vettünk fel ezen kritériumok alapján.

A gazdasági szakontológia számára ezzel a módszerrel felvettünk további 940 fogalmat a gazdaság, vállalkozás és a kereskedelem szakterületéről is.

3.4 Domain synsetek

A PWN 2.0-s verziójában bevezetett domain-relációk segítségével reprezentálni lehet a korábbi szemantikai relációkkal (főneveknél: hipernímia, holonímia, antonímia) ki nem fejezhető kapcsolatokat, illetve szerepük lefedi a hagyományos (értelmező) szótárak tárgyterületi, nyelvhasználati minősítő kódjainak funkcióját is. A reláció egy *domain synset* mint összefogó kategória és egy vagy több *domain term synset* mint elem között ábrázol tematikus/nyelvhasználati kapcsolatot. A domain relációnak három fajtája van: tartalmi/tematikus/szemantikai kapcsolatot kifejező (category), térbeli kapcsolatot kifejező (region), valamint nyelvhasználati kategóriát kifejező (usage).

Annak érdekében, hogy a HuWN PWN-re támaszkodó későbbi bővítése során akadálymentesen lehessen a domain-relációkat az angolból átvenni, felvettük az összes, PWN-ben szereplő category és region domain synsetet. A region domain fogalmak körét kiegészítettük speciálisan magyar régiók gyűjtésével is. A usage domain relációk használatát elvetettük, a PWN-ben tapasztalt inkonzisztenciák miatt: néhol egy synset-re vonatkozó usage minősítés nem minden szinonimára (literálra) érvényes, és ez a kódolás nem teszi egyértelművé, hogy melyek azok. (A PWN-ben ugyanakkor előfordul az is, hogy pont a usage minősítés alapján választanak szét két synsetet.) Az általunk bevezetett, literál-szintű nyelvhasználati kódolás, melynek segítségével a synset minden egyes elemére külön-külön megadhatunk minősítő kódokat, rugalmasabb megoldást biztosít.

3.5 Tulajdonnevek

A nemzeti wordnetek kisebb-nagyobb számban tartalmaznak named entity (NE, „névvel rendelkező entitás”, itt most tulajdonnév) jellegű synseteket is. Ezek között vannak „univerzálisak”, pl. a világ országai, fővárosai; világirodalmi jelentőségű alkotók, híres képzőművészek, tudósok, politikusok, és vannak az adott nemzethez, országhoz köthető, helyi fontosságú NE-k, pl. az adott ország megyéi, régiói, települései; a nemzeti irodalom, képzőművészet, tudomány, politika nagyságai. A Magyar WordNet ilyen irányú bővítéséhez tematikus NE-listákat gyűjtöttünk. Ezek az alábbi fő kategóriákba sorolhatók:

- földrajzi nevek (ország, megye, település, egyéb (hegy, víz stb.))
- intézmény jellegű nevek (cégek, nevezetességek, kórházak, színházak, mozik, légitársaságok stb.)
- személynevek (keresztnevek, családnevek, híres emberek nevei (művészek, történelmi alakok stb.))
- címek (újságok)

- márkanevek (termékek, árucikkek)

A listákat áttekintve az alábbi feladatokat határoztuk meg:

1. Egységesítés (formátum- és kódkonverziók)
2. Szelekció (mely kategóriákat, ill. az adott kategóriából mely NE-ket integráljuk)
3. Kollekció (a kiválasztott NE-knek milyen alakváltozatait, szinonimáit, körülírásait vegyük fel)

3.5.1 Synset-szintű beépítésre kijelölt NE-k

Bizonyos tematikus listák kiválasztott elemei közvetlenül bekerülnek a Magyar WordNetbe. A preferálandó kategóriák:

- földrajzi nevek:
  - országnevek
  - magyar megyék
  - magyar települések
  - világvárosok
- intézménynevek:
  - magyar nevezetességek
- személynevek:
  - magyar keresztnevek
  - híres emberek

Ha egy tulajdonnévnek van a magyarban meghonosodott írásmódja, literálváltozata, akkor alapelv, hogy minden esetben a magyar írásmódú alakokat tüntessük fel.

A synset-szintű beépítéssel kapcsolatos részfeladatok:

1. A szelektálandó állományok kézi leválogatása (a beépítendő NE-k megjelölése), a kiválasztott NE-k írásképének ellenőrzése, javítása.
2. Szelektálás közben az „ömlesztett” anyag finomítása, pl. a híres ember kategórián belül megadni az albesorolás(oka)t (festő, író, költő, hadvezér, politikus, fizikus stb.).
3. Annak ellenőrzése, hogy a kiválogatott literálok szerepelnek-e az angol nyelvű wordnetben (automatikus ellenőrzésnél probléma lehet, ha eltér a magyar és az angol írásmód, pl. Róma – Rome – Roma, Olaszország – Italy). Itt az is járható út, ha az angol wordnetes ellenőrzés előtt automatikusan generáljuk ezeket a synseteket, és a kézi ellenőrzéskor történik meg az angol wordnetbe való bekötés, ha ott már létezik ez a synset.
4. A bekötési synset (hipernima) meglétének ellenőrzése, szükség esetén felvétele és magyarítása.

3.5.2 Csak táblázatszintű, pointeres beépítésre javasolt NE-k

A szelektáláskor „kihulló" elemek is vannak annyira értékesek, hogy elérhetővé tegyük őket a wordnetből. Ezek nem önálló synsetként jelennek meg az ontológiában, hanem listaszerűen. Ilyen módon kerülnek be az alábbi témakörökből NE-k:

- intézménynevek:
  – az összes fennmaradó
- személynevek:
  – magyar családnevek
- címek
- márkanevek

Az ezzel kapcsolatos feladatok:

1. A formátum egységesítése (egységes ékezethasználat, kiegészítő információk törlése stb.)
2. A pointeres megoldás technikai részleteinek meghatározása (új reláció bevezetése: synsetekről tematikus listák felé)

4. A bővítési módszerek kiértékelése

4.1 Vizsgálati módszer

Egyrészt arra voltunk kíváncsiak, hogy a Magyar WordNet központi részét alkotó synsetek (BCSHu) mennyire relevánsak a magyar beszélők számára, azaz valóban a mag-ontológiában van-e a helyük. Másrészt az egyes bővítési módszerek hatékonyságát is számszerűsíteni szerettük volna. Ehhez az egyes bővítési eljárások során felvett synsetekből választott véletlen mintákat értékeltettünk két magyar beszélővel. A mérés a következőképpen folyt le:

a) A vizsgálandó synset-halmazokból 200-as véletlen mintákat vettünk.
b) A minták synsetjeit egyenként értékelte a két magyar anyanyelvű személy egymástól függetlenül. Minden synsetet 1-től 10-ig kellett pontozniuk. A nagyobb szám a fogalom nagyobb relevanciáját jelenti az értékelő személy számára. A két annotátor közötti egyetértés az összes értékelésre átlagolva 78,67%-os volt (egy adott synset értékelésében az egyetértést 100%-osnak vettük, ha mind a két annotátor ugyanazt a pontot adta, 0%-osnak, ha a különbség maximális (9 pontnyi) volt. Az értékeket átlagoltuk a synsetek és az összes minta felett.)
c) A két értékelő pontszámait synsetenként átlagoltuk, majd kiszámoltuk a 200 synset pontátlagainak átlagát és szórását.

4.2 Eredmények, értékelés

Az 1. és a 2. táblázat oszlopai jelentik azokat a főnévi wordnet-szegmenseket, melyek mindegyikéből 200 synsetes véletlen mintát generáltunk. Az egyes szegmensek:

NONBCS: az angol wordnet BCS-en kívüli synset-állománya
BCS1: a mag-ontológia 1-es szintje
BCS2: a mag-ontológia 2-es szintje
BCS3: a mag-ontológia 3-as szintje
CONC_1: az 1. koncentrikus bővítési körben felvett synsetek
TREE: a teljes részfák felvételekor bekerült synsetek
CONC_2_CAND: a 2. koncentrikus bővítési kör jelöltjei: a hipernima-reláció mentén elérhető synsetek
LIT_FREQ: korpuszokból készült szóalak-gyakorisági listák alapján felvett synsetek a 2. koncentrikus bővítési kör jelöltjei közül
ILI_OVL: a nyelvek közötti átfedések száma alapján felvett synsetek a 2. koncentrikus bővítési kör jelöltjei közül

	NONBCS	BCS1	BCS2	BCS3	CONC_1	TREE
átlag	4,51	6,56	6,21	5,03	5,71	4,21
szórás	2,48	2,78	2,20	2,45	1,71	2,61

1. táblázat

Az 1. táblázat NONBCS eredménye igazolja várakozásainkat. Megállapítható, hogy ha a bővítés során a bekerülő fogalmakat véletlenszerűen választottuk volna ki az angol wordnetből, akkor átlagosan kisebb relevanciájú fogalmak kerültek volna a Magyar WordNetbe. Ez tulajdonképpen a CONC_1 halmazának kontrollcsoportja.

A BCS1, BCS2 és BCS3 mérések kilógnak a sorból, abban az értelemben, hogy ezeket a synset-halmazokat nem mi válogattuk össze, hanem a EuroWordNet és a BalkaNet ontológiákból „örököltük”. Itt azt mérhettük meg, hogy az említett projektek által kulcsfontosságúnak tekintett fogalmak mennyire relevánsak a két magyar értékelő számára. A kapott eredmények átlagai és szórásai is elmaradtak várakozásainktól. Magasabb relevanciára és kisebb szórásértékekre számítottunk.

A CONC_1-ben, azaz az első koncentrikus bővítési körben olyan synseteket építettünk be a Magyar WordNetbe, melyek a mag-ontológiából egy lépésben elérhetők a hiponima-reláció mentén. A közepesnek mondható átlag a BCS viszonylag gyengébb eredményéből származtatható. Az alacsony szórásérték a synsetek értékességének homogenitását mutatja.

A TREE gyenge eredménye könnyen megmagyarázható. Egy teljes részfa beépítése azt jelenti, hogy felveszünk minden synsetet, amely egy bizonyos kijelölt csomópont alatt található. Pl. népek, nyelvek, pénznemek teljes részfái. Ezek között – a kiemelt jelentőségű fogalmak mellett – nagyon sok kevéssé ismert, illetve ritkán használt fogalom is szerepel. A TREE magas szórásértéke ezt a feltételezést alátámasztja.

	CONC_2_CAND	LIT_FREQ	ILI_OVL
átlag	4,25	5,26	8,32
szórás	2,27	1,74	1,25

2. táblázat

A CONC_2_CAND halmaz kb. 35 ezer főnévi synset-jelöltet tartalmazott a 2. koncentrikus bővítési körre. Az e halmazon elvégzett mérés azt mutatja, hogy ha véletlenszerűen választottuk volna ebből a 2. kör synsetjeit, akkor milyen értékes fogalmak kerültek volna be a Magyar WordNetbe. Ez tulajdonképpen a másik két oszlop halmazának kontrollcsoportja.

A LIT_FREQ eredményei a gyakorisági listák hasznosságát igazolják.

Az ILI_OVL-en mért átlag és szórás egyaránt kiemelkedő volt. Ilyen jó eredményeket nem vártunk ezen a szegmensen. A véletlen mintához viszonyított majdnem kétszeres átlag és alig több mint félszeres szórásérték e módszert mutatja messze a leghatékonyabbnak.

Összegzésként elmondható, hogy érdemes munkát fektetni a bővítés szisztematizálásába. A Magyar WordNet egyik értékét a synsetek magas relevanciája jelenti.

Bibliográfia

1. Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M.: Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. In: Proceedings of the Third International WordNet Conference, Seogwipo, Jeju Island, Korea (2006) 291–292.
2. Christodoulakis, D. N. (ed.) (2004): Design and Development of a Multilingual Balkan Wordnet. BalkaNet Final Report. http://www.ceid.upatras.gr/Balkanet/deliverables/finalreport_sub.pdf
3. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998)
4. Miháltz, M.: Magyar EuroWordNet projekt: bemutatás és helyzetjelentés. III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005) 68–78
5. Tufiş, D., Cristea, D., Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue, vol. 7, no. 1-2 (2004)
6. Vossen, P. (ed.): EuroWordNet General Document, Version 3. University of Amsterdam (1999)

A melléknevek beillesztése a Magyar WordNetbe

Gyarmati Ágnes[1], Almási Attila[2], Szauter Dóra[2]

[1] MTA Nyelvtudományi Intézet, Nyelvtechnológiai Osztály
1068 Budapest, Benczúr u. 33.
aagnes@nytud.hu
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Aradi Vértanúk tere 1.
{vizipal, szauter.dora}@freemail.hu

Kivonat: A wordnet lexikai adatbázisban a melléknevek rendezése merőben eltér a főnévi és igei struktúrától: a melléknevek közti legmeghatározóbb reláció nem a hipo-hipernímia viszony, hanem az antonímia, mely reláció mentén (kiegészítve a jelentéshasonlósági similar_to relációval) a deskriptív melléknevek clusteres struktúrába szervezhetőek. A Magyar WordNet melléknévi részének építésekor felmerültek nyelvspecifikus, valamint általános, a wordnet rendszerét érintő problémák. Nyelvspecifikus probléma az egyes szavak beilleszthetőségének kérdése, valamint a különböző „wordnetes" nyelvek közötti megfeleltetések megadása. Az antonímia reláció szigorúan két synset között van definiálva, mégis találhatók a wordnetben „antonimahármasok." Cikkünkben ezekre a kérdéseknek a rendezésére kínálunk konstruktív megoldást, mellyel a szemantikai viszonyok árnyaltabbá válnak a wordnetben.

1 Bevezetés

A *wordnet* lexikai adatbázisban az egyes szavak helyét szemantikai relációk határozzák meg. A melléknevek rendezése merőben eltér a főnévi és igei struktúrától. A melléknevek közti legmeghatározóbb, a melléknévi struktúra jellegét kialakító reláció nem a főneveknél és az igéknél megismert *hipo-hipernímia* viszony, hanem az *antonímia*. Ennek eredményeként a melléknevek nagy része ún. *clusteres* rendszerbe szerveződik.

A Magyar WordNet (a továbbiakban HuWN)[51] melléknévi részének készítésénél több különböző szempontot kell figyelembe venni. Jellegéből adódóan a Princeton WordNet (PWN) mintájára készül, és a magyar nyelv keretein belül igyekszik megtartani a PWN struktúráját, mind a literálok, mind a relációk felvételénél. A két nyelv között azonban komoly lexikai, valamint asszociációs különbségek vannak, ennek

[51] Az adatbázis a 2005 tavaszán indult Magyar Ontológia Építése projekt keretén belül készül, mely a Szegedi Tudományegyetem, a MorphoLogic Kft. és a Nyelvtudományi Intézet közös projektuma (GVOP – 2004 – 3.1.1.). A szerzők köszönetüket fejezik ki a projektben dolgozó kollégáiknak, elsősorban Kuti Juditnak és Varasdi Károlynak a HuWN melléknévi rendszerének kialakításában, ezáltal -- közvetve vagy közvetlenül -- a jelen tanulmány megírásában nyújtott segítségükért.

következtében a PWN melléknévi részének puszta lefordításával nem kapjuk meg a HuWN megfelelő részét.

Tanulmányunk a HuWN melléknévi részének építésekor felmerült általános és nyelvspecifikus problémákról ad számot, melyek nyomán szükségessé vált a szófaji kategorizáció vizsgálata, valamint néhány újabb reláció felvétele a wordnetbe.

2 A Princeton WordNet melléknévrendszere

A PWN melléknévi synsetjei többségükben mellékneveket tartalmaznak, de módosítóként szereplő főnevek, egyes igei származékok (folyamatos és befejezett melléknévi igenevek), valamint elöljárós szószerkezetek is kerültek közéjük. A PWN jelenleg 19500 melléknevet tartalmaz, melyeket 10000 szinonimacsoportba (a wordnet terminológiájával synsetbe) rendeztek. Ezek nagy részét deskriptív és relációs melléknevek alkotják, referenciamódosító melléknevek kisebb számban jelennek meg köztük [1].

Az alábbiakban röviden bemutatjuk a PWN melléknévi struktúrájának azt a részét, melynek ismerete a HuWN 3. fejezetbeli tárgyalásához elengedhetetlenül szükséges. Ennek értelmében a következőkben *deskriptív* melléknevekkel foglalkozunk, és csak megjegyezzük, hogy a r*eferenciamódosító* melléknevek egy része szintén beilleszthető a wordnetbe a *deskriptív* melléknevek közé, valamint hogy a *relációs* melléknevek kezelése merőben eltér a deskriptívekétől (ez utóbbi két melléknévosztályról, és a wordnetbe való integrálásukról ld. [1]).

2.1 Deskriptív melléknevek

A 'melléknév' szó hallatára rendszerint a deskriptív melléknevekre gondolunk. A deskriptív melléknév egy attribútum értékét adja a főnévnek. Pl. az *alacsony* és *magas* melléknevek a MAGASSÁG attribútumának (ellentétes pólusán levő) értékeit képviselik, és adják át azután a vonatkozó főnévnek, főneveknek. [1]

A deskriptív melléknevek szemantikai szerveződése teljességgel eltér a főnevek vagy az igék szerveződésétől, ugyanis melléknevek esetében nem beszélhetünk hiponímia relációról, ami hierarchiába rendezné őket. Deskriptív melléknevek közti alapvető szemantikai kapcsolat az *antonímia*, melynek jelentőségére először szóaszszociációs kísérletek során derült fény. Felnőtt beszélők esetén pl. a *jó* melléknévvel kapcsolatban elhangzott leggyakoribb asszociáció a *rossz* melléknév, a *rossz* esetén pedig a *jó* volt. Az asszociáció kölcsönössége a melléknevekkel kapcsolatos adatok egy szembetűnő tulajdonsága, érdemes ezt a lexikai információt a wordnetben is jelölni.

Antonim kapcsolat csak szavak, *literálok* között állhat fenn, erre utalnak a szóaszszociációs kísérletek eredményei is. Tegyük fel, hogy a főnevekhez és igékhez hasonlóan a (deskriptív) mellékneveket is szinonimacsoportokba, azaz synsetekbe osztjuk jelentésük alapján. Mint láttuk, ezek a synsetek nem szervezhetők hierarchiába *hipo-hipernímia* relációk mentén, erre sokkal alkalmasabb az *antonímia*.

Tekintsük a következő két lehetséges synsetet: {*heavy, weighty*} ('nehéz', 'súlyos') és {*light,* weightless} ('könnyű','súlytalan'), jelentésük alapján ez a két synset oppozí-

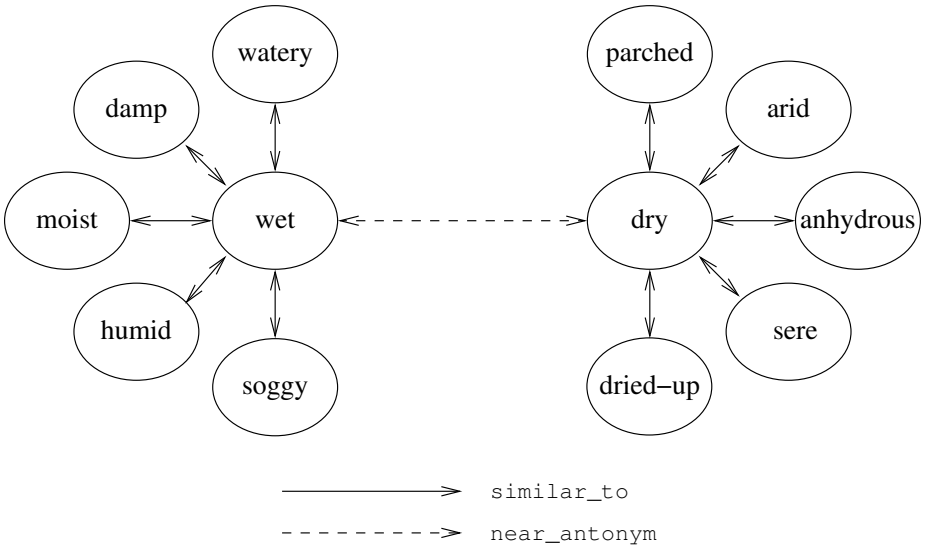
cióban áll egymással. Ennek ellenére nem mondhatjuk, hogy egy antonim párral állunk szemben, hiszen a synset elemei egyenjogú tagjai az őket tartalmazó synsetnek, így ha egy reláció a synset egyik literáljára fennáll, akkor definíció szerint minden további literál esetében fenn kell állnia a relációnak. Tehát ha a *heavy-light* antonim párt elfogadjuk, akkor létezőnek kell feltételeznünk a *heavy-weightless*, valamint a *weighty-light* párok antonimaságát is, márpedig ezek a viszonyok nem valósak.

Az antonímia relációnak a literálok közötti definiálásához érv továbbá az is, hogy az angolban a deskriptív melléknevek többségének ellentéte morfológiai szabály alkalmazásával jön létre. A jelentés polaritásának megváltozását egy negatív prefixumnak a szóhoz való hozzákapcsolása váltja ki (pl- *un-*, *in-* , és allomorfjai az *il-*, *im-*, *ir-*). A morfológiai szabályok szóalakokra vonatkoznak, nem pedig szójelentésekre. Ennek következményeként is el kell vetnünk, hogy az antonímia viszony a jelentések között húzódik.

Az antonímia a wordnet terminológiája szerint nem szemantikai, hanem lexikális reláció. A wordnet tehát nem konceptuális oppozícióként értelmezi az antonimaságot, hanem asszociatív relációként, amely (a prototipikalitás-elmélet szerint) központi helyet elfoglaló lexikalizálódott fogalmak között áll fenn. A konceptuális oppozíciót *indirekt antonímia* viszonyként jeleníti meg, az alábbiakban ismertetett módon.

A deskriptív melléknevek tipikusan kétpólusú attribútumok értékeit rendelik hozzá a főnévi fejekhez, ellentétes értelmű melléknevek egy adott attribútum ellentétes értékeit fejezik ki. Ezek az antonímia viszonnyal rendelkező melléknevek kitüntetett szereppel rendelkeznek az attribútumértékek halmazában, azaz abban a dimenzióban, melyre vonatkozóan értékeket jelölnek. Ezek a speciális helyzetű, a tartománynak mintegy a két pólusát lexikalizáló melléknevek lesznek az adott dimenzió *központi* vagy *fokális synsetei,*[52] melyek *near_antonym* relációval vannak összekötve. Köréjük csoportosulnak az adott dimenzió egyéb melléknevei, melyeknek nincsen antonim párjuk, azaz nincs *közvetlen antonimájuk*. Ezeket az ún. *szatelit synseteket* jelentéshasonlóság alapján, a hozzájuk közelebb álló fokális synsethez kapcsoljuk *similar_to* relációval, így a szatelit synseteknek is lesz – a hozzájuk kapcsolódó fokális synseten keresztül – közvetett antonimájuk. Ezt a clusteres szerkezetet szemlélteti a következő ábra (1).

[52] A wordnet terminológiájában a synset szinonimacsoportot jelent, itt csak az egyszerűség és a hagyomány kedvéért tartjuk meg az elnevezést, bár a szinonimák, mint a következőkben látni fogjuk, nem a synstben, hanem körülötte helyezkednek el.



1. ábra: A wordnet kétpólusú melléknévi szerkezete

Vannak olyan clusterek, melyeket mintegy mesterségesen egészítettek ki egy fokális synsettel, miután a két pólus közül az egyikhez nem találtak sem közvetlen, sem közvetett antonimát, de magának a dimenziónak a létjogosultsága nem volt megkérdőjelezhető. Ilyen például az *angry* melléknév esete, mely ezáltal az *angry-unangry* dimenzióban az egyetlen valóban lexikalizált fokális synset.

Miután bemutattuk a wordnet melléknévi struktúrájának alapjait, rátérhetünk a HuWN melléknévi rendszerének tárgyalására.

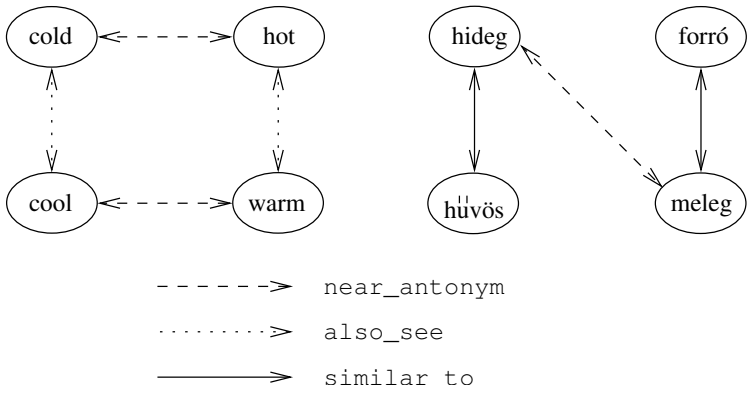
3 Melléknevek a Magyar WordNetben

A Magyar WordNet melléknévi része nem a PWN-nek egyszerűen magyarra fordított változata, elkészítése körültekintő munkát igényel. Erre elsősorban a két nyelv közötti lexikai, ill. asszociációs különbségek miatt van szükség, emellett igyekszünk a PWN-ben előforduló következetlenségeket kiküszöbölni. Ebben a fejezetben csak deskriptív melléknevekkel foglalkozunk.

3.1 Nyelvi sajátosságok

Mint azt a 2. fejezetben, a PWN melléknévi rendszerének ismertetésekor láttuk, az antonímia reláció konkrét szavak, nem pedig az általuk jelölt konceptuális tartalmak között állhat fenn. Ezt szem előtt tartva nem meglepő, hogy a HuWN struktúrája néhány kisebb eltérést mutat a PWN struktúrájához képest. Az alábbi (2) ábra egy ilyen esetet szemléltet.[53]

[53] Az ún. also_see reláció kapcsolja össze azokat az egymáshoz nagyon hasonló dimenziókat meghatározó fokális synseteket, melyek a hagyományos wordneti értelemben egyetlen közös



2. ábra: A hideg-meleg tartomány nyelvi különbözősége

Konceptuálisan csak egy ellentétpár létezik ebben a dimenzióban: a *hideg* és a *meleg* tartomány. Lexikális szinten azonban az angolban két oppozíció is jelen van: *cold-hot*, ill. *cool-warm*. Mind a négy melléknévnek találunk lexikalizálódott magyar megfelelőt (*hideg, forró, hűvös,* ill. *meleg*), de a közöttük lévő relációkat nem vehetjük át, ugyanis a magyarban csak egyetlen ellentét realizálódott. Az angol példa alapján feltételezhető *hideg-forró* pár csak a *forró* szó irányából létezhetne, hiszen a *forró* szóról egy magyar nyelvi beszélő a *hideg* (esetleg a *jéghideg*) szóra asszociál, míg a *hideg* esetében a *meleg*re.

Magában a szókészletben is mutatkoznak eltérések. Vannak az angolban (így a PWN-ben is) olyan melléknevek, melyekhez nincs megfelelő magyar melléknév. Ez az eset többféleképpen állhat elő. Egyfelől eredhet ez a lexikalizálódott magyar megfelelő teljes hiányából (pl. *unattractive* – 'nem vonzó'). Másfelől egy, az angolban melléknévvel kifejezett tartalom a magyarban akár más szófajú szóként is lexikalizálódhat, pl. *afraid* (mn) – *fél* (ige). Az ilyen típusú megfelelések jelölésére vezettük be újonnan az *eq_xpos_synonym* relációt, mely a szófaji határokon átívelő szinonimitást hivatott jelezni.

Egy-egy új szónak a lexikai adatbázisba való felvételekor több szempontot is meg kell vizsgálni. Egyfelől meg kell győződni arról, hogy valóban létezik ilyen szó, és ezáltal a szótárba való beillesztése indokolt. Másfelől meg kell határozni az adott szó szófaját. Ezekben a vizsgálatokban támpontot nyújthatnak a hagyományos szótárak (pl. a Magyar Értelmező Kéziszótár (ÉKSZ)). Azonban egy itt elért negatív eredmény még nem jelenti automatikusan egy szó felvételének kizárását, ebben az esetben segítségünkre lehetnek korpuszok is (pl. a Magyar Nemzeti Szövegtár [4]).

A szófaj meghatározása még akkor is lehet problémás, ha az adott szó szerepel egy hagyományos szótárban. Az *alvó* szó az ÉKSZ szerint főnév vagy melléknévi igenév lehet, 'aki éppen (javában) alszik', valamint 'hálószoba' jelentéssel. Érvelésünk szerint vulkánokra vonatkoztatva ez egy lexikalizált melléknév. A mellékneveknek más szófajoktól való elkülönítésére tesztek adhatók, melyek a szótárszerkesztők segítségére lehetnek [2], [3].

---

synsetbe kerülnének, de egy antonímia reláció révén saját „fokális synset”-ségük kétjogosultságot nyert.

A teljesség igénye nélkül bemutatunk néhány tesztet, melyek a melléknevek és a melléknévi igenevek megkülönböztetésére szolgálnak:

-- Állítmányi szerepben csak melléknevek állhatnak, pl. *ez a hír megdöbbentő* szemben azzal, hogy **ez a hír Pétert megdöbbentő*.[54]

-- Az alapige vonzatait a melléknevek nem tarthatják meg (a szabad határozó ez alól kivétel), pl. a *Pétert megdöbbentő hír*, illetve **ez a hír Pétert megdöbbentő.*

-- Csak melléknevek fokozhatóak, melléknévi igenevek nem, nem mondhatjuk, hogy *Pétert megdöbbentőbb hír.* (További tesztekhez ld. [3])

Látható, hogy a tesztek nem minden esetben megbízhatóak, csupán támpontot adnak. Ezek alapján az *alvó* szót még nem tekinthetnénk melléknévnek. A következő teszt indokolja döntésünket, melyek a lexikalizált jelzős szerkezetek (pl. *bontott tégla, vágott virág*) felismerését segítik.

-- A szerkezet jelentése sajátos (a *bontott tégla* kifejezés nem egyenértékű azzal, hogy 'tégla, melyet (le/szét-stb) bontanak/bontottak').

-- A jelzett szó elveszti élhangsúlyát (*'bontott tégla*, nem pedig *'bontott 'tégla*).

-- További jelzők csak a szerkezet egészére vonatkozhatnak, ezért egyrészt nem szakíthatják meg a szerkezetet, másrészt nem értelmezhetők kizárólag a jelzett szóra vonatkoztatva (**bontott piszkos tégla*, **piszkos, bontott tégla*).

-- A jelzős szerkezetnek megfelelő névszói állítmányú mondat agrammatikus (**Az udvar közepén levő tégla bontott*).

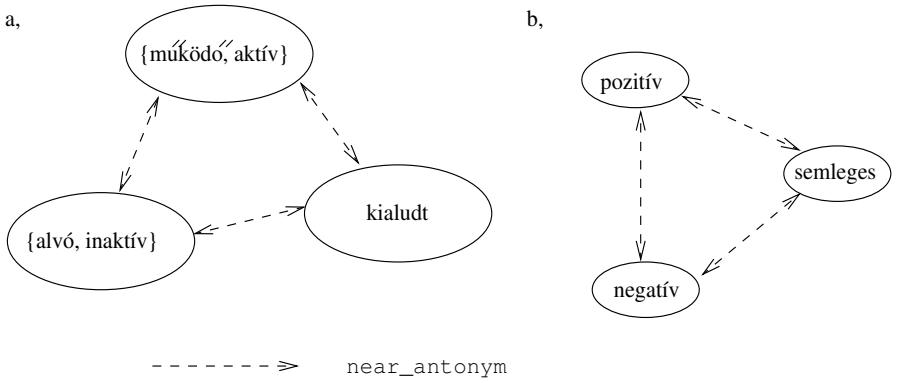
Ezek alapján az *alvó* szó melléknévként való felvétele megalapozott, hiszen érvelésünk szerint az *alvó vulkán* lexikalizálódott jelzős szerkezet. Nem mondhatjuk, hogy **a szigeten álló vulkán alvó*, sem azt, hogy **alvó nagy vulkán*, és a szerkezet egyetlen egységet alkot a hangsúly szempontjából.

Hasonló gondolatmenet alapján beillesztésre kerültek a wordnetbe pl. az agyhalott, ázott melléknevek is, melyek egyébként sem az ÉKSZ-ben, sem A magyar nyelv nagyszótárában nem szerepelnek.

3.2 A tipikustól eltérő dimenziók

A Princeton WordNetben vannak olyan *near_antonym* relációval rendelkező melléknevek is, melyek nem egy bizonyos kétpólusú dimenzió valamely végpontját jelölik. Az alábbi ábra (3) két ilyen esetet szemléltet.

[54] A * jelentése itt: rosszulformált.



3. ábra: A tipikustól eltérő melléknévi dimenziók

Ezekben a dimenziókban a központi synsetek mintegy háromszöget alkotva állnak egymással ellentétben. A PWN alapján azt feltételezhetnénk, hogy itt többdimenziós tartományokról van szó, melyeknek egyaránt három kitüntetett „végpontjuk" van. Amellett érvelünk, hogy a wordnetben nem helyes ez az ábrázolási mód.

A wordnet melléknévrendszere nagyrészt a lexikalizált oppozíciós párokra épül, ezek *kétpólusú* struktúrát határoznak meg. Felmerülhet az az igény, hogy ez a kétpólusú rendszert kiterjeszthető legyen úgy, hogy a fenti példák beilleszthetőek legyenek a wordnetbe anélkül, hogy a „kivételes eset" bélyeget kellene magukra venniük.

Mint ismeretes, két adott szó esetén az antonímia reláció felvételének szükséges feltétele a kettejük közti asszociáció megléte. Vizsgáljuk meg ebből a szempontból a (3.b) ábrán látható példát.

A *pozitív-negatív* pár egyértelmű ellentétpár, mely a szóban forgó dimenzió[55] két végpontját realizálja. De mi a helyzet a *semleges* melléknévvel? A *pozitív-semleges* pár oppozíciós viszonya hasonló módon zárható ki, mint az előző pontban a *hideg-forró* esetében: habár a *semleges* szó asszociálhatja a *pozitív* szót, fordítva ez nem áll fenn. A *negatív-semleges* pár helyzete még bizonytalanabb. Hiába érezhetünk egyfajta ellentétes viszonyt bármely kettő között, a lehetséges három antonímia reláció közül csak egyetlenegy vehető fel a wordnetben.

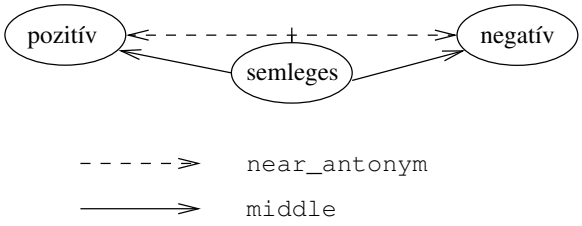
Ebből a konkrét példából is jól látható, de elvi megfontolások alapján is elvethető, hogy a wordnet filozófiájához hűen ki lehessen terjeszteni a *near_antonym* reláció használhatósági körét. Egy adott szó egy adott jelentésben csak egyetlen ellentétre vonatkozó asszociációt idézhet fel, ellenkező esetben ezt az asszociációs módszert eleve nem lehetett volna alkalmazni a wordnet melléknévi struktúrájának kialakítására.[56] Ekkor ugyanis nem lehetett volna a dimenziók végpontjait egyértelműen és megbízhatóan meghatározni. Nézetünk szerint nem a wordnet alapjainak megváltoztatásában kell keresni a kiutat, hanem egyfajta továbbfejlesztésében. A következőkben erre kínálunk megoldást.

A *pozitív* és *negatív* melléknevek közötti antonímia reláció létjogosultsága vitathatatlan, ez a két melléknév alkotja egy dimenzió két végpontját. Ez a dimenzió azon-

[55] Valójában jelentéstől függően több dimenzió is szóba jöhet, de ez a gondolatmenetünk szempontjából nem jelent problémát.

[56] Fontos kiemelni, hogy egy adott jelentésről van szó. A többszörös asszociációnak éppen a több jelentés megkülönböztetésében lehet szerepe (ld. a 2. fejezetben).

ban eltér a (metaszinten) prototipikusnak mondható tartományoktól, hiszen nem jelennek meg benne fokozatok, sem a két végpont között, sem körülöttük nem helyezkednek el jelentésükben csupán graduális különbségeket mutató szavak. Ezzel szemben a van lexikalizált kifejezés, mely ennek a durva fokozatú skálának pontosan a közepét nevezi meg, ez a *semleges* melléknév (nem kapcsolható *similar_to* relációval egyik végponthoz sem). A „szabályos" dimenziókban is léteznek olyan szavak, melyek a tartományuk közepére utalnak (a *hideg-meleg* dimenzióban pl. a *langyos*), de ezek vonatkozása bizonytalan, míg a *semlegesé* határozott. Ezt a kitüntetett szerepet javaslatunk szerint nem az elméleti síkon is megtámadható *near_antonym* reláció felvételével kell a tartomány két központi végpontjához kapcsolni, hanem egy új, a dimenzió határozott középpontját jelző *middle* relációval. Ezáltal az ellentétviszonyok nem alkotnának egy háromszöghöz hasonló szerkezetet, mint ahogy az a PWN rendszerében leolvasható, hanem a többi tartományhoz hasonló szerkezetet hoznának létre. A különbség csak annyi, hogy ebben a dimenzióban három kitüntetett szerepű synset is van, egyfajta elfajuló háromszöget alkotva (4. ábra).



4. ábra: A middle reláció

A *middle* reláció definiálásával nemcsak a „mindenki mindenkinek ellentéte" anomáliás helyzetet lehet kiküszöbölni, hanem a szemantikai viszonyokat is jobban tükrözi a rendszer. Szükséges még megjegyezni, hogy ezzel nem csak egy egyedi, elszigetelt esetet oldottunk meg. A *middle* reláció minden olyan dimenzióban felvehető (és felveendő), melyek középső pontjára (esetleg pontszerűnek tekintett belső intervallumára) lexikalizálódott megnevezés létezik az adott nyelvben, mint pl. *alsó-felső-középső*. (Azonban ha egy melléknév nem pontosan a középpontot fejezi ki, akkor az eddigiekben megszokott *similar_to* relációval kapcsolódik az egyik végponthoz.)

Első közelítésben a (3.a)-ban bemutatott példa esetében is alkalmazható lenne a *middle* reláció. A dimenzió két ellentétes végpontja a {*működő, aktív*}, ill. a *kialudt*, az {*alvó, inaktív*} pedig a középső tartományt jelöli. Itt azonban a középpont nem pontszerű, és nem is tekinthető annak. Sőt, ami ennél meglepőbb, az {*alvó, inaktív*} synset akár *similar_to* viszonyban lehetne, hiszen az *alvó* jelző 'most éppen nem működő' vulkánra utal, tehát jelentése közelebb áll a .{*működő, aktív*} synsethez. Azonban nem hagyhatjuk figyelmen kívül az *aktív* és *inaktív* szavak között feszülő ellentétet.

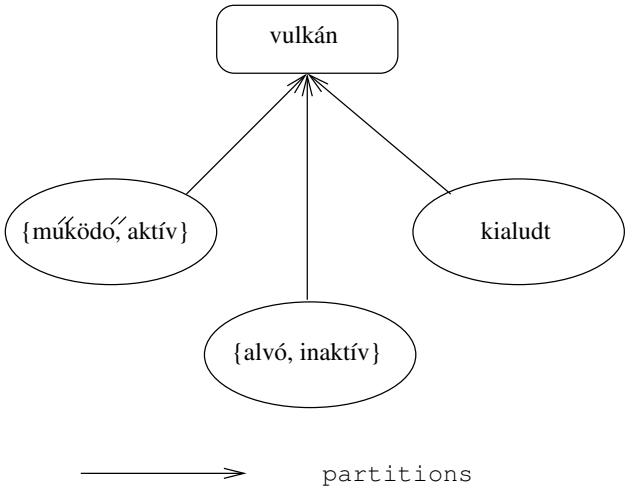
A {*működő, aktív*}, *kialudt* és {*alvó, inaktív*} synsetek által meghatározott tartomány nemcsak az egyes elemek között húzódó hasonlóság és ellentét kettőssége miatt tér el az eddig tárgyalt dimenzióktól. Ezek a melléknevek szemantikailag megszorítják, mire vonatkozhatnak: konkrét esetünkben *vulkánokra*. A wordnetnek számot kell tudnia adni erről a szemantikai kapcsolatról is. A PWN és a BalkaNet ezeket a mel-

lékneveket az antonímia relációval kapcsolja össze, és sokszor az általuk kizárólagosan módosított főnévvel nem is jelez kapcsolatot.

A megoldás ismertetése előtt megjegyezzük, hogy most sem csupán egyetlen tartományben lép fel ez a szituáció. Gondoljunk például a kizárólag *növényekre* jellemző *egynyári-kétnyári-évelő* tulajdonságokra, melyek a PWN-ben szintén háromszöget alkotó *near_antonym* relációval szerepelnek.

A szóban forgó melléknevek *particionálják* a főnév terjedelmét, azaz a főnév alá eső tárgyak halmazát az általuk jelölt tulajdonságok mentén diszjunkt részhalmazokra osztják. Az általunk felvett új reláció neve éppen ezért *partitions* (ld. 5. ábra).

A *partitions* hasonlít a wordnet *category_domain* relációjára, de nem azonos vele. A *category_domain* reláció azt a célt szolgálja, hogy megadja, egy bizonyos jelentés milyen témakörhöz kapcsolódik (pl. {*monovalent*:2 'egyvegyértékű, monovalens'} – {*chemistry*:1 'kémia, vegyészet'}), de nem mond semmit arról, hogy az adott melléknév mely főnevet módosíthatja, akkor sem, ha az kizárólagos. A *partitions* reláció definiálásával tehát teljesebbé és pontosabbá tehetjük a szemantikai kapcsolatok ábrázolását a wordnetben.



5. ábra: A partitions reláció

4 Értékelés

A HuWN melléknévi részének építése közben elért eredményeink elsősorban nem mennyiségi, hanem minőségi jellegűek. Habár a 3.1 fejezetben említettük, hogy a wordnet szókészlete bővebb (vagy bővebbé tehető), mint a hagyományos szótáraké, fontosabbnak tartjuk a minőség javítását. A 3.2 fejezetben megmutattuk, hogy hogyan lehet a wordnetben meglevő, a wordnetben elvileg tarthatatlan hármas melléknévi ellentéteket kiküszöbölni. A rendszerben lévő következetlenség felszámolásán túl a látszólag azonos, de két különböző probléma megoldásával, azaz két új reláció

definiálásával egy eszközt adtunk, mellyel a wordnetben a szemantikai viszonyok árnyaltabban ábrázolhatók.

Bibliográfia

1. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
2. Kiefer, F.: Jelentéselmélet. Corvina, Budapest (2000)
3. Komlósy, A.: Régensek és vonzatok. In: Kiefer, F. (ed.): Strukturális magyar nyelvtan I. Mondattan. Akadémiai Kiadó, Budapest (1992) 299—527
4. Magyar Nemzeti Szövegtár. http://corpus.nytud.hu/mnsz

IV. Szemantika

Hol fáj? – A jelentésreprezentáció nehézségei egy kórlapkitöltő rendszerben

Grőbler Tamás és Szőts Miklós

Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy J. u. 7.
{grobler,szots}@all.hu

Kivonat: Szövegek jelentésének formális reprezentálására szolgáló módszeren dolgozunk. A szövegek morfológiai és szintaktikai elemzése után a jelentésreprezentációt ontológiában állítjuk elő. A szemantikai elemzés feladata a világmodellt tartalmazó ontológia példányosítása a szöveg alapján, vagyis a szöveg jelentésének megfelelő individuumok és a közöttük fennálló relációk létrehozása. A szemantikai elemző tudásbázisát a kutatás keretében épülő ontológia adja, amelyet az OWL ontológia-leíró nyelven fogalmazunk meg. A cikkben áttekintjük a rendszer felépítését és működését, és eközben körbejárjuk azokat a pontokat, amelyek a jelentésreprezentáció szempontjából különleges bánásmódot igényelnek. Az egyes problémák kezelésére javaslatot teszünk. A konkrét feladat, amely keretében a kísérletet végezzük a kórházi panaszfelvételek során lejegyzett magyar nyelvű szövegek egységes kórlap-reprezentációvá alakítása.

1 Bevezetés

Az a kutatás, amelyről beszámolunk[57], nem nyelvészeti ihletből született, hanem az ontológiák tanulmányozásából. A kutatás célja, hogy egy szöveg jelentését formálisan reprezentáljuk. Ennek több praktikus felhasználási területe lehet, a jelenlegi projekt orvosi szabad szövegeket formális kórlapstruktúrába interpretál, de az igazi célkitűzés szemantikus kereső fejlesztése.

Noha, a kutatás nem nyelvészeti ihletből született, a kísérlet során egyre szorosabb nyelvészeti kapcsolatokra döbbentünk rá. A legfontosabb: egy diskurzusreprezentációt építünk, az egyetlen újdonság az, hogy nem csak a szövegből nyelvészeti módszerekkel kinyerhető információra alapozunk, hanem az ontológiában strukturálisan tárolt háttértudásra is támaszkodunk.

Az ontológiát magát sokan mint nyelvi elemek (terminusok) rendszerét értik. Azonban valójában az ontológia fogalmak rendszere, amely fogalmakat természetesen nyelvi kifejezésekkel nevezünk meg. A nyelvi elemek és az ontológia fogalmait már sokan megkülönböztetik, pl. a DOLCE csúcsontológiájához már kidolgozták a

[57] A kutatást a GVOP-3.1.1-2004-05-0363/3.0 sz. és az NKFP-2/042/04 (MEO) pályázat támogatja.

WordNet lexikonnal való kapcsolatot [3]. Többen használják a szóelőforduláson alapuló visszakeresés javítására ([2], [5], [8]) vagy szakmai szövegek generálására.

Mi az ontológiát **világmodellnek** fogjuk fel, és a szöveg jelentését ebben reprezentáljuk. Ehhez természetesen szükséges a nyelvi tudás tárháza (a továbbiakban lexikonnak nevezzük), valamint ennek elemeinek az ontológiába való leképzése.

Pragmatikus okok miatt feltételezzük, hogy a szöveg egy jól meghatározott témakörre korlátozódik („universe of discourse"), és csak leíró jellegű szövegeket kezelünk. Elvileg ezek a korlátok átléphetőek, de a megfelelő ontológia elkészítése rendkívüli erőfeszítést igényelne.

Az ontológia mint „világmodell" tartalmazza azon fogalmakat[58], amelyek előfordulásaira a reprezentálandó szövegek referálnak, legalábbis azokat, amelyek a szöveg megértése szempontjából fontosak – ezt tárgyaljuk a 2.fejezetben.

A nyelvi tudástár elemei mindazon elemi nyelvi objektumok (szavak, ragok, idiómák stb.), amelyek jelentése az ontológiában valamilyen módon reprezentálva van – erről a 3. fejezetben szólunk.

A továbbiakban ismertetjük azt a rendszert, amelyen dolgozunk, illetve bemutatjuk, hogyan működik a már megvalósított rendszer orvosi szövegeken.

Az utolsó fejezetben röviden összefoglaljuk a továbbfejlesztés terveit.

2 Ontológia – a világról szóló tudás

Ahhoz, hogy az ontológiában egy tudásterület szövegeinek jelentését reprezentálhassuk, az ontológiának képesnek kell lennie leírni azokat a szituációkat, amelyek a várt szövegekben tipikusan előfordulnak. Ehhez elsősorban egy megfelelően kidolgozott **csúcsontológiára** van szükség, amely megszabja mi az, ami kifejezhető, és mi az, ami nem. A következőkben néhány fontos kérdést veszünk sorra. Ezeket a kérdéseket általános szinten vetjük fel, de felhívjuk az olvasó figyelmét arra, hogy a szakontológiák kategóriái radikálisan eltérhetnek az általános célú ontológiákéitól. Példa: egy orvosi ontológiában a PÁCIENS és az ELLÁTÓ SZEMÉLYZET TAGJA fogalmak közös nemét, az EMBER fogalmat teljesen felesleges felvenni.

2.1 Eseményszerűségek reprezentálása

Az eseményszerűség (occurrence, perdurant: az események, folyamatok stb. gyűjtőfogalma) a természetes nyelvfeldolgozás szempontjából kulcsfontosságú: ezek azok a fogalmak, amelyekre (általában) igével referálunk. A csúcsontológiában az ESEMÉNYSZERŰSÉG és az ENDURANT (térben anyagi és absztrakt létező) kategóriák közt értelmezzük a *résztvevője* relációt, amely fajtái lesznek a szereprelációk, és kijelölik milyen fogalom előfordulása lehet pl. egy esemény *aktora*, *tárgya* stb. – lásd [7].

[58] A MEO projektben kidolgozott ontológiamodell szellemében mind az osztályokat, mind a relációkat fogalmaknak értjük.

2.2 Tulajdonságok

Hogyan reprezentáljuk azt, hogy az „ég kék”, vagyis hogy „az ég színe kék”? Vagy azt, hogy „a páciens vérnyomása 2006. nov. 8-án ülő helyzetben, bal karon mérve 220/178 Hgmm”? Míg az első példa egyszerűnek látszik – bár ott is van egy rejtett időfüggés –, a másodikban már nyilvánvaló a probléma: egy fogalomhoz (PÁCIENS) egy tulajdonság (VÉRNYOMÁS) egy értékét (*vérnyomás*a) kell rendelni, amely különböző paraméterektől (TESTHELYZET, MÉRÉSI HELY, IDŐ) függ**.** Nyilvánvaló, hogy különböző tulajdonságoknál különböző paraméterekkel kell számolnunk. Azt a megoldást választjuk, hogy a „tulajdonsága” relációt reifikáljuk, és azt a *hordozója*, *értéke*, valamint a paramétereknek megfelelő relációk (példánkban: *ideje*, *mérési helye*, *testhelyzetben*) kapcsolják a megfelelő objektumok, értékek osztályaihoz.

2.3 Idő

Az időkezelés minden tárgykör jellemzője, viszonylag egyszerű: megkülönböztetünk időintervallumot és időpontot, valamint egy lineáris skálát rendelünk az időpontokhoz. Az IDŐINTERVALLUM és IDŐPONT fogalmak közt értelmezve van a *kezdőpontja* és a *végpontja* relációk. Azonban az egyszerűséget elrontja a granularitási probléma: az egyes időfogalmak mint NAP, ÓRA stb. hol időintervallumot, hol időpontot jelentenek[59]. Ennek feloldására több út van, mi azt választottuk, hogy minden ilyen időfogalom az IDŐPONT fajtái, amelyek előfordulásai egy eseményszerűséghez rendelt IDŐINTERVALLUM előfordulás kezdőpontja ill. végpontja lehetnek. A szövegből kiindulva a szemantikus elemző határozza meg, hogy milyen fogalom előfordulásai szerepelnek mint időintervallum-határok.

2.4 Hely

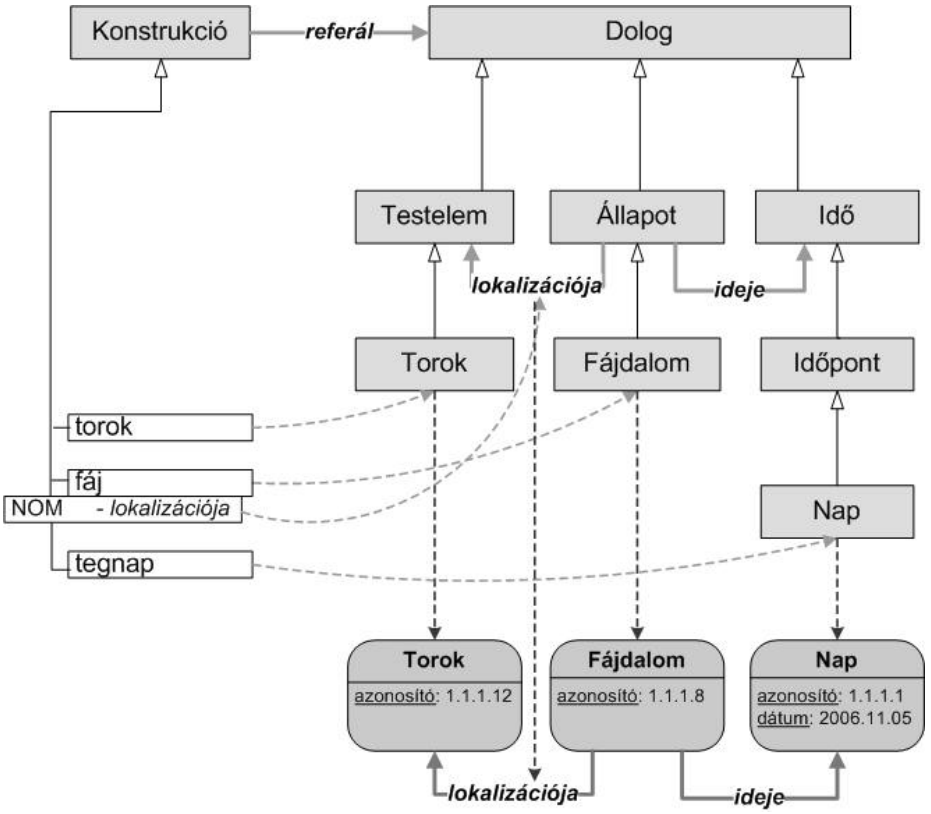
A hely kezelése lényegében különbözik az időtől: nincs, illetve általában nem használatos egy jól meghatározott koordinátarendszere. Egyéb fogalmak szolgálnak helymeghatározásra. Az orvosi témakörben két, egymástól teljesen különálló helymeghatározással találkozunk: egyrészt az emberi szervezet részeihez kapcsolódnak orvosi fogalmak (pl. „daganat a májon”), másrészt az egyes ellátó egységek mint az ellátás helyei. Két különböző relációval modellezzük ezt a két összefüggést.

3 A nyelvi tudás és az ontológia kapcsolata

A MEO modellnek [6] megfelelően az ontológiának két rétege van: a fogalmi réteg (a tulajdonképpeni ontológia), és a nyelvi réteg, amelyben a fogalmakat megnevező nyelvi elemek vannak. A nyelvi réteg egyelőre nem több, mint egy lexikon, amelyben az egyes lexémák (idiómák, ragok stb.) alatt ezek egyértelműsített változatai szerepelnek. Ezeknek a *referál* függvény adja meg jelentésüket (1. ábra). A lexémák egy csoportja és az igei vonzatkeretekben szereplő morfológiai jegyek közvetlenül is

[59] Az egy külön probléma, hogy időtartam-egységet is jelentenek.

utalnak egy-egy fogalomra ill. relációra. Elemzés közben a lexémák által referált fogalmak egy-egy előfordulását (példányát) hozzuk létre, amelyeket a fogalmak közötti megfelelő relációval kötünk össze. Eközben a példányok adat típusú tulajdonságokat is felvehetnek.



1. ábra A „Tegnap fájt a torka.” mondatnak megfelelő ontológiarészlet és jelentésreprezentáció. Az üres nyilak a generikus relációnak felelnek meg. A lexémáknak megfelelő fogalmakat a *referál* relációra kimondott megszorítások kötik a világmodell megfelelő fogalmaihoz. Az alsó sorban a fogalmak egy-egy előfordulása adja a jelentésreprezentációt.

A jelentésreprezentáció generálásában rejlő igazi probléma természetesen nem az, hogy egyes szavak jelentését megadjuk, hanem az, hogy reprezentálhatók legyenek azok a szintaktikai viszonyok, amelyek az egyes szavakat, morfémákat értelmes mondattá szervezik, azaz a régens-bővítmény viszonyokat. Ezek reprezentálása két módon történhet, de mindkét módon a többértelműség problémájával kell megküzdenünk.

1. **vonzatkeret reprezentálása** A vonzatviszony olyan reláció, amely szóelőfordulások közt értelmezett – ennek az ontológiában a szereprelációk felelnek meg. Ez a megfeleltetés nem univerzális: például a nominatívusz vonzat – bár általában az *ágense* vagy *elszenvedője* szereprelációnak felel meg – jóformán bármely szereprelációnak megfelelhet, az orvosi szövegekben például sokszor a *lokalizáció*-nak. Ezért egyértelműsített lexikonelemekként külön ad-

juk meg a vonzatkeretet: a vonzatoknak relációkat feletetünk meg, amelyek kijelölik a lexikonelem megfelelő szerepreláció(já)t. Például, a „fáj" szó nominatívusz vonzata a FÁJDALOM fogalom *lokalizáció* szereprelációjának felel meg. Ha egy szóhoz több különböző vonzatkeret is járulhat, ezeket különkülön egyértelműsített változatként kezeljük.

2. **szabad határozók** A legtöbb rag, amely a magyarban a szabad határozókat jelzi, általában megfeleltethető az ontológiában szereplő relációknak: pl. a -bAn ragnak egyrészt az *ideje*, másrészt a *helye* reláció egy-egy fajtája. Ez az egyik olyan terület, amelyen az OWL által nyújtott definíciós lehetőségek nem elegendőek ahhoz, hogy az ontológiában leríjuk egy nyelvi elem pontos értelmezését.

Természetesen vannak olyan nyelvi elemek, amelyeknek az ontológiában nem feleltethető meg referátum. Ilyenek a névelők, névmások, tagadószók stb. Ezeket az alaktani és szintaktikai elemzéshez csatolódva külön szabályokkal kell kezelnünk (l. 5. fejezet).

4 A kórlapkitöltő rendszer

A kifejlesztett rendszer feladata a kórházi panaszfelvételek során lejegyzett magyar nyelvű szövegek egységes kórlap-reprezentációvá alakítása. A kórlapnak tartalmaznia kell a beteg adatait, a rá vonatkozó panaszfelvételek körülményeit, és az egyes panaszfelvételek során leírt tüneteket, panaszokat, azok tulajdonságait, fellépésük és megszűnésük idejét, ill. az esetleg fontos egyéb körülményeket.

Egy szöveg jelentésének meghatározása a morfológiai és szintaktikai törvényszerűségek figyelembe vétele nélkül elképzelhetetlen. Ugyanakkor a nyelvi (elsősorban szintaktikai) elemzés is csak akkor lehet elegendően pontos, ha kiegészül szemantikai információval. Ebből a szempontból ideális helyzet az lenne, ha a morfológiai, szintaktikai és szemantikai elemzés egyszerre lenne elvégezhető. Erre a magyar nyelv esetében is van példa [1]. Egy ilyen elemző megvalósítása hosszabb távú terveinkben szerepel, addig azonban a morfológiai és szintaktikai elemzést előfeldolgozásnak tekintjük.

A szöveg nyelvi előfeldolgozását a MorphoLogic kft. morfológiai és szintaktikai elemzője [4] végzi. A szintaktikailag elemzett szöveget XML formátumú elemzési fa formájában kapjuk meg. Az elemzés fejközpontú: minden nagyobb egység, csoport, frázis (lehetőség szerint) hivatkozik arra a terminális elemre, amelyik a fejének tekinthető. A terminálisok releváns morfológiai jegyei is fel vannak tüntetve, ezek nagymértékben segítik az egyes szövegelemek közötti szemantikai relációk felderítését.

A projekt keretében rendelkezésre álló orvosi szövegek esetében különös jelentősége van a szövegnormalizálásnak. Kezelni kell a szakterületre jellemző idegen (jellemzően latin) szavakat, a rövidítéseket ill. ezek legkülönfélébb változatait, a számokat és a szövegek sietős lejegyzése miatt feltűnően gyakori hibákat, elgépeléseket. A jelenleg működő rendszer egyelőre normalizált bemenettel dolgozik, a szövegnormalizálást a következő fázisban szintén a MorphoLogic-kal közösen valósítjuk meg.

Szoftverrendszerünk tehát tartalmazza a nyelvi előfeldolgozót, a jelentésreprezentációt létrehozó szemantikai elemzőt, és a reprezentációból a kórlap egyes mezőinek megfelelő információt kiolvasó kórlapkitöltőt. A nyelvi előfeldolgozó saját szótári és

nyelvtani adatbázissal rendelkezik, a szemantikai elemző adatbázisa pedig a fent tárgyalt ontológia. Az ontológiát a Protégé ontológiaszerkesztővel szerkesztjük, és a Java nyelven írott programkódunk a Jena fejlesztői programcsomag segítségével használja.

5 Eddigi eredményeink

A jelentésreprezentáció ontológiában való megvalósítása számos problémát vet fel. Ezek egy részét az ontológia megfelelő kialakításával és használatával kezelni tudjuk, más esetekben viszont ki kell lépnünk az ontológia megszabta keretek közül, és külön kell kezelnünk a problémát. Reményeink szerint a korábban említett lexikalista nyelvtanba integrálódva ezek az esetek is a többivel együtt lesznek kezelhetők.

5.1 Egyértelműsítés, anafora-feloldás

A szabad szövegek elemzésekor rendre felmerülő egyik jelentős kihívás a rendszerben több szinten megjelenő többértelműség. A nyelvi elemzés eredménye akár a lexikális, morfológiai, akár a szintaktikai szinten is többértelmű lehet (l. a címben szereplő „Hol fáj?” kérdést). A lexémáknak az ontológiában több konstrukció is megfelelhet, amelyek különböző fogalmakra referálnak.. A fogalmak közötti kapcsolatrendszerben is többféle reláció állhat fenn két fogalom között.

Egyfajta többértelműségnek tekinthetjük az anaforák jelenlétét is, hiszen általában több, korábban előforduló fogalom közül kell kiválasztanunk, hogy melyikre utalnak. Ilyenkor a már feldolgozott fogalmak közül az anafora által meghatározott tulajdonságúakat keressük ki.

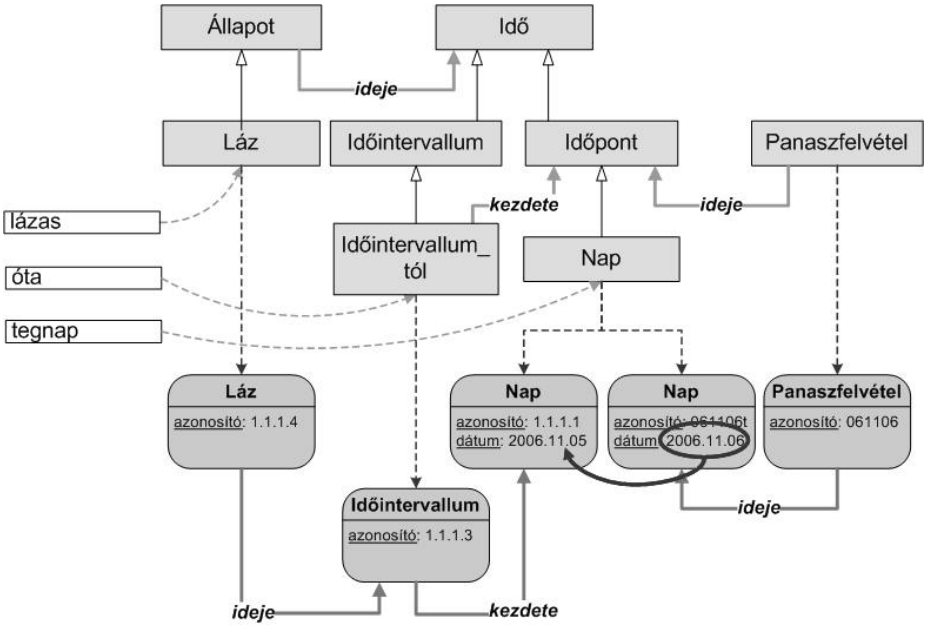
Az egyértelműsítés alapja minden esetben az, hogy melyik reprezentáció a teljesebb abban az értelemben, hogy a szövegből több információt tud kinyerni. Ennek mérésére a fogalmak közti relációkat aszerint értékeljük, hogy az általuk összekötött előfordulások szempontjából mennyire specifikusak. Azt a reprezentációt választjuk ki, amelyik minél több és minél specifikusabb relációt tartalmaz. Az egyértelműsítést praktikus okokból mindig a lehető legalacsonyabb szinten végezzük el.

5.2 Ismeretlen szavak

Bármekkora ontológiát is építünk, a feldolgozandó szöveg szükségképpen fog tartalmazni ismeretlen szavakat. Megtehetnénk, hogy ezeket kihagyjuk a jelentésreprezentációból, de előfordulhat, hogy éppen egy ismeretlen szó forrasztja össze a jelentés egyébként szétesőő részeit, hiszen a mondat szintaktikai szerkezete egyértelműen kijelölheti a szerepét. Éppen ezért az ismeretlen szavak reprezentálására a VALAMI csúcskategória egy előfordulását hozzuk létre, illetve az ismeretlen kapcsolatok jelölésére bevezettük a *dummyReláció* nevű relációt.

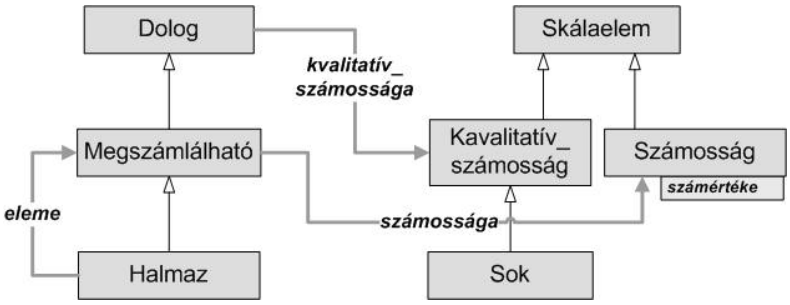
5.3 Idő és számosság

A szöveg sokszor egyáltalán nem vagy csak az igeidővel jelzi, hogy az egyes események vagy állapotok mikor következnek be, ill. állnak fenn. Máskor pedig relatív időhatározók vannak, pl. „ma”, „két napja” stb.



2. ábra A „Tegnap óta lázas.” mondatnak megfelelő ontológiarészlet és jelentésreprezentáció. A „tegnap óta” kifejezést egy időintervallum reprezentálja, amelynek kezdete az a nap, amelynek dátuma egy nappal előzi meg a panaszfelvétel napját.

A kórlap kitöltéséhez azonban szükséges, hogy a lehető legpontosabban meghatározzuk az egyes panaszok kezdetének, fennállásának és/vagy végének az idejét. Ezért az algoritmus tartalmaz egy modult, amely minden panaszhoz a megfelelő időt próbálja hozzárendelni a panaszfelvétel idejének ismeretében. Ennek támogatására az ontológiában a feladathoz illeszkedően terveztük meg az idő ábrázolását (l. 2. fejezet).



3. ábra A számosság és a halmazok ábrázolása az ontológiában.

Az egynél nagyobb számosságú objektumok kezelése is kihívást jelent az ontológiaszerkesztők számára. Ebben a projektben az a megoldás tűnt a leginkább kezelhetőnek, hogy minden megszámlálható dolog egyben halmaz is lehet, amelynek elemei ugyanolyan típusú objektumok, mint önmaga. Egy előfordulás azáltal válik halmazzá, hogy vagy a számosságát vagy az elemeit meghatározzuk. A számmal kifejezhető SZÁMOSSÁG mellett megkülönböztetjük a KVALITATÍV_SZÁMOSSÁGot (pl. sok, kevés) is (l. 3. ábra).

5.4 Ontológián kívüli módszerek

A természetes nyelvi szöveg számos olyan elemet tartalmaz, amelyek nem a világmodell fogalmaira referálnak, hanem a többi fogalom jelentését vagy egymáshoz való viszonyát módosítják. Ezért létrehoztuk a nem tartalmas szavak szótárát, amely egyszerű szintaxissal megírt utasításokat ad a szemantikai elemzőnek az egyes szavak kezelésére vonatkozóan.

A legegyszerűbb esetben az adott szóval egyáltalán nem foglalkozunk (pl. kötőszók, „csak" stb.). Ilyenkor vagy a szintaktikai elemzés hordozza a megfelelő információt (pl. mellérendelés a kötőszók esetében), vagy a szóban rejlő információt nem kívánjuk reprezentálni (pl. szubjektív minősítés a „csak" esetében).

A tagadószókat és fosztóképzőket már a nyelvi előfeldolgozás során megjelöljük, és ezt figyelembe véve kell a létrehozott példányok igazságértékét megállapítani. Szintén még az előfeldolgozás során a mellérendelő szerkezeteknél, felsorolásoknál a közös bővítményeket a felsorolás minden elemére kiterjesztjük.

Vannak speciális jelentéssel bíró kifejezések, amelyek az adott mondat komplex kezelését kívánják meg. Például az „egyéb panasza nincs" kifejezés azt jelenti, hogy a panaszok halmaza (amelynek korábbi panaszok az elemei) lezárt, abba újabb elemet nem lehet felvenni.

A 4. ábra egy konkrét példába sűrítve mutat meg néhányat a fenti jelenségek közül.

6 Hogyan tovább?

A jelenleg elkészült kórlapkitöltő rendszer számos kérdésre adott választ, de az itt bemutatott víziónk szempontjából egyelőre csak egy demó, amelynek számos korlátja van. További feladatok az alábbi területeken adódnak.

A rendszer jelenlegi architektúrája hosszabb távon nem tartható: a szintaktikai és a szemantikai elemzés szétválasztása a módszer szellemével ütközik. Az elsődleges kutatási feladat olyan nyelvtan kidolgozása, amely akár a szemantikus elemzéssel párhuzamosan, egymást segítve képes szöveget elemezni. A szintaktikai és a szemantikai elemzés párhuzamosítása hatékonyabbá tenné a többértelműségek feloldását. Mindenképp valamilyen lexikális nyelvtanra gondolunk, így azokat a szabályokat, amelyeket most procedurálisan építünk a szemantikai elemzésbe, leíró szabályokként használhatnánk.

```
"Tegnap óta lázas, és fáj a torka, ma sokat köhögött. Egyéb panasza
nem volt."
Panaszfelvétel ("") [id=Test_20061106]
  :eredményezi-->
    Panasz ("panasza") [id=Test_20061106_1.2.1.3]
      :ideje-->
        Nap ("ma") [id=Test_20061106_1.1.1.15]
          :dátuma--> 2006.11.06.
      :zárt-e--> true
      :eleme-->
        Köhögés__Állapot ("köhögött") [id= Test_20061106_1.1.1.17]
        Fájdalom ("fáj") [id=Test_20061106_1.1.1.8]
        Láz ("lázas") [id=Test_20061106_1.1.1.4]
    Köhögés__Állapot ("köhögött") [id=Test_20061106_1.1.1.17]
      :aktora-->
        Páciens ("") [id=Test]
      :kvalitatív_számossága-->
        Sok ("sokat") [id=Test_20061106_1.1.1.16]
      :ideje-->
        Nap ("ma") [id=Test_20061106_1.1.1.15]
          :dátuma--> 2006.11.06.
    Fájdalom ("fáj") [id=Test_20061106_1.1.1.8]
      :ideje-->
        Időintervallum_tól ("óta") [id=Test_20061106_1.1.1.3]
          :kezdőpontja-->
            Nap ("Tegnap") [id=Test_20061106_1.1.1.1]
              :dátuma--> 2006.11.05.
      :lokalizációja-->
        Torok ("torka") [id=Test_20061106_1.1.1.12]
    Láz ("lázas") [id=Test_20061106_1.1.1.4]
      :ideje-->
        Nap ("Tegnap") [id=Test_20061106_1.1.1.1]
          :kezdi-->
            Időintervallum_tól ("óta") [id=Test_20061106_1.1.1.3]
          :dátuma--> 2006.11.05.
  :elszenvedője-->
    Páciens ("") [id=Test]
  :ideje-->
    Nap ("") [id=Test_20061106_time]
      :dátuma--> 2006.11.06.
```

4. ábra Egy példaszöveg reprezentációja a program által adott karakteres kimeneten.

A jelenleg használt Protégé ontológiaszerkesztő adottságai sok mesterséges megoldást kényszerítettek ránk, amelyek kissé elbonyolították az ontológia szerkezetét. Majd ha a MEO projekt során fejlesztett ontológiaszerkesztő [6] használatra kész lesz, sokkal természetesebb ontológiát használhatunk.

Valódi alkalmazás rendkívüli méretű ontológiát igényel – egy ilyen megkonstruálása külön több éves projekt. Olyan megoldáson gondolkodunk, amely – legalábbis néhány alkalmazás esetén – nem kíván teljes ontológiát; illetve az alkalmazás használata során épülne az ontológia. A legfontosabb ilyen alkalmazás olyan keresőprogram kifejlesztése lenne, amely a kereső-kifejezés jelentésének megfelelő találatokat adna.

Bár a bemutatott rendszer is bizonyítja módszerünk alkalmazhatóságát, számos kérdés vethető fel bonyolultabb szövegek elemzésénél. A módszer továbbfejlesztése is előttünk áll.

Bibliográfia

1. Alberti G., Balogh K., Kleiber J., Viszket A.: A totális lexikalizmus elve és a GASG nyelvtan-modell. In: Maleczki M. (szerk.): A mai magyar nyelv leírásának újabb módszerei V. SZTE, Szeged (2002) 193–218.
2. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic annotation, indexing, and retrieval. Journal of Web Semantics 2 (2005)
3. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: Ontology Library. WonderWeb Deliverable D18. URL: www.loa-crn.it/Publications.html
4. Merényi Cs., Tihanyi L.: A MetaMorpho fordítóprogram projekt 2006-ban. In: MSZNY 2006, Szeged (2006)
5. Nagypál, G.: Improving Information Retrieval Effectiveness by Using Domain Knowledge Stored in Ontologies. OTM Workshops 2005, LNCS 3762, Springer-Verlag (2005) 780-789.
6. Szakadát I., Szőts M., Gyepesi Gy.: MEO - Ontology Infrastructure. In: Magyar G., Knapp G., Wojtkowski, W., Wojtkowski, G., Zupancic, J., Wrycza, S. (eds.): Advances in Information Systems Development: New Methods and Practice for the Networked Society, Proceedings Information Systems Development, Springer, in press
7. Szőts M., Lévay Á.: Szerepfogalmak az ontológiában – az OntoClean metodológia továbbfejlesztése. In: MSZNY 2005, Szeged (2005)
8. Tomassen, S.L.; Gulla, J.A.; Strasunskas, D.: Document Space Adapted Ontology: Application in Query Enrichment. Proceedings of 11th International Conference on Applications of Natural Language to Information Systems (NLDB'2006), Klagenfurt, Austria, LNCS 3999, Springer-Verlag (2006) 46-57.

Argumentumstruktúrák gépi azonosítása (Szemantikai modul a Hunpars elemzőhöz)

Babarczy Anna[1], Gábor Bálint[1], Hamp Gábor[2], és Rung András[3]

[1] Kognitív Tudományi Tanszék, BME, 1111 Budapest, Stoczek u. 2.
{babarczy, bgabor}@cogsci.bme.hu

[2] Szociológia és Kommunikáció Tanszék, BME, 1111 Budapest, Stoczek u. 2.
hampg@eik.bme.hu

[3] Nyelvtudományi Intézet, MTA-ELTE, 1068 Budapest, Benczúr Gy. u. 33.
runga@artitude.hu

Kivonat: A Hunpars projekt folytatásaként a III. MSzNy Konferencián bemutatott mondattani elemző alkalmazásunkat szemantikai modullal egészítjük ki A fejlesztés elsődleges célja tagmondatszintű szemantikai tudások beemelésére, a frázisstruktúra tematikai címkézése. Erre a célra egy strukturált vonzatkerettárat fejlesztünk. Bár első lépésben az igék argumentumszerkezetére helyezzük a hangsúlyt, a fejlesztés olyan általános elméleti alapokon nyugszik, melyekkel bármely predikátumfunkciót betöltő nyelvi elem kezelhető. Az argumentumszerkezetek leírásában a Role and Reference Grammar fogalomrendszerét ötvözzük a FrameNet projekt módszereivel és a Konstrukciós nyelvtan egyszintű, lexikalista filozófiájával.

1 Bevezetés

A Hunpars-projekt folytatásaként a III. MSzNy Konferencián bemutatott mondattani elemző alkalmazásunkat [1] szemantikai modullal egészítettük ki. A Hunpars jelenleg implementált moduljai automatikusan végzik bármilyen értelmezhető magyar mondat szintaktikai elemzését. Az elemző a frázisstruktúra nyelvtanok alapelveinek felhasználásával a mondat szavait hierarchikus szerkezetekbe, frázisokba szervezi, és szintaktikai jegyekkel felcímkézett, zárójelezett mondatszerkezetet ad kimenetként.

A projekt második szakaszában célunk elsősorban tagmondatszintű szemantikai tudások beemelése, a frázisstruktúra tematikai címkézése. Tematikai címkézés alatt egy olyan rendszert értünk, melyben a mondatban szereplő predikátumokhoz tartozó főnévi, illetve határozói frázisokat tagmondatbeli szerepüket jelölő címkékkel látunk el. A fejlesztés jelenlegi szakaszában a hangsúly nem annyira az elemzés technikai részletein, mint inkább az erőforrásként használt tematikai nyelvtan kidolgozásán van. A projekt olyan általános elméleti alapokon nyugszik, melyekkel bármely predikátumfunkciót betöltő nyelvi elem kezelhető, bár első lépésben az igék keretrendszerének kidolgozása a cél. A generatív nyelvelméletben elterjedt, a tematikus szerepek fogalmi keretére épülő megközelítést a klasszikus funkcionalista igetipológiával [2,

3], a konstrukciós nyelvtan lexikalista elveivel [4] és ezek korpusznyelvészeti alkalmazásaival [5] társítjuk.

2 A tematikai elemző-modul szerkezete

Az elemző elsődleges erőforrása a vonzatkerettár. A vonzatkerettár szerkezeti felépítése az 1. ábrán bemutatott példán látható. A tár alapeleme a Frame, melyet egy meghatározott vonzatkeret definiál. A vonzatkeret meghatározásában a morfoszintaktikai és a tematikai jegyek azonos súllyal szerepelnek. Egy-egy Frame egy vagy több, a vonzatkeretébe illeszthető nyelvi elemet, vagy konstrukciót, foglal magába – ezeket lexikális tételnek nevezzük. Egy lexikális tétel állhat egyetlen szóból, de lehet szónál kisebb elem (például igekötő, képző) vagy többszavas kifejezés is (idióma, kollokáció). A tematikai elemzés a Frame és a zárójelezett, morfoszintaktikailag annotált mondat illesztéséből áll.

FRAME	**Ösztönöz**	
	LEXICAL ENTRIES	ösztönöz bátorít buzdít biztat unszol sarkall
	ACTOR	ösztönző <CAS<NOM>>
	UNDERGOER	ösztönzött <CAS<ACC>>
	NON-MACROROLE	goal <CAS<SBL>>; <CLAUSE<SUBJ-IMP>> manner <CASE<INS>>

[*A vállalatvezetők*]$_{\text{ACTOR}}$ [a magasabb profit érdekében] [minden évben] [*prémiummal*]$_{\text{MANNER}}$ [***ösztönzik***] [*jobb munkára*]$_{\text{GOAL}}$ [*a dolgozókat*]$_{\text{UNDERGOER}}$

PERIPHERY	location <CAS<INE;…>> source <postpp/ÉRDEKÉBEN;…> (…)

1. **ábra** Frame és periféria a vonzatkerettárban

2.1 A tematikus szerepek

A Frame-et definiáló vonzatkeret leírása a Van Valin nevéhez fűződő Role and Reference Grammar (RRG) [3] fogalmaira épít. A RRG megkülönböztet két kitüntetett szerepű argumentumot, azaz két makroszerepet: az ACTOR-t és az UNDERGOER-t. A két tematikai funkciót a vonzatkeret formális logikai szerkezete alapján határozzuk meg: informálisan fogalmazva, az actor az esemény aktív szereplője, míg az undergoer a viszonylag passzív résztvevő. Bár egy tipikus tranzitív szerkezetben az actor az alany, az undergoer pedig a tárgy szintaktikai funkcióhoz rendelhető, ettől a mintától eltérő vonzatkeretek is előfordulhatnak (pl. az állapotot vagy visszaható eseményt kifejező egy-argumentumú igék alanya UNDERGOER funkciót tölt be, míg a személytelen igék dativus argumentuma bizonyos esetekben ACTOR szerepet kap).

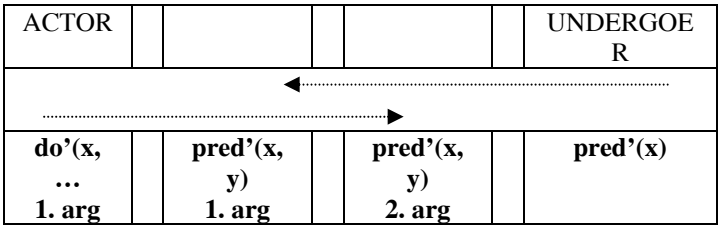
A predikátum logikai szerkezetét az Aktionsart típusa határozza meg. A Hunpars szempontjából lényeges Aktionsart kategóriák és a hozzájuk rendelt intranzitív és tranzitív logikai struktúrák az (1) táblázatban láthatók. A táblázatot követő példák a táblázat sorait illusztrálják sorrendben.

1. Táblázat: Logikai szerkezetek Aktionsart típus szerint

Aktionsart	Logikai szerkezet
állapot	**pred'**(x) vagy (x, y)
atelikus cselekvés	**do'**(x, [**pred'**(x) vagy (x, y)]
állapot változás	BECOME **pred'**(x) vagy (x, y)
telikus cselekvés	**do'**(x, [**pred'**(x, (y))]) & BECOME **pred'**(z, x) vagy (y)

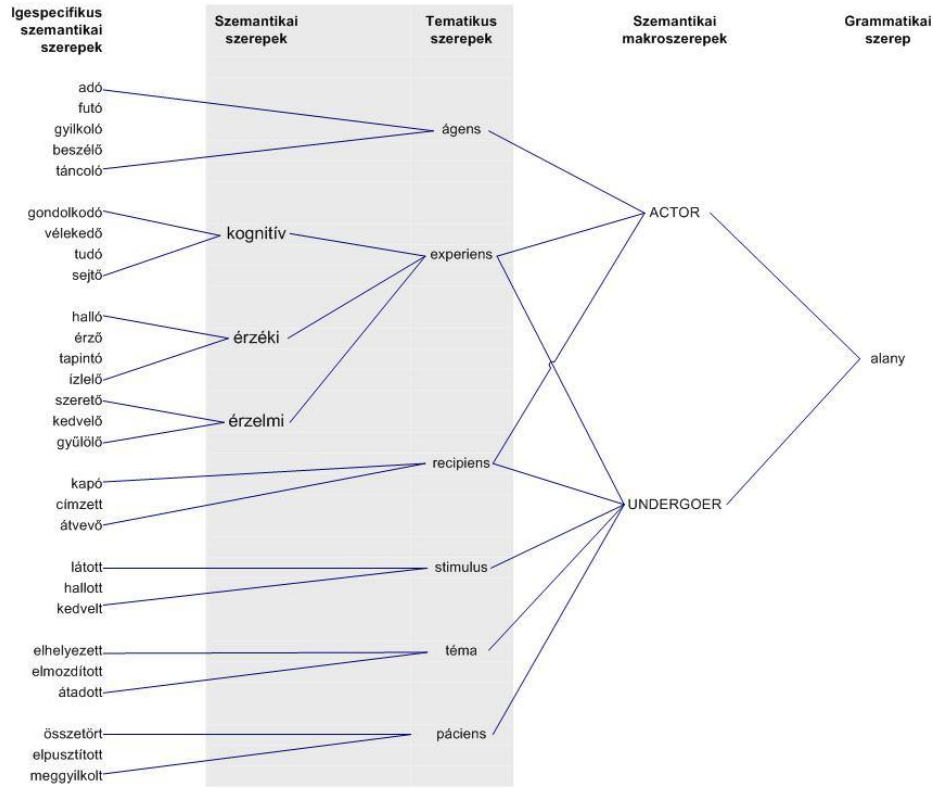
1. Vilmos álmos. Vilmos ágyban van.
2. Edit szánkózik. Feri almát hámoz.
3. Vilmos felébredt. Edit a lejtő aljára ért.
4. Edit visszaliftezett a pálya tetejére. Feri kenyeret süt.

A két makroszerep kiosztása az actor-undergoer hierarchia szerint valósul meg, amint a 2. ábrán látható:



2. ábra. Az ACTOR és az UNDERGOER makroszerepek kiosztását vezérlő hierarchia. Forrás: [9], 177. o.

A logikai szerkezetben előforduló egyéb, nem-makroszerepű argumentumokat tematikai címkékkel jellemezzük.. Döntő kérdésnek tekintettük, hogy hol kell meghúzni a tematikai kategóriák határait. A Hunpars célja a kategóriák számának minimalizálása: új címkét csak akkor vezetünk be, ha ez szükséges ahhoz, hogy kifejezhető legyen egy-egy predikátum argumentumai közötti kontraszt. Ezzel a döntéssel a nehezen definiálható, sokszor „megérzéseken" alapuló tematikai osztályok alkalmazásából adódó kategorizációs problémákat igyekeztünk elkerülni, illetve azokat minimálisra csökkenteni. (Lásd 3. ábra)



3. ábra. A szemantikai szerepek hierachiáját mutatja a [8] (31. o.) alapján készült ábra, amelyen a szürkével jelölt tartomány nem jelenik meg Hunpars-mondatelemzésben.

Annak, hogy az elemzés csak kevés szerepcímkét használ, van néhány következménye. Például a *vár* ige tárgyesetű és -rA ragos bővítménye azonos szerepcímkét kap, mert mindkettő nem fordulhat elő egy argumentum szerkezetben. A -tÓl ragos argumentumot ezzel szemben megkülönböztetjük, mivel szerepelhet a tárggyal együtt:

5. Mit$_{[\text{UNDERGOER}]}$ vármy?
6. Mire$_{[\text{UNDERGOER}]}$ várs?
7. *Mire mit várs?
8. Mit$_{[\text{UNDERGOER}]}$ várs tőle$_{[\text{SOURCE}]}$?

Bár a *vár* ige esetében a két UNDERGOER címkével jelölt argumentum kétségtelenül hasonló tematikai funkciót tölt be a mondatokban, a tárgyesetű és valamilyen oblique esetű alternatívák azonos címkézése nem szemantikai, hanem disztribúciós alapon történik. A *tud* ige tárgy és oblique argumentumai éppen ezért két különböző tematikai címkét kapnak:

9. Mit$_{[\text{UNDERGOER}]}$ tudsz?
10. Miről$_{[\text{THEME}]}$ tudsz?
11. Miről$_{[\text{THEME}]}$ mit$_{[\text{UNDERGOER}]}$ tudsz?

Ezek alapján az ACTOR és az UNDERGOER mellett hatféle tematikus szerepet különböztetünk meg: CO-ACTOR, MANNER, SOURCE, LOCATION, GOAL, THEME, és ezek mindegyikének egy absztrakt és egy konkrét típusát, ami szükség illetve lehetőség szerint elválasztja például a téri elhelyezkedést (konkrét) az időbeli elhelyezkedéstől (absztrakt), vagy az eszközhasználatot (konkrét) a módtól (absztrakt):

12. Lassan$_{[\text{MANNER/ABSTRACT}]}$, nagy gonddal$_{[\text{MANNER/ABSTRACT}]}$ írt.
13. Lassan$_{[\text{MANNER/ABSTRACT}]}$ írt a tollal$_{[\text{MANNER/CONCRETE}]}$.
14. Lassan$_{[\text{MANNER/ABSTRACT}]}$ írt nagy gonddal$_{[*\text{MANNER/CONCRETE}]}$.
15. Nagy gonddal$_{[\text{MANNER}]}$ írt.

A fenti példákban a vonzatkeretnek megfelelően a módhatározót absztrakt módként címkézzük. Ha ez mellérendelő viszonyban áll egy –vAl esetű NP-vel, az utóbbi is absztrakt címkét kap, hiszen nagyvalószínűséggel azonos szerepű argumentumokat koordinálunk (4). Ha a két argumentum nincs mellérendelő viszonyban, a –vAl ragos főnevet eszköznek tekinthetjük, bár ez ritka esetben téves elemzéshez vezethet (6). Magában álló –vAl esetű NP semleges MANNER címkét kap (7), itt csak a szavak vagy a szövegkörnyezet részletes szemantikai elemzése dönthetne a két értelmezés között.

A fenti nyolc tematikus szerepet kiegészíti egy alacsonyszintű argumentum-leírás. Az *ösztönöz*-keretben az actor talán triviálisan az *ösztönző* leírást kapja, az undergoer pedig az *ösztönzött* lesz attól függetlenül, hogy a Frame melyik igéje szerepel a mondatban. Ennek jelentőségét a későbbiekben tárgyaljuk.

Ennyiben az elemző lexikális alapú. A magyar NooJ tematikai moduljához [6] hasonlóan, a lexikális megkötéseket egy default mapping rendszer egészíti ki, amely a magyar nyelv sajátosságainak megfelelően elsősorban esetragokra, névutókra és határozó típusokra épül, de bármilyen más formai feltételrendszer kifejezésére is alkalmas. A rendszer beépül a tematikai modul Frame szerkezetébe: a default mintát egy predikátum-független supra-lexikális Frame-nek tekintjük. A Frame típusok magyarázatát ld. alább. A default címkék szerepe kettős: egyrészt a vonzatkerettárban nem szereplő predikátumokat tartalmazó mondatok elemzését teszik lehetővé, másrészt a szabad határozók definiálásának kritériumát adják és ezek annotálását végzik. Szabad határozónak tekintünk minden olyan NP-t és határozót, amely az esemény logikai szerkezetén kívül esik, azaz amelynek a morfológiai vagy szintaktikai jegyei egy olyan default tematikai szerepet határoznak meg, amely bármilyen mondatba funkcióváltozás nélkül beilleszthető. Ezek a szereplők/körülmények a perifériába tartoznak, és nem játszanak szerepet a Frame definiálásában.A szabad határozók és

argumentumok megkülönböztetésében James Pustejovsky [7] definícióját követjük: eszerint valódi szabad határozónak csak olyan mondatrészek tekinthetők, amelyek szabadon előfordulhatnak bármilyen mondatszerkezetben. Ide tartoznak tehát az eseményt időben és térben elhelyező határozók és az esemény okát megnevező határozók. Bármilyen más bővítmény argumentumnak tekintendő, attól függetlenül, hogy a megnevezése kötelező-e, hiszen csak a mondat adott predikátuma engedheti az előfordulását, vagyis beépül a predikátum logikai szerkezetébe.

2.2 A Frame és lexikális tételek

Ez természetesen nem kell hogy azt jelentse, hogy minden cselekvést kifejező ige vonzatkeretébe külön-külön vesszük fel például a módhatározó argumentumot. Egy-egy Frame leírása, a konstrukciós nyelvtan filozófiáját követve, bármilyen kötöttségi szinten megvalósítható. Lehet a Frame viszonylag szűken definiált, mint például az 1. ábrán látható *ösztönöz*-keret, amelybe csak néhány (de legalább egy) lexikális tétel illeszthető, de alkothatunk tágan definiált, alulspecifikált ige-osztályokra épülő kereteket is, és sublexikális morfémákat jellemző kereteket is. Az előbbire példa a cselekvést jelentő vagy a mozgást jelentő ige-keret, ahol a vonzatkeret meghatározása csak olyan jegyeket tartalmaz, amelyek minden cselekvést illetve mozgást jelentő igét jellemeznek. Az utóbbi típusba az argumentumstruktúra szempontjából meghatározó igekötők és képzők keretei tartoznak. Mivel egy keretbe olyan lexikális tételek illeszthetők, amelyek közös tematikai szerepekkel és közös morfo-szintaktikai szerkezettel rendelkeznek, és a tematikai szerepek maximális hatáskörűek, a keret-tagság lehetséges létszámát a morfoszintaktikai megkötések szigorítása vagy lazítása vezérli. Ennek megfelelően a morfoszintaktikai feltétel rendszer kötetlen: a pontos szóalaktól kezdve a legtágabb szófaji kategóriáig minden szinten specifikálhatjuk a Frame vonzatainak formai követelményeit.

Egy lexikális tétel lehet egy vagy több Frame instanciája is. Ha egynél több kerethez tartozik, a tematikai elemzést a keretek összevonása adja, a következő szabályok szerint. (1) A Frame-eknek három típusa van: lehetnek alapszintűek, bővítők vagy módosítók. (2) Alapkeretnek tekintjük a legszűkebben definiált lexikális Frame-eket, vagyis azokat, melyek lexikális tételei megengednek azonos alacsonyszintű argumentum-leírásokat, azaz szinonimáknak tekinthetők. Egy lexikális tétel legfeljebb egy alapkeret instanciája lehet. (3) Az alapszintnek meg nem felelő Frame-ek operátor keretek, amelyek bővítik és/vagy módosítják az alapkeret adott vonzatstruktúráit. Az operátorkeretek nem határoznak meg önálló alacsonyszintű argumentum-leírásokat. (4) Az igeosztályokat leíró Frame-ek és a szupralexikális default Frame bővítő operátor keretek. Ha egy lexikális tétel egy alapkeret és egy vagy több bővítő keret instanciája is, a tematikai elemzést a keretek unifikációja határozza meg, azaz a bővítő keretben talált új szerepű argumentumokkal kiegészítjük az alapkeret által meghatározott bővítményeket. Az alábbi (leegyszerűsített) példa az *ír* ige keretének az összetételét mutatja: az első sorban meghatározott végső Frame az ige alapkeretének (16a), egy transzfer igeosztály bővítőkeretének (16b) és a cselekvés igeosztály bővítőkeretének (16c) unifikációjából áll össze.

16. ír{ír} [ACTOR/író<NOM>, UNDERGOER/szöveg<ACC>, GOAL<DAT>, MANNER<INS>]
  a. ír{ír} [ACTOR/író<NOM>, UNDERGOER/szöveg<ACC>]
  b. benefit{olvas, ír, …} [ACTOR/<NOM>, UNDERGOER<ACC>, GOAL<DAT>]

c. cselekvés{eszik, ír, …} [ACTOR/<NOM>, UNDERGOER<ACC>, MANNER<INS >]

A vonzatkerettár felépítése lehetővé teszi a gyors fejlesztést: besorolhatunk lexikális tételeket igeosztály Frame-ekbe, anélkül, hogy alapkereteket határoznánk meg. Ilyen esetben a bővítő keret(ek)ből jön létre a predikátum Frame-je. A viszonylag széles eseménykörben előforduló vonzatok egyszerű feltérképezése és a vonzatkerettár gyors fejlesztése mellett a bővítő operátor keretek a ritka vagy szokatlan vonzatszerkezetek azonosításában is szerepet játszanak. A *megkérdez* ige alapszintű Frame-je például a <CAS<DEL>> (*-rÓl*) morfoszintaktikai jegyet rendeli az esemény THEME argumentumához, és nem illeszkedik az alábbi nem-kanonikus mondatra:

17. A mellékhatások tekintetében kérdezze meg orvosát, gyógyszerészét.

Az ige azonban lehet instanciája a tudás-transfer igeosztály bővítő keretének, ahol a THEME morfoszintaktikai specifikációi megengedőbbek, tartalmazzák a <tekintetében>, <vonatkozólag>, stb. névutós szerkezeteket. A többszintű Frame-alkotás egyrészt gazdaságossági okokra vezethető vissza, másrészt azt a pszicholingvisztikai folyamatot próbálja implementálni, miszerint egy lexikális konstrukció rögzített szerkezete bizonyos körülmények között közeledhet a lexikális tétel szemantikai szomszédainak szerkezeti felépítéséhez, illetve analógiás úton kölcsönözheti ezek egyes elemeit.

(5) Az operátor keretek másik típusa egyben bővítő és módosító operátor. Az itt leírt argumentumok nemcsak új bővítményekkel egészíthetik ki az eredeti keretben megadott bővítményeket, hanem módosíthatják is azokat. Ezek a keretek alkalmasak az igekötők és a képzők vonzatkeretre gyakorolt hatásának leírására: egy ilyen sublexikális keret instanciája lehet egy vagy több sublexikális morféma (képző vagy igekötő). Ebben az esetben önmagukban nem alkotnak végleges Frame-t, hanem az elemzés alatt álló mondat predikátumának töve által előhívott alapkerettel és/vagy osztálykeret(ek)kel együttesen határozzák meg az elemzés vonzatstruktúráját. A sublexikális operátor keretben meghatározott argumentumok felülírják a lexikális keret azonos tematikai szerepű argumentumainak morfoszintaktikai specifikációit:

18. fürdet{fürdet, mosdat, ...}[ACTOR<NOM>, UNDERGOER/fürdő<ACC>]
   a. fürdik{fürdik, mosdik, …}[UNDERGOER /fürdő<NOM>]
   b. tat{VERB<CAUS>}[ACTOR<NOM>, UNDERGOER<ACC>]

(6) Amennyiben konfliktus merül fel az alkalmazható Frame-típusok között, vagyis több illesztési lehetőség is van, a „nyertes” Frame-kombináció az lesz, amelyik az elemzés alatt álló mondat legnagyobb számú argumentum frázisára illeszthető.

3 Összefoglalva

A Hunpars-elemzőben alkalmazott tematikai címkézés rugalmasnak tekinthető. Érveltünk amellett, hogy ez a rugalmasság rendszerelméleti szinten is józan próbálkozásnak látszik. Emellett azonban praktikus megfontolássokkal is lehet indokolni a Hunparsban alkalmzott megoldást. Míg az automatikus szemantikai elemzés precizi-

tásának növelése munka- és időigényes szűken definiált kereteket kíván, a használható lefedettség eléréséhez észszerű egy gyorsabb fejlesztési folyamat lehetőségét is megteremteni. Ennek érdekében terveink között szerepel az igék automatikus keretbesorolása morfoszintaktikailag elemzett korpuszból kinyert statisztikai minták alapján.

Bibliográfia

1. Babarczy, A., Gábor, B., Hamp, G., Kárpáti, A., Rung, A., Szakadát, I.: Hunpars: mondattani elemző alkalmazás. In: III. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2005) 20–28.
6. Gábor Kata, Héja Enikő: Vonzatok és szabad határozók szabályalapú kezelése. In: III. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2005) 245-256.
4. Goldberg, Adele**:** Constructions. A Construction Grammar approach to argument structure. Chicago: University of Chicago Press (1995)
7. Pustejovsky, James: The generative lexicon. Cambridge, MA: MIT Press (1996)
8. Van Valin, Robert: An introduction to syntax. Cambridge: CUP (2001)
9. Van Valin, Rober & LaPolla, Randy: Syntax. Structure, meaning and function. Cambridge: CUP (1997)
2. Vendler, Zeno: Linguistics in philosophy. Ithaca: Cornell Univ. Press (1967)
3. http://linguistics.buffalo.edu/research/rrg.html
5. http://framenet.icsi.berkeley.edu

Szemantikai igeosztályok tesztelése az MNSz-ben

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 nyolcvanas évektől megnőtt a lexikon szerepe a szintaxiselméletekben. Egyre több megközelítés feltételezte, hogy az ige lexikális jelentése nagymértékben meghatározza azt a szintaktikai környezetet, amelyben az ige előfordulhat. Ezt az előfeltevést elfogadva, hosszú távú célunk egy olyan szabályalapú szintaktikai elemző fejlesztése, amely az igével egy tagmondatban szereplő esetragos főnévi csoportok annotálása során az ige releváns szemantikai tulajdonságaira támaszkodik. Jelen cikkben a Magyar Nemzeti Szövegtár adatai alapján azt vizsgáltuk, hogy milyen mértékben alkalmazhatók a magyarra Levin [6] szemantikai metapredikátumokkal definiált angol igeosztályai.

1. Bevezetés

Cikkünkben arra a kérdésre keressük a választ, hogy alkalmazhatók-e Levin [6] igeosztályai egy szabályalapú magyar szintaktikai elemzőben. Munkánk arra a feltevésre épít, hogy az angol vonzatkeret-alternációkat univerzálisan érvényes szemantikai metapredikátumok irányítják, melyek – legalábbis bizonyos szempontok szerint – a magyar nyelvben is egymáshoz hasonló viselkedésű igecsoportokat jelölnek ki. Levin az angol igék argumentumrealizációs lehetőségeit vizsgálva az igék szintaktikai tulajdonságaiból indul ki, és ebből következtet közös lexikai szemantikai tulajdonságaikra. Mi ezt az irányt megfordítva arra voltunk kíváncsiak, hogy az általa felépített lexikai szemantikai reprezentációt magyar igékre átültetve megkapjuk-e azokat az igeosztályokat, melyek azonos bővítménykeretekkel rendelkeznek. A bővítménykeretet [3] alapján Levinnél tágabban értelmezzük, mivel véleményünk szerint az igék lexikális jelentéskomponensei nemcsak kötelező vonzataik, hanem a mellettük kitehető szabad határozók körét és morfoszintaktikai tulajdonságait befolyásolják. Ezért nemcsak a feltételezett vonzatokat, hanem az igéket módosító, NP kategóriájú szabad határozók egy részét is az ige lexikai szemantikai reprezentációja által meghatározottnak tekintjük.

Hipotézisünket a Magyar Nemzeti Szövegtárból vett adatokon teszteltük. Kiválasztottunk két angol igecsoportot, és a korpuszból olyan mondatokat gyűjtöttünk, melyekben az igék magyar megfelelői állítmányként szerepelnek. Ezután az INTEX/NooJ korpuszfeldolgozó eszköz segítségével a mondatokat tagmondatokra bontottuk, és futtattuk a szövegen a legfelsőbb szintű NP-ket felismerő nyelvtanunkat. Az eredményül kapott tagmondat-vázakon azt vizsgáltuk, hogy melyek azok az

esetragok, melyek az igecsoportok mellett szereplő felső szintű NP-ken leggyakrabban előfordulnak.

A továbbiakban bemutatjuk az argumentumrealizációs elméletek közös előfeltevéseit és a számunkra releváns *predikátumdekompozíciós elmélet* módszereit [2.fej.], majd ismertetjük, hogy milyen szerepet tulajdonítunk a predikátumdekompozíciós jelentésleírásnak a szintaktikai elemzésben [3.fej.]. Ezután bemutatjuk a hipotézis ellenőrzéséhez használt módszerünket [4.fej.] és a vizsgálat eredményeit két, általunk választott igecsoportra [5.fej.]. Végül szót ejtünk a munka lehetséges további irányairól [6.fej.].

2 Az igei jelentés szintaktikai vonatkozásai

A nyolcvanas évektől megnőtt a lexikon szerepe a szintaxiselméletekben. Egyre több megközelítés feltételezte (pl.: LFG [4], GB [1], Role and Reference Grammar [2]), hogy az ige jelentése szerepet játszik az argumentumainak felszíni szintaktikai realizációjában, azaz az igei jelentés nagymértékben meghatározza azt a szintaktikai környezetet, amelyben az ige előfordulhat. Ez a feltevés abból a megfigyelésből indult ki, hogy a hasonló jelentésű igék vonzatkerete hasonló mintát mutat.

A szóban forgó elméletek közös előfeltevése, hogy az ige szemantikai reprezentációjában szereplő argumentumok a szintaxisban vonzatokként realizálódnak. Céljuk olyan szabályszerűségek megállapítása, amelyek alapján megjósolhatóvá válik egy adott argumentum szintaktikai funkciója. Így az argumentumok szintaxisban betöltött szerepét megjósló elméletek sikeressége elsősorban két tényezőn múlik: egyfelől az ige megfelelő szemantikai reprezentációján, másfelől pedig a leképezési eljárás (*linking/mapping theory*) megfelelő kialakításán.

Az ige szemantikai reprezentációja szempontjából az argumentumrealizációval foglalkozó elméletek (legalább) két nagy csoportra oszthatók. Az első csoportba tartozó elméletek feltételezik, hogy az ige szintaktikai funkcióját az ige által kiosztott thematikus vagy szemantikai szerepek határozzák meg. A másik csoportba tartozó elméletek ezzel szemben azt feltételezik, hogy az ige jelentése tovább bontható, és az igei jelentést karakterizáló metapredikátumok felelősek azért, hogy az argumentumok hogyan realizálódnak. Ezeket az elméleteket az alábbiakban predikátumdekompozíciós elméleteknek fogjuk nevezni. Levin [5][60] a predikátumdekompozíciós elméletek mellett érvel a thematikus szerepeket használó elméletekkel szemben.

A predikátumdekompozíciós eljárás célja, hogy az igék jelentésében megtalálja azokat a közös jelentéskomponenseket vagy metapredikátumokat, amelyek az adott igékre jellemző szintaktikai viselkedésért felelősek. Mára általánosan elfogadottá vált az a nézőpont, hogy az igék reprezentációjában el kell választani egymástól a több igére jellemző általánosabb jelentéskomponenseket és az ige jelentésének idioszinkratikus részét. Így tehát az ige szemantikai reprezentációja az általánosabb, metapredikátumokkal karakterizált eseménytípus és a beágyazott idioszinkratikus rész kompozíciójaként áll elő. Ekkor azonban már nem tartható az a feltevés, hogy az ige jelentése *kizárólag* az ige argumentumainak – a vonzatoknak – a szintaktikai megvalósulásáért felelős. Mint tudjuk, a predikátum jelentése, pontosabban a predi-

[60] 2.2, 5. fej., 6. fej.

kátum jelentéséből elvonható eseményséma meghatározhatja, hogy milyen típusú időhatározók jelenhetnek meg a mondatban.

Azt gondoljuk tehát, hogy az ige jelentése nemcsak a vonzatainak felszíni megjelenését determinálja, hanem – az eseménységén keresztül – a szabad határozók esetleges megjelenését is az ige engedélyezi. Egy ige akkor engedélyezi egy szabad határozó megjelenését a mondatban, ha az igei jelentésben van egy olyan metapredikátum, amely *kompatibilis* a szabad határozó jelentésével. Az elképzelés alapján nemcsak az időhatározók disztribúcióját jósolhatjuk meg. További példát szolgáltatnak az állapotváltozást jelentő mozzanatos igék, amelyek mellett a *–ra* esetragos főnév megadhatja az ige által leírt cselekvés közvetlen előzményét („*puskalövésre elindult*"). Ezzel szemben, ha a mozzanatosság metapredikátuma nincs jelen az ige szemantikai reprezentációjában, akkor a szóban forgó szemantikai szerep sem jelenhet meg, hiszen nincsen olyan metapredikátum, amely engedélyezné. (Pl.: *"*puskalövésre üldögélt*").

A predikátumdekompozíciós elméletek hátránya, hogy nehéz megtalálni az ige szintaktikai viselkedése szempontjából releváns metapredikátumokat. Levin[1993] szerint a metapredikátumokat az igecsoportok szintaktikai viselkedéséből kiindulva lehet feltérképezni. Előfeltevése szerint azok az igék rendelkeznek azonos metapredikátumokkal, amelyek azonos vonzatkeret-alternációkban vesznek részt.

Ezt az elképzelést [5] a kauzatív alternációval szemlélteti.

The wind opened the door.	*The door opened.*
A szél kinyitotta az ajtót.	Az ajtó kinyílt.
Mary broke the window.	*The window broke.*
Mari betörte az ablakot.	Az ablak betört.
Bill cooled the soup.	*The soup cooled.*
Bill lehűtötte a levest.	A leves lehűlt.

Levin szerint a kauzatív alternáció arra a szemantikai tényre vezethető vissza, hogy vannak olyan igék, amelyek jelentésreprezentációjában az ige által leírt esemény külső okának van argumentumhelye. Külső ok alatt Levin az állapotváltozást elszenvedő dologtól független entitást ért, amely közvetlen hatást gyakorol a tárgy által jelölt entitásra (a fenti mondatokban '*a szél', 'Bill'* és *'Mary'*).

Ezzel párhuzamosan Levin azt is feltételezi, hogy ha van egy olyan – alternációk által kijelölt – igeosztályunk, amely bizonyos alternációkban egyezik, ám más alternációkban eltér egy másik osztálytól, akkor van olyan jelentéskomponens, amely mindkét igeosztály tagjaiban közös. Ezt az állítást a hangkibocsátó igék ([6] 43.2, 234.o.) példájával szemlélteti. A csoportba tartozó igék egy része részt vesz a kauzatív alternációban, míg egy másik része nem:

The train rumbled.	**Peter rumbled the train.*
A teherautó zakatol.	*Péter zakatoltatta a vonatot.
The tea kettle whistled.	**The boiling water whistled the tea kettle.*
A teáskanna fütyült.	*A víz fütyültette a teáskannát.

The teacups clattered.	*I clattered the cups as I loaded the sink.*
A csészék csörögtek.	Csörgettem a csészéket, ahogy ... .
The windows rattled.	*The storm rattled the windows.*
Az ablak zörgött.	A szél zörgette az ablakot.

Ha a kauzatív alternációt megalapozó szemantikai bázis helyes volt, akkor annak alapján magyarázhatónak kell lenniük a fenti adatoknak is. Ha az alternációt az ige által jelölt esemény külső ok szerepű argumentuma teszi lehetővé, akkor az elmélet azt jósolja, hogy azon igék esetében nem lesz jól formált az alternáció tranzitív tagja, amelyeknél a hangkibocsátás oka az eseményben résztvevő entitás belső, inherens tulajdonságainak a következménye. A fenti példák pontosan ezt igazolják.

3 A jelentésreprezentáció szerepe a szintaktikai elemzésben

Munkánk kiindulópontjaként Levin [6] szolgál, melyben a fenti módszer alapján meghatározott angol igeosztályok szerepelnek. A vizsgálat eredményét egy szabályalapú magyar szintaktikai elemző [9] továbbfejlesztésében szeretnénk hasznosítani, elsősorban a legfelsőbb szintű mondat-konstituensek és az igei állítmány közti dependencia-viszonyok automatikus felismeréséhez. Mint azt [3]-ban ismertettük, a szintaktikai elemzés nélkülözhetetlen részének tartjuk a szemantikai igeosztályokra való hivatkozást, mert véleményünk szerint a magyarban a névszói bővítmény és az ige közti *szintaktikai* függőségi reláció jellege (azaz a névszó vonzat/adjunktum státusza) sem dönthető el az ige bizonyos szemantikai tulajdonságainak ismerete nélkül. Míg az adjunktumok közös tulajdonsága, hogy egy-egy igecsoport mellett ugyanolyan formai jegyekkel (esetraggal) jelölve ugyanazt a szemantikai szerepet töltik be, a vonzatok szemantikai szerepe mindig egyedi, és nem határozható meg az ige jelentésére való hivatkozás nélkül. A cikkben megválaszolandó fő kérdés, hogy a Levin által meghatározott – végső soron szemantikailag definiált – angol igeosztályok mennyiben mutatnak a magyarban is hasonló szintaktikai viselkedést, azaz mennyire hasonlít az a szintaktikai környezet, amelyben előfordulnak. A *szintaktikai környezet* fogalmát tágabban értjük, mint Levin: az igével előforduló valamennyi főnévi csoportot megvizsgáljuk. Mint a 2. fejezetben már említettük, véleményünk szerint az ige szemantikai tulajdonságai, vagyis az igei jelentés releváns komponensei elsősorban éppen az adjunktum szerepű NP-k disztribúciójáért felelősek. Amennyiben bebizonyosodik, hogy a Levin által használt igeosztályok magyar megfelelői is osztoznak szintaktikai tulajdonságokban, úgy az igeosztályokra érvényes általánosításokat fogalmazhatunk meg szabályok formájában, melyeket beépíthetünk a szintaktikai elemzőbe.

4 A vizsgálat módszere

Kiinduló hipotézisünk szerint tehát a közös jelentéskomponensekkel (metapredikátumokkal) jellemzett igék egymáshoz hasonló szintaktikai környezetekben fordulnak elő. Két további feltételezéssel élünk:

1) Levin igeosztályait *univerzálisan* érvényes szemantikai metapredikátumok határozzák meg;

2) Ezek az igeosztályok átültethetők magyar nyelvre, vagyis az adott igék jelentésének megfelelő magyar igék szintén egységes csoportot fognak kijelölni.

Természetesen várható, hogy egyes, az angolban egységes csoportok a magyarban különbözőképp viselkedő alcsoportokra oszlanak, illetve a szemantikailag homogén osztályokon belül is számíthatunk lexikai kivételekre. További kérdést jelent, hogyan értelmezzünk a korpusz adatait annak megítélésében, hogy bizonyos szerkezetek *lehetségesek* a magyarban, hiszen abból, hogy egy szerkezet nem fordul elő az MNSz-ben, nyilvánvalóan nem következtethetünk arra, hogy nem is létezik a nyelvben. Mindazonáltal a jelen cikkben leírt munka kereteiben azt vizsgáltuk, hogy a metapredikátumokkal definiált igecsoportok elemei jellemzően egymáshoz hasonló (illetve a többi csoporttól eltérő) szintaktikai környezetben fordulnak-e elő a korpuszban. Így csak azokat a szintaktikai mintákat vettük figyelembe, melyek egy-egy predikátum előfordulásainak legalább tíz százalékában megjelennek.

A feltételezéseink alapján azt várjuk, hogy ha az igeosztályok a magyarban is releváns csoportokat alkotnak, akkor elemeik osztoznak egyfelől legalább egy jelentéskomponensben, másrészt – ebből következően – engedélyezik egy vagy több, az osztályra specifikus szemantikai szerepű és esetragú NP megjelenését. Ha a csoportok igéi valóban egységes viselkedést mutatnak, akkor a szintaktikai elemzőt olyan szabályokkal bővíthetjük, melyek kimenete nemcsak felismeri, hanem szemantikai szerepet jelölő címkékkel is ellátja az igék egyes bővítményeit.

A hipotézist a Magyar Nemzeti Szövegtár [7] adatai alapján teszteljük. Az első szakaszban az ige bővítményeként szereplő esetragos főnévi csoportok disztribúcióját vetettük össze.

Az MSZNY 2005 konferencián elhangzott előadásunkban [3] az instrumentális (*-val*) esetrag lehetséges szemantikai szerepeit vizsgáltuk. Ezek között találtuk – többek között – a *mentális állapotváltozás oka* (‘*megdöbbent vkit* <u>*vmive’l*</u>) és az általunk *nem default eszköz*nek (’*beszennyez vmit* <u>*vmivel’)*</u> nevezett szemantikai szerepeket. Az alábbiakban azt a két igeosztályt fogjuk megvizsgálni, amelyek ezeket a szemantikai szerepeket engedélyezik a mondatban. A korpuszból lekérdeztük azokat a mondatokat, melyek az igék valamelyikét tartalmazzák, és olyan mintát állítottunk össze, mely igénként legfeljebb 3,000 mondatot tartalmaz. Az eredményül kapott szöveget az Intex korpuszfeldolgozó eszköz [7] magyar moduljával [9] részleges szintaktikai elemzésnek vetettük alá. A mondatokat tagmondatokra bontottuk, a legfelsőbb szintű NP-ket tagekkel helyettesítettük, melyek csak az NP esetragját/névutóját kódolják. Az így kapott tagmondat-vázakból azokat vettük figyelembe, melyek állítmányként (finit alakban) tartalmazzák az adott igecsoport valamelyik elemét. Ezután minden igére megszámoltuk a vele azonos tagmondatban szereplő esetragos NP-k előfordulásait.

5 Eredmények

5.1 Mentális állapotváltozást jelentő igék

A mentális állapotváltozást jelentő igéket [6] 2.13.4 igeosztályával azonosítottuk, amely osztályt a tranzitív birtokos alany alternáció definiálja:

Mark terrified me with his singlemindedness.
Mark megijesztett engem a korlátoltságával.
Mark's singlemindedness terrified me.
Mark korlátoltsága megijesztett.

A 221 elemű angol igeosztály elemeit összesen 77 magyar igének feleltettük meg. Az osztályba tartozó angol igék tranzitívak. Magyarra fordításukkor azonban úgy döntöttünk, hogy átalakítjuk őket intranzitívvá, mert azt találtuk, hogy a magyar intranzitív alakok morfológiailag egyszerűbbek, a tranzitív alak belőlük képezhető (pl. *megijed-megijeszt*). Az igék fordításánál kritérium volt, hogy az igékhez tartozó eseményséma minél hasonlóbb legyen az eredetihez. Ha több lehetőség közül kellett választani, a nem kauzatív, befejezett aspektusú változatot részesítettük előnyben.
A korpuszból egy 13,023 mondatból álló mintát vettünk, melyek mindegyike tartalmazza az osztály valamelyik igéjét. Az igék többségénél azt tapasztaltuk, hogy leggyakrabban bővítmény nélkül, illetve egy (nominatívuszi) bővítménnyel fordulnak elő. Ez feltehetőleg annak tudható be, hogy a mentális állapotváltozást jelentő igék leginkább személyes közlésekben, vagyis az MNSz internetes fórum alkorpuszában szerepelnek, melyre a sok rövid tagmondatból álló, sok igét tartalmazó, hosszú mondatok jellemzőek. Az esetragok gyakoriságának vizsgálata során azokat az igei lemmákat vettük figyelembe, amelyek ötnél többször fordultak elő a korpuszban. Az esetragok között sorrendet állítottunk fel annak alapján, hogy hány különböző igei lemmával szerepelnek együtt a lemma előfordulásainak legalább 10 százalékában. (A nominatívuszt ezen a ponton nem vizsgáltuk.)
Ennek eredményeképpen az alábbi sorrendet kaptuk:

1. Táblázat: Gyakori esetragok a mentális igék mellett

Eset	Igék száma
INE	12
SUP	12
ABL	10
SUB	7
INS	5
DAT	2

Amint a táblázat mutatja, a leggyakoribb esetragok is csak az igék 15 százaléka mellett fordulnak elő legalább az esetek 10%-ában, tehát nem állíthatjuk, hogy egységes viselkedésű csoportot találtunk. Ezért megvizsgáltuk, hogy az esetraggyakoriság alapján milyen lehetséges alcsoportokat állapíthatunk meg, és hogy ezek szemantikailag koherens osztályokat alkotnak-e. Ha az igei lemmákat a mellettük

előforduló leggyakoribb esetrag szerint csoportosítjuk, akkor több esetragra (SUP, ABL, DAT) is olyan alosztályokat kapunk, ahol a szóban forgó esetrag szemantikai szerepe ugyanaz: minden osztály tagjai az ige jelentésében foglalt állapotváltozás okát kódolják az adott raggal[61].

Inesszíivusz (-ban):
elfásul, kifárad, kimerül, elfárad
Datívusz (-nak):
m*egörül, megörvend*
Ablatívusz (-tól) és szuperesszívusz (-on) váltakozik:
m*egrémül, felélénkül, elborzad, megrészegül, megijed, fellelkesül ,elkábul, elhűl, meghökken, ,meglepődik, elképed, megsértődik, megütközik, stb.*
(Össz.: 25 ige)

Az adatok tehát arra utalnak, hogy a mentális állapotváltozást jelentő igék a magyarban szűk alcsoportokra oszthatók az alapján, hogy milyen esetraggal fejezik ki az *ok* szemantikai szerepű bővítményüket. Ezek közül az INE és a DAT esetraggal járó osztályok egy-egy szinonimacsoportot alkotnak, mag az ABL/SUP váltakozással jellemzett osztály jelentés szempontjából is heterogén. További vizsgálatra szorul, hogy milyen igei jelentéskomponensek irányítják az esetragok közti választást, és vajon ezek a jelentéskomponensek a magyarban tényleg csak az adott alcsoportot határozzák-e meg.

5.2 "Spray/load" igék

Az alcímben szereplő igeosztály talán a vonzatkeret alternációkkal foglalkozó szakirodalom legtöbbet tárgyalt jelensége. Levin felosztásában ez a 2.3.1 osztály. Az alternációt az alábbi példák szemléltetik:

a) *John loaded the truck with hay .*	*John loaded hay on the truck.*
János megrakta a szekeret szénával.	János szénát rakott a szekérre.
b) *John sprayed the wall with paint.*	*John sprayed paint on the wall.*
János befújta a falat festékkel.	János festéket fújt a falra.

Amint a példamondatok magyar megfelelői is mutatják, az eltérő bővítményszerkezet a magyar igék esetében gyakran igekötő megjelenésével jár együtt. Ezért feltételeztük, hogy az angol igék magyar megfelelőinek vizsgálatába érdemes bevonni az igék igekötős változatait is. A 49 elemű angol igeosztálynak 51 igekötő nélküli magyar igét feleltettünk meg, és ezek igekötős változatainak előfordulásait is lekérdeztük. Az igekötős igék bevonása mellett szól továbbá, hogy az igecsoport elemeinek jelentése tartalmazza a *helyváltoztatás* jelentéskomponenst, ami alapján feltételezhetjük, hogy az irányt jelölő igekötők – legalábbis sok esetben – produktívan kapcsolódnak az igéhez, átlátszó szerkezetet eredményezve.

[61] Az INE toldalék az esetek többségében idő vagy hely szerepű szabad határozót jelöl, de kijelöl egy jól körülhatárolt szemantikai alosztályt is, amely mellett *ok*ot fejezhet ki.

Az igekötős változatokkal együtt összesen 295 ige előfordulásait kérdeztük le, így 59,941 mondatot vizsgáltunk. Várakozásunk az volt, hogy az angol load-spray alternáció magyar megfelelője nagy számban fog előfordulni:

V	N.ACCi	INSj	*Vki tölt vmit vmivel.*
V	N.ACCj	LOCi	*Vki tölt vmit vhová.*

Ahol a LOC irányt jelölő esetrag (ILL, SUB vagy ALL (-hoz)).

Az igék osztályként jelentős eltérést mutatnak a mentális állapotváltozást jelentő igékhez képest. Ezt úgy vizsgáltuk, hogy ebben a csoportban is kiválasztottuk azokat az esetragokat, amelyek valamely ige mellett legalább az előfordulásaink a 10 százalékában szerepelnek, és rangsort készítettünk olyan szempontból, hogy hány ige mellett érik el ezt a gyakoriságot. Nem foglalkoztunk az egyetlen ige mellett szereplő esetragokkal, valamint az alany- és tárgyesettel sem. Várakozásunknak megfelelően azt találtuk, hogy az INS a leggyakoribb esetrag, de majdnem ugyanilyen gyakori az ILL és a SUB együttvéve. Érdekes, hogy a harmadik LOC szerepű rag, az ALL sehol nem éri el a 10% gyakoriságot.

2. Táblázat: Gyakori esetragok a spray/load igék mellett

Eset	Igék száma
INS	98
SUB+ILL	96
INE	57
SUB	53
ILL	43
SUP	41

A táblázat első két sora arra utal, hogy az angol spray-load alternáció magyar megfelelője jellemző erre az igecsoportra. Érdemes azonban azt is megvizsgálni, hogy a két szintaktikai minta ugyanazon igék mellett váltakozik-e, vagy a különböző bővítménykeretek az adott ige igekötős változataihoz kötődnek. Azt mondhatjuk, hogy 28 alternáló igét találtunk (amely mindkét mintában szerepel), 70 ige csak az ACC+INS, míg 51 csak az ACC+LOC mintában fordul elő. Ahol az alapige és igekötős változatai is megjelennek a mintázatban, ott a legtöbb esetben az alapige az egyetlen, amely mindkét mintával kompatibilis és 10% fölötti gyakorisággal elő is fordul bennük (pl. *locsol*, *dobál*, *töm*, *ken, permetez, borít*)[62]. Igekötős változataik a két minta közül legfeljebb egyben fordulnak elő 10%-nál gyakrabban. A *meg* és a *tele* – mint nem irányt jelölő igekötők – mindig az ACC+INS szerkezetet hívják elő. Az irányt jelölő igekötők esetében nem általánosítható, hogy melyik szerkezetet részesítik előnyben, például a *ki* a borít igével az ACC+LOC, a töm igével az ACC+INS mintát, a *be* a ken mellett az ACC+INS, a szúr mellett az ACC+LOC mintát hívja elő. Kérdés, hogy

[62] Az INS és a LOC típusú bővítmény váltakozása jellemző a *szúr* igére is, de abban különbözik az említettektől, hogy a LOC típusú bővítmény az INS-es szerkezetben nem tárgyként, hanem ILL (*-ba*) esetben jelenik meg.

az alapigék szemantikai tulajdonságai alapján megjósolhatók-e az igekötő-használat szabályai, vagy az igekötők nem átlátszó összetétellel kapcsolódnak ezekhez az igékhez.

6 Konklúzió és további teendők

A két csoport igéi mellett előforduló főnevek esetragjainak gyakorisága a csoportok között jelentős eltérést mutat, ami igazolni látszik azt a feltevést, hogy az igék lexikai szemantikai tulajdonságai befolyásolják a felszíni szintaktikai környezetüket. Mindazonáltal a leggyakoribb esetragokról sem mondható el, hogy a csoport igéinek többsége mellett előfordulna. Különösen igaz ez a mentális igék csoportjára, melyek mellett tagmondatonként csak átlagosan egy NP szerepel a korpuszban (az alanyesetűeket is beleértve). Az esetragok előfordulásainak alacsony száma arra is rávilágít, hogy a magyarban kevés a kötelező vonzat[63]. A mentális állapotváltozást jelentő igékről bebizonyosodott, hogy nem érdemes őket szintaktikailag egységes csoportként kezelni, ám találhatunk szűkebb, szemantikailag és szintaktikailag is koherens alcsoportokat. Kérdés, hogy ezek jelentéséből elvonható-e olyan metapredikátum, melyet érdemes beépíteni a lexikai szemantikai reprezentációba, mert más szintaktikai jelenségek leírásában is használhatók, vagy az igék csak idioszinkratikus szemantikai és szintaktikai tulajdonságaikban hasonlítanak. Az általánosítás lehetőségét az veti fel, hogy az igeosztály alcsoportjai ugyanazt a szemantikai szerepet fejezik ki a mellettük leggyakoribb esetragokkal.

A spray/load igék 10 százaléka részt vesz az angol csoportra jellemző alternációban, és összesen több mint 50 százalékra jellemző a két minta legalább egyike. Ahhoz, hogy igeosztályként elfogadjuk ezt a csoportot, először meg kell vizsgálnunk, miért nem mutat hasonló adatokat a csoport igéinek másik része: valóban nem vehetnek részt az alternációban, vagy esetleges másfajta használatuk/másik jelentésük torzította a korpusz adatait?

A szemantikai igeosztályok tesztelésekor elsősorban arra voltunk kíváncsiak, hogy találunk-e szintaktikai hasonlóságot az angol adatok alapján egy osztályba sorolt igék között. A két igeosztály eredményeinek összehasonlítása alapján erre igennel válaszolhatunk, ám az osztályokon belüli egyes igék közti eltérések és a minták előfordulásainak csekély száma mindenképp azt vonja maga után, hogy az osztályok kialakításánál kézi ellenőrzésre is szükség van. Tervezzük egyéb igeosztályok magyarra fordítását és a csoportok szintaktikai viselkedésének tesztelését is. A vizsgálatoktól azt várjuk, hogy kiderüljön, vannak-e és melyek azok az igei jelentéskomponensek, amik a magyar igék szintaxisának leírásában hasznos általánosítások megragadására szolgálhatnak. Ha bebizonyosodik az, hogy az angol igeosztályokat karakterizáló metapredikátumok a magyarban is szerepet játszanak az ige szintaktikai környezetének meghatározásában, akkor a szintaktikai elemző számára kulcsfontosságú igeosztályok létrehozásakor támaszkodhatunk az angol csoportokra, és ezáltal szükségtelenné válik a magyar alternációk részletes feltérképezése.

[63] Természetesen találtunk kivételeket, például a “megbékél” ige mellett előfordulásainak több, mint 40 százalékában, a “megbékül” mellett pedig 51 százalékban szerepel az INS esetrag.

Bibliográfia

1. Chomsky, N.: *Lectures on Government and Binding*, Foris, Dordrecht 1981
2. Foley, W. A.., R. D. Van Valin, Jr. *Functional Synatx and Universal Grammar*, Cambridge University Press, Cambridge 1984
3. 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 .
4. Kaplan, R. M. and J. Bresnan:.Lexical-Functional Grammar: a Formal System for Grammatical Representations. In J. Bresnan (ed.) pp. 173-281. [1982]
5. Levin, B., M. R. Hovav: Argument Realization. Cambridge University Press, Cambridge. 2005
6. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. Int. J. Digit. Libr. 1 (1997) 108–121
7. Silberztein, M.: Dictionnaires électroniques et analyse automatique de textes: Le systeme Intex. Masson, Paris 1993.
8. Váradi, T.: The Hungarian National Corpus. Proceedings of the Third International Conference on. Language Resources and Evaluation, 2002. Las Palmas pp.385-389
9. Váradi, T., Gábor, K.: A magyar INTEX fejlesztésről. In: *Alexin Z., Csendes D. (szerk): A Második Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete,* Szeged Egyetemi nyomda, 2004. Szeged, pp. 3-11.

Főnevek szemantikai jegyei és kódolásuk a MetaMorpho projektben

Orosz Kata

MorphoLogic Kft., 1126 Budapest Orbánhegyi út 5.
MTA Nyelvtudományi Intézet – Korpusznyelvészeti Osztály
orosz@morphologic.hu

Kivonat: A cikk a MetaMorpho gépi fordító rendszer magyar-angol változatában a lexikonban található főnevek szemantikai jegyeit, és a jegyek kódolásának folyamatát mutatja be. A cikk első fele a főnévi szemantikai jegyeket osztályozza, valamint tárgyalja, hogy mi a funkciója ezeknek a jegyeknek a MetaMorpho projektben. Ezután a jegyek definiálásával kapcsolatos problémákról, illetve a jegyek egymáshoz való viszonyának meghatározásáról esik szó. A dolgozat a jegykódolási folyamat bemutatásával, a kódolás során felmerülő, részben elméleti, részben gyakorlati problémák illusztrálásával, és az eddig elért eredmények összegzésével zárul.

1 Szemantikai jegyek a MetaMorpho projektben

A MetaMorpho projekt[64] keretében fejlesztés alatt álló gépi fordító rendszer kétféle szemantikai jegyet használ: igei és főnévi jegyeket. (Az igék szemantikai jegyeiről ebben a dolgozatban nem esik szó.) A főnévi jegyek is két csoportba oszthatók: az igei vonzatkeret-leírásoknál használt jegyekre, és olyan jegyekre, amelyeket az igei vonzatkeret-leírások nem, csak a vonzatkereteket működtető szintaxis [1] használ.

1.1 A szintaxis által használt szemantikai jegyek

Az igei vonzatkereteket működtető szintaxis két főnévi szemantikai jegyet használ: az **encnt** és a **gender** jegyeket. Az **encnt** nevű jegy a magyar főnevek angol fordításának megszámlálhatóságáról hordoz információt. Ez a jegy az angol fordítás generálásánál játszik szerepet, amennyiben ez alapján rendeli hozzá a szintaxis az angol főnévhez a megfelelő kvantort (pl. *many* vagy *much*). A jegy kétféle értéket vehet fel: [YES] (megszámlálható) és [NO] (megszámlálhatatlan). A **gender** jegy szintén a generálásnál játszik szerepet; az angol főnevekhez tartozó, az adott szó grammatikai nemének megfelelő személyes névmás kiválasztását teszi lehetővé. Ez a jegy az angol

[64] A MetaMorpho projektről bővebben ld. Tihanyi László: A MetaMorpho projekt története. *I. MSzNy*, Szeged (2003), A MetaMorpho projekt 2004-ben. *II. MSzNy*, Szeged (2004), A MetaMorpho fordítóprogram projekt 2005-ben. *III. MSzNy*, Szeged (2005).

nyelv grammatikai nemeinek megfelelően háromféle értéket vehet fel: [M] (hímnem), [F] (nőnem), [N] (semlegesnem).

1.2 Az igei vonzatkeret-leírásoknál használt szemantikai jegyek

Az igei vonzatkeretek leírásánál 11 szemantikai jegyet használunk. Ezek (ábécérendben): **abstract**, **animate**, **bodypart**, **dynamic**, **company**, **currency**, **human**, **mass**, **measure**, **time**, **weather**. Ezeknek a jegyeknek az a funkciója[65], hogy a fordítandó, magyar nyelvű mondatok elemzése során segítsék a megfelelő igei vonzatkeret kiválasztását, és ezzel csökkentsék a nem megfelelő elemzések számát. (Ebből következik, hogy a kétféle – azaz az igei vonzatkeret-leírásoknál használt, illetve a szintaxis által használt – főnévi szemantikai jegycsoport között az az alapvető különbség, hogy az egyik csoportba tartozó jegyek **/encnt**, **gender/** a generálásban játszanak szerepet, míg a többi 11 jegy az elemzést támogatja.)
Az igei vonzatkeret-leírásoknál használt szemantikai jegyek háromféle értéket vehetnek fel: [YES], [NO] és [NIL]. A főnevek szemantikai jegyének értéke határozza meg, hogy az adott főnév milyen vonzatkereteket képes kiválasztani. Az adott jegyre nézve [YES] vagy [NO] értékkel rendelkező főnevek képesek olyan igei vonzatkeretek kiválasztására, amelyek az adott jegyre nézve a főnévével megegyező értékű szemantikai megkötést tartalmaznak. (A kiválasztás természetesen csak abban az esetben történik meg, ha az igei vonzatkeret-leírás által megkövetelt jegyre nézve [YES] vagy [NO] értékkel rendelkező főnév a fordítandó mondatban a potenciálisan választható igei vonzatkeret meghatározott vonzataként szerepel.) Például:

(1)
***VP=fel|ad**

HU.VP = SUBJ + TV(:lex="fel|ad") + **OBJ**(pos=N, case=ACC, **abstract=NO**, human=NO)

EN.VP = SUBJ + TV[lex=**"post"**] + OBJ

Vmi felad valamit. (János feldta a levelet.)

(2)
***VP=fel|ad**

HU.VP = SUBJ + TV(:lex="fel|ad") + **OBJ**(pos=N, case=ACC, **abstract=YES**)

EN.VP = SUBJ + TV[lex=**"give"**] + PART[lex=**"up"**] + OBJ

Vmi felad abstractvmit.(János feladta a hobbiját.)

[65] A MetaMorpho projektben használt szemantikai jegyek funkciójának definíciója, és az igei vonzatkeretek kiválasztásában játszott szerepük elméleti és technikai kidolgozása Merényi Csaba érdeme.

A fenti példákban a HU.VP-vel kezdődő sorok az adott igei vonzatkeret-leíráshoz tartozó, mmd-formalizmussal [1] kódolt elemzősorok; az EN.VP-vel kezdődő sorok a generálósorok; a dőlt betűvel szedett sorok a vonzatkeret-leírások értelmezését segítő példamondatok (egy általános és egy konkrét példamondat). A főnévi lexikonban a *csomag* [NO], a *remény* [YES] értékkel rendelkezik az **abstract** jegyre nézve. Tegyük fel, hogy a gépi fordítónak az alábbi mondatok fordításához kell kiválasztania a megfelelő igei vonzatkereteket:

(4) Tibi feladta a csomagot.
(5) Tibi feladta a reményt.

A *csomag* és a *remény* is tárgyi vonzata (OBJ) a *felad* igének, tehát elméletileg mindkét főnév választhatná mindkét igei vonzatkeretet, így a következő fordítások születnének:

(4/1) Tibi posted the package.
(4/2) *Tibi gave up the package.
(5/1) *Tibi posted hope.
(5/2) Tibi gave up hope.

Azonban az (1) és a (2) példában a tárgyi vonzat szerepét betöltő főnévre nézve a leírások tartalmaznak egy szemantikai megkötést: az (1) példában csak [abstract=NO] jeggyel rendelkező főnév, a (2) példában csak [abstract=YES] jeggyel rendelkező főnév választhatja ki magának az adott vonzatkeretet. Tehát amikor a gépi fordítónak a *Tibi feladta a csomagot.* mondathoz kell kiválasztania a megfelelő vonzatkeretet, az elemzés során észleli, hogy a tárgyi vonzat szerepét betöltő főnév az [abstract=NO] jeggyel rendelkezik, és ezért az (1) példában szereplő vonzatkeretet választja ki; így csak a helyes fordítást generálja (*Tibi posted the package.*)

Egy főnév akkor vesz fel [NIL] értéket, ha adott szóalaknak poliszémia vagy homonímia következtében több jelentése (és esetleg, de nem szükségszerűen több fordítása) lehetséges. Például:

(6) Az anyák kiborultak.

Az *anya* szónak a (6) mondatban két jelentése is lehet: édesanya (**anya1**, angolul *mother*), és anyacsavar (**anya2**, angolul *screw nut*). Attól függően, hogy melyik értelemben szerepel az *anya* szó, a *kiborul* igének is két különböző jelentése – és ebben az esetben két különböző fordítása – lehetséges: leesik, szétszóródik (**kiborul1**, angolul *spill*) és ideges lesz (**kiborul2**, angolul *become upset*). A **kiborul1** igei vonzatkeretben az alanyi vonzaton az [animate=NO], a **kiborul2**-ben az [animate=YES, human=YES] szemantikai megkötés szerepel. Az *anya* szó a lexikonban (erős egyszerűsítéssel) a következőképpen fog szerepelni:

(7)

```
*NX=anya

HU.NX (animate=NIL) = N(lex="anya")
```

```
EN.NX (animate=YES, human=YES) = N[lex="mother"]

EN.NX (animate=NO) = N#1[lex="screw"] + N[lex="nut"]
```

Amikor a gépi fordító elemzi a (6) mondatot, látja, hogy az alany szerepét betöltő főnév [NIL]-es értékű, ezért mindkét (**kiborul1** és **kiborul2**) vonzatkeretet kiválasztja. Azonban nem mind a négy lehetséges fordítást adja meg (köztük két helytelen fordítással, nevezetesen: **The mothers spilled.* és **The screw nuts became very upset.*), hanem a főnevek megfelelő fordítását a több lehetséges angol generálósor közül (EN.NX) már a vonzatkeret-leírások megkötéseinek megfelelően választja ki, és ezeknek megfelelően csak két fordítást generál: *The mothers became upset*. és *The screw nuts spilled*.

A főnevek szemantikai jegyei akkor tudják igazán növelni az elemzés hatékonyságát (t.i. akkor tudják kiválasztani a megfelelő igei vonzatkeretet), ha a vonzatkeretek leírásánál használt szemantikai megkötések, és a lexikon főnevein kódolt szemantikai jegyek összhangban vannak egymással. Az összehangoláshoz pedig szükség van a jegyek minél pontosabb definiálására.

1.3 A szemantikai jegyek definiálásának problémája

A szemantikai jegyek definiálásánál több probléma is felmerült. Az egyik ilyen dilemma az, hogy a gépi fordító hatékonysága szempontjából nagyságrendileg hány jegyet célszerű bevezetni. A másik, hogy pontosan mik legyenek azok a jegyek, amikre az igei vonzatkeretek hivatkoznak.

A második probléma megoldása statisztikai alapon történt. Az igei vonzatkeretek megírásával foglalkozó munkatársak[66] lettek felkérve, hogy még a pontos jegydefiníciók megalkotása előtt írjanak össze olyan, intuitív (ontológiai) definíciókon alapuló jegyeket, amiket alkalmasnak találnak azonos az azonos alakú, de eltérő jelentésű és/vagy eltérő vonzatkerettel rendelkező igék szemantikai alapú elkülönítésére. A gépi fordító angol-magyar változatában még csak két, intuitív definíción alapuló jegy játszott kitüntetett szerepet: az **animtype** (élő) és **humantype** (akaratlagos cselekvésre képes) jegyek. A magyar-angol változat készítésekor ez a két jegy **animate** és **human** néven tovább élt, azonban további jegyek bevezetésére mutatkozott igény. A vonzatkeret-leírások írói által javasolt jegyek közül végül 9 – **abstract** (elvont fogalom), **bodypart** (testrész), **dynamic** (valamilyen „működésre" képes gép), **company** (intézmény), **currency** (pénznem), **mass** (anyagnév), **measure** (mértékegység), **time** (idő-mértékegység), **weather** (időjárással kapcsolatos jelenség) – került azon az alapon elfogadásra, hogy az igék feldolgozásának akkori fázisában hány igei vonzatkeret-leírásnál érezték szükségesnek az adott jegyre való hivatkozást. Az **1. táblázat** azt mutatja, hogy 2006 júliusában az addig megírt közel 20000 igei vonzatkeret közül hány hivatkozott az adott jegyre.

---

[66] Az igei vonzatkeretek megírásával a Nyelvtudományi Intézet munkatársai foglalkoznak. (A MetaMorpho projekt a MorphoLogic Kft., a Nyelvtudományi Intézet és a Szegedi Egyetem alkotta konzorcium keretei között folyik.)

1. táblázat:
Szemantikai jegyekre való hivatkozás gyakorisága az igei vonzatkeret-leírásokban

A jegy neve	Hivatkozások száma
HUMAN	8682
ABSTRACT	3469
ANIMATE	2836
BODYPART	303
MEASURE	150
TIME	99
DYNAMIC	70
MASS	61
COMPANY	9
CURRENCY	6
WEATHER	1

A hivatkozások gyakorisága mellett fontos szempont volt a hivatkozást tartalmazó igei vonzatkeretek korpuszbeli gyakorisága is – így kerülhetett be a jegyek közé a **company**, a **currency** és a **weather**. Ezek ugyan kevés hivatkozást kaptak, ám az a néhány vonzatkeret, ami mégis hivatkozott rájuk, olyan gyakran használatos a magyar nyelvben, hogy megtartásuk mégis célszerűnek tűnt.

A jegyek listájának megalkotása többé-kevésbé az első kérdésre is megadta a választ; nevezetesen, hogy hány főnévi szemantikai jegyet érdemes bevezetni a gépi fordító hatékonyságának szempontjából. A MetaMorpho projektben érvényesített, az 1.2 pontban vázolt jegyfelfogás nem teszi célszerűvé a jelenleginél nagyságrendileg több főnévi szemantikai jegy bevezetését, hiszen az igei vonzatkeret-leírások szemantikai alapú elkülönítés már a fent említett 11 jeggyel is kivitelezhető. Ezért még ha a projekt folyamán a jegyek listája részben változik is, a jegyek számának radikális növekedése kevéssé valószínű.

A jegyek listájának ilyen módon történő összeállítása egy gyakorlati hátránnyal járt: a jegyek „definíciói” egy kisszámú embercsoport nyelvi intuícióin alapultak, ezért a lexikon főnévi elemeinek kódolásához is csak intuitív támpontot adtak. Ez azonban már átvezet a jegykódolás során felmerült problémákhoz.

2. A jegyek kódolása a magyar-angol lexikonban

A magyar-angol irányú fordítóprogram lexikonába két lépésben kerültek fel a jegyek. Első lépésben a lexikon kb. 72.500 főnevének körülbelül 25%-a lett részben (azaz csak néhány, és nem az összes jegyre) kódolva. Ez nagyrészt az angol-magyar lexikonban **animtype** és **humantype** jeggyel rendelkező szavak magyar megfelelőinek az **animate** illetve a **human** jeggyel való automatikus – számítógépes módszerekkel történő – felruházását jelentette. Szintén az első fázisban, manuális módszerekkel, és lényegében az intuitív jegydefiníciók alapján további 5000 főnév is részleges kódolásra került – ezúttal már a 9 új jegyet is felhasználva.

A második fázisban szükségesnek mutatkozott, hogy a (lehetőségekhez mérten) pontosan definiáljuk, milyen típusú főnevek tartozzanak az adott jegyekhez. A jegyek definiálása végül a WordNet ontológia segítségével, az intuitív jegydefiníciók figyelembevételével történt.

2.1 A jegyek WordNet-alapú definiálása

A WordNet ontológiának mind az angol (Princeton WordNet), mind a jelenleg is fejlesztés alatt álló magyar verzióját [2] felhasználtuk a jegyek definícióinak megalkotásához. A WordNetben olyan csomópontokat kerestünk, melyek az adott jegyhez intuitíve tartozó szavak átfogó kategóriáiként szolgálhattak. Egy jegy jellemzően több, egymással alá-fölérendeltségi viszonyban nem álló csomópontból állt össze. (Néhány jegy WordNet-alapú definícióját a **2. táblázat** tartalmazza.)

A WordNet-alapú jegydefiniálás két szempontból is hasznosnak bizonyult. Egyrészt a WordNet-csomópontokban megadott definíciók alapján egy intuitíve nehezen kódolható szóról a WordNet ontológia segítségével gyorsan és egyszerűen eldönthető, hogy melyik jegy(ek)hez tartozik. A gyakorlatban pl. a VisDic program [2] segítségével ellenőrizhető, hogy adott szó mely WordNet-es csomópontokon „halad át" – ha olyan csomópont alá tartozik, ami egy, a MetaMorpho projektben használt szemantikai jegy definícióját adja, akkor a szó az adott jegyre nézve [YES] értékkel kódolandó.

2. táblázat:

Példa a MetaMorpho projekt szemantikai jegyeinek WordNet-alapú definíciójára

A jegy neve	WordNet Synonym	WordNet ID
ABSTRACT	psychological feature:1; abstraction:6	ENG20-00020333-n; ENG20-00020486-n
ANIMATE	living thing:1, animate thing:1; cell:2	ENG20-00003009-n; ENG20-00004824-n
BODYPART	body part:1	ENG20-04919813-n
COMPANY	organization:1, organisation:3	ENG20-07523126-n
CURRENCY	medium of exchange:1, monetary system:1; monetary unit:1	ENG20-12615184-n; ENG20-12627781-n

A WordNet-alapú jegydefiníciók másik haszna, hogy segítségükkel minden olyan főnévről, ami a WordNet ontológiában és a MetaMorpho projekt magyar-angol főnévi lexikonában *is* szerepel, automatikusan – gépi heurisztikák alkalmazásával – megállapítható, hogy milyen jegyekre nézve kell [YES]-re kódolni; azaz listákat lehet készíteni azokról a szavakról, amik egy adott jegyre nézve [YES] értékkel rendelkeznek. A szavak listázásnál a számítógépes ontológiákban – így a WordNet-ben is – alkalmazott hipernímia-hiponímia relációt [2], [3] használtuk ki.

A jegykódolásnak ebben a fázisában a magyar WordNet főnévi állománya még viszonylag kicsi volt; ezért döntöttünk úgy, hogy a lényegesen nagyobb szóállomány-

nyal rendelkező angol WordNetből is megpróbálunk szemantikai információt kinyerni, és az angol szólisták automatikus fordításával növeljük az adott jegyre nézve [YES] értékre kódolandó magyar szavak listáját. Ez természetesen problémát okozott a poliszém és a homonim szavaknál, ezért az így nyert listákat minden esetben manuálisan kellett korrigálni.

Az angol WordNet alapján készült listák esetében a poliszémia és a homonímia vezetett nem megfelelő szavak listázásához. Például az angol *bay* szó öblöt és pej lovat is jelent. Utóbbi jelentése az angol WordNet-ben az **animate** definíciója alá tartozik, a másik nem. Amikor viszont a WordNet-ből kinyert angol szavak listája gépi módszerekkel össze lett vetve az angol szavak magyar fordításával, nem csak a *pej ló* összetett főnév került az [animate=YES] értékkel bíró szavak listájára, hanem az *öböl* szó is (hiszen a gépi illesztés csak a szóalak azonosságát vette figyelembe).

Szintén manuálisan kellett eltávolítani a listákról azokat a szavakat, amik a WordNet-definíciók és az intuitív definíciók közötti apró eltérések következtében kerültek fel tévesen a listákra. Például a kitalált személyek a WordNet-ontológiákban az **abstract** jegy alá tartoznak, az igei vonzatkeretek szempontjából viszont Pán Péter épp úgy viselkedik, mint Kovács Béla, hiszen mindketten akaratlagos cselekvésre képes lények, azaz inkább a **human** jegyhez tartoznak.

A poliszémiából, homonímiából, és a definíciós különbségekből adódó hibaarány még így is 15% alatt maradt, ezért a manuális korrekció szükségessége ellenére is elmondható, hogy ez az eljárás a kódolási folyamatot nagy mértékben felgyorsította. A **3. táblázat** azt mutatja, hogy hány, a MetaMorpho projekt főnévi lexikonában szereplő szóról tudtuk megállapítani a WordNet ontológiák segítségével, hogy az adott jegyre nézve [YES] értékkel rendelkezik.

3. táblázat:
A WordNet alapján készült listák szavainak száma, jegyekre lebontva

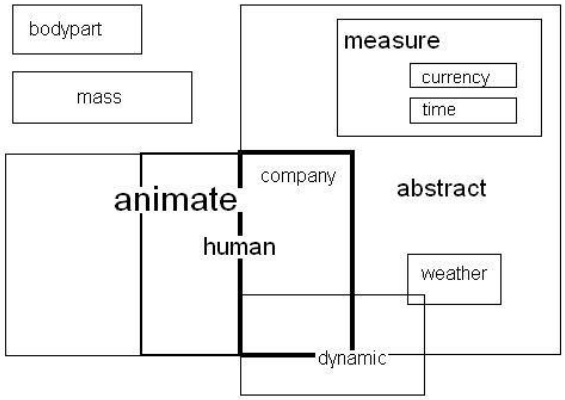
A jegy neve	**Angol WordNet-ből kinyert szavak száma**	**Magyar WordNet-ből kinyert szavak száma**
ABSTRACT	6688	372
ANIMATE	3067	121
BODYPART	663	67
COMPANY	514	42
CURRENCY	90	7
DYNAMIC	4	168
HUMAN	1823	36
MASS	3322	223
MEASURE	748	63
TIME	407	35
WEATHER	88	6
	ÖSSZESEN:	**18554 szó**

Természetesen ez a szám (18554) nem azt jelenti, hogy ennyi szó teljes kódolása megoldódott volna. A szólisták csupán annyit árulnak el, hogy a rajtuk szereplő szavak biztosan [YES] értékűek az adott jegyre nézve; a többi jegy tekintetében viszont

nem hordoznak információt. Pontosabban nem mindegyikük, hiszen ösztönösen is érezhető, hogy a jegyek között átfedés van. A jegyek egymáshoz való viszonyának megállapítása azért vált szükségessé, hogy leírhatóak legyenek az automatikus jegyöröklődések, és ezáltal a jegykódolás folyamata még gyorsabbá és hatékonyabbá váljon.

2.2 A jegyek egymáshoz való viszonyának megállapítása

A jegyek közötti kapcsolatokat a Cruse-féle halmazelméleti szemantikai modellt alkalmazva, az *átfedés, inklúzió, különbözőség* kategóriái [3] mentén fogalmaztuk meg. A jegyek közti kapcsolat megállapításánál figyelembe vettük az igei vonzatkeretek leírásánál már használt, intuitív jegydefiníciókat is. A **1. ábrán** látható a főnevek szemantikai jegyeinek egymáshoz való viszonya.



1. ábra: A MetaMorpho projektben használt szemantikai jegyek egymáshoz való viszonya.

A jegyek inklúziós kapcsolatából az alábbi automatikus jegyöröklések következnek:

(8) **[human=YES] → [animate=YES]**
(9) [company=YES] → [human=YES] → [animate=YES]
(10) [company=YES] → [abstract=YES]
(11) [weather=YES] → [abstract=YES]
(12) **[measure=YES] → [abstract=YES]**
(13) [currency=YES] → [measure=YES] → [abstract=YES]
(14) [time=YES] → [measure=YES] → [abstract=YES]

2.3 A jegyek kódolása: eredmények, további feladatok

A MetaMorpho projekt magyar-angol gépi fordító rendszerének fejlesztése során a szemantikai jegyek kódolása jelenleg harmadik fázisánál tart. Az első fázisban az angol-magyar változatban használt két jegyet alapul véve a magyar-angol változat főnévi lexikonában közel 18000 szó lett kódolva az **animate** és **human** jegyekre; valamint manuális módszerekkel további 5000 szó lett kódolva az időközben bevezetett 9 jegy valamelyikére. A második fázisban a WordNet ontológiák segítségével további 18500 jegyérték került azonosításra; ez az automatikus jegyörökléseket figyelembe véve mintegy 14700 szót érint. Mivel a WordNet alapján készített listákra már csak olyan szavak kerültek fel, melyek semmilyen jegyre nézve nem tartalmazta korábbi információt a MetaMorpho főnévi lexikonában, némi kerekítéssel azt mondhatjuk, hogy a főnevek több, mint 50%-áról (37700 szó) van már valamilyen (kódolt) szemantikai információnk.

Sajnos ez nem azt jelenti, hogy ezzel a kb. 37700 szóval már ne kellene foglalkozni, hiszen a jegyöröklődések csak egy irányban működnek; azaz azoknál a szavaknál, amelyekről csak az **animate** vagy **abstract** jegyek vonatkozásában vannak információink, manuálisan kell egyértelműsíteni, hogy milyen értékkel rendelkeznek a **human**, **company**, **measure**, **time** és **currency** jegyek tekintetében. Ilyen szóból pedig közel 33500 van. Ami azt jelenti, hogy jelenleg a teljes főnévi állományból csak 4200 szónak rendelkezünk a teljes szemantikai leírásával (vagyis adott szóról mind a 11 jegy esetében tudjuk, hogy [YES] vagy [NO] értéket vesz-e fel) – ez a teljes szótári állománynak csupán 6%-a.

A kódolási folyamat második szakaszának fontos eredményeként könyvelhető el ugyanakkor, hogy a szemantikai jegyek WordNet-alapú definiálásával, illetve a jegyek egymáshoz való viszonyának feltérképezésével létrehoztunk egy olyan elméleti keretet, ami a gyakorlatban is megkönnyíti a kódolási folyamatot. A WordNet-ontológiákhoz készített gépi heurisztikák pedig a későbbiekben is felhasználhatóak; tehát a WordNet ontológiák állományának bővülésével újabb és újabb szemantikai információk nyerhetők ki, amik felhasználhatók a jegykódolási folyamatban.

A jegykódolás jelenleg is zajló, harmadik szakaszában a manuális jegykódolás jut főszerephez; először azokat a szavakat ellenőrizzük, amikről az első két kódolási fázis eredményeképpen már rendelkezünk valamilyen szemantikai információval[67]; továbbá zajlik azoknak a szavaknak a (szintén manuális) kódolása, amelyek még teljesen „érintetlenek” ebből a szempontból.

A [NIL] értékű szavak feldolgozása is megkezdődött. A poliszém illetve homonim főnevekről a projekt egy korábbi szakaszában már készültek listák. Az elsődleges feladat ezeknek a listáknak a kibővítése; majd az egy szóalakhoz tartozó összes lehetséges jelentés, és a különféle jelentéseknek megfelelő fordítás megtalálása. Ezután a szójelentéseket egyesével kell kódolni a szemantikai jegyek szempontjából; ezek a szójelentés-sorok végül gépi módszerekkel lesznek „összeadva”, hogy az egyes szójelentések, fordításaik és szemantikai jegyeik a (7) példához hasonló lexikon-elemekké álljanak össze. Pl. az **ülés1** (*seat*) [abstract=NO] és az **ülés2** (*session*) [abstract=YES] sorok egyetlen lexikon-elemmé, az *ülés* [abstract=NIL]-lé fognak összevonódni.

[67] A jegykódolásnak ezt a részét a Szegedi Tudományegyetem Mesterséges Intelligencia Tanszék Kutatócsoportjának munkatársai végzik.

A jegykódolás utolsó szakaszában az addig külön listákon tárolt szemantikai információkat gépi módszerekkel, az mmd-formalizmusnak megfelelően kell kódolni a főnévi lexikonban.

Köszönetnyilvánítás

Szeretnék köszönetet mondani Kiss Gabriellának, Merényi Csabának, Miháltz Mártonnak és Tihanyi Lászlónak, hogy a projektben végzett munkámhoz segítséget nyújtottak, és jelen cikk megírásakor is hasznos tanácsokkal láttak el.

Bibliográfia

1. Merényi, Cs.: A MetaMorpho magyar-angol gépi fordító rendszer igei vonzatkereteit működtető nyelvtan. In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005) 108–115
2. Miháltz, M..: Magyar EuroWordNet projekt: bemutatás és helyzetjelentés. In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005) 68–78
3. Szakadát, I.: Réteges struktúrák, alaprelációk. In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005) 43–55

V. Gépi fordítás

A MetaMorpho fordítóprogram projekt 2006-ban

Tihanyi László, Merényi Csaba

MorphoLogic, 1126 Budapest Orbánhegyi út 5
{tihanyi,merenyi}@morphologic.hu

Kivonat: Az előadásunkban négy témát érintünk. Egyrészt ismertetjük a magyar-angol fordítóprogram nyelvészeti fejlesztésének fontosabb kérdéseit, másrészt beszámolunk a létrejött nyelvi adatbázisok jellegéről és mennyiségéről, majd bemutatjuk az idei évben megvalósult termékfejlesztéseket és szolgáltatásokat, valamint ezek fogadtatását a piacon és az interneten. Végül röviden beszámolunk a jövő évi terveinkről is.

1 Nyelvészeti fejlesztések

Az elmúlt év folyamán a magyar-angol fordítóprogram nyelvtanának fejlesztésére összpontosítottunk. Fő céljaink a mondatszintű nyelvi jelenségek minél szélesebb körének kezelése, valamint az igei vonzatkeretek minél kifinomultabb leírásához szükséges eszközök kifejlesztése volt. Emellett a frázisszintű nyelvtan fejlettsége is elérte azt a szintet, hogy a rendszer a legtipikusabb szerkezetekkel kipróbálható legyen, azaz egyszerű „tankönyvi" mondatokra értékelhető fordítást adjon.

Ebben a részben a magnyelvtan jelenlegi állásáról számolunk be. Magnyelvtannak azt a szabályhalmazt nevezzük, amely általánosságban írja le a magyar nyelvi formákat és azok alapértelmezett fordítását, azaz a nyelvi jelenségeknek egy sematikus generatív nyelvtannal megfogható körét kezelő alrendszert. A magnyelvtan szabályait két csoportba sorolhatjuk. Vannak közöttük hagyományos környezetfüggetlen nyelvtanok szabályaira hasonlító minták, melyek a különböző nyelvi kategóriák illeszkedési lehetőségeit írják le, de nagy számban vannak közöttük olyan „technikai" szabályok is, amelyek nem közvetlenül nyelvi jelenséget reprezentálnak, hanem a rendszer igényei szerint manipulálják a már létrejött reprezentációkat. Ez utóbbi típusra jellemző példa az a tagmondatot reprezentáló kategórián működő szabályhalmaz, melynek feladata a vonzatokat leíró jegycsoportok permutálása.

Ilyen technikai jellegű szabályokra, amelyek tulajdonképpen programszerű működést valósítanak meg, azért van különösen nagy szükség, mert el szeretnénk kerülni a nyelvleírásban a redundanciát. Redundancia két helyen jelentkezhet a rendszerben. Amint az korábbi beszámolóinkból [4] kiderült, a MetaMorpho-formalizmusnak két szintje van. A lexikális erőforrásokat előállító szabályírók a magasabb szintű *mmd* formalizmust használják, ebből egy konverterprogram állítja elő a rendszer számára közvetlenül értelmezhető *mmo* nyelvű leírást. Az egymással szisztematikusan összefüggő nyelvi formákat ezen szintek egyikén sem tartjuk célszerűnek külön szabályok

formájában tárolni. Az *mmd* leírás szintjén azért nem, mert feleslegesen végeztetnénk automatizálható munkát emberekkel, ami nyilvánvalóan gazdaságtalan. Felmerül az a lehetőség, hogy az egymásból levezethető szerkezeteket a konverter generálja ki egy alapalakból, de ennek a megoldásnak is komoly hátrányai vannak. Az *mmo* szabályok száma így feleslegesen megnőne. Ez egyrészt a fordítórendszer végleges méretét olyan nagyra növelheti, hogy az erre épülő termékek terjesztése – különösen az interneten keresztül – nehézkessé válna, másrészt a fejlesztést is nehezítené az ilyen méretű állományokkal való munka lassúsága.

A fent említett okok miatt az igei vonzatkeretekből levezethető szerkezetek kezelésének azt a módját választottuk, hogy az *mmd* és az *mmo* leírás szintjén is csupán egy szabály reprezentálja a vonzatkeretet. Minden olyan szerkezet fordításához, amely ebből levezethető, ezt az egy leírást alkalmazzuk úgy, hogy a ténylegesen előforduló nyelvi alak elemzését és fordítását technikai szabályok segítségével transzformáljuk.

A további fejezetekben példákkal ábrázolva felsorolásszerűen bemutatjuk, hogy milyen mondatszintű jelenségeket elemez és fordít már le a MetaMorpho rendszer, valamint néhány jelenség esetén kissé részletesebben kitérünk arra, hogy milyen technikai megoldások segítségével érjük el azt, hogy az adott nyelvi forma az általános vonzatkeret-leírással fordítható legyen.

1.1 A kezelt nyelvi jelenségek bemutatása

1.1.1 Szabad határozók

Az igei vonzakeretre épülő tagmondatban az ige, a segédigék, a vonzatok és azok kimozgatott összetevői mellett természetesen szerepelhetnek szabad határozók is. Nyelvi modellünk fent említett fő osztályai közül a szabad határozók mondatba való beelemzését hagytuk utoljára. Mivel ezek a topikalizáció és a fókuszba emelés lehetőségeit figyelembe véve a vonzatokhoz hasonlóan szinte tetszőleges mondatpozíción előfordulhatnak, a teljes vonzatkeretet reprezentáló kategória felépítése közben, még a vonzatkeret-azonosítás előtt szükséges gondoskodni beelemzésükről. A vonzatkeret reprezentációjával kapcsolatban lásd [4]. Hasonlóan a többi mondatösszetevőhöz, a szabad határozókat is pointer típusú jegyekben, illetve az azokat kísérő jegyhalmazokban jelenítjük meg. A vonzatkeretek azonosításánál a határozók esetleges jelenléte nem okoz gondot, mivel a VP-t leíró szabály már a [4]-ban leírt, magas absztrakciós szintű VPP nevű kategóriára épít, amelyben minden összetevőt jegyhalmazok képviselnek. A VP-s minta egészen egyszerűen nem tesz utalást a határozókat leíró jegyekre.

A számtalan különféle határozói kifejezés szótári leírása még előttünk áll, de a magnyelvtan már biztosítja a nem szerkezeti esetekben álló, illetve névutós főnévi csoportok alapértelmezett határozói fordítását, (ezeket sok esetben majd szótári szabályoknak felül kell bírálnia). A mondatszintaxis szintjén a szabad határozók kezelését gyakorlatilag megoldottnak tekinthetjük. Következzen néhány példa:

- egyszerű határozószót, illetve névutós NP-t tartalmazó szabad határozók

```
Moose[Hu-En]>tegnap a kutya a ház mögött aludt.
1: [the dog slept behind the house yesterday.]
```

- esetragos, vagy névutós utalószóval álló 'hogy'-os mellékmondat (ennél a példánál az is megfigyelhető, hogy a szabad határozóknak is lehetnek távoli összetevőik)

```
Moose[Hu-En]>a kutya amiatt ugatott, hogy énekeltem.
1: [the dog barked because I sang.]
```

- kötőszóval álló határozói mellékmondat, akár közbeékelve is

```
Moose[Hu-En]>a kutyám, miközben szundikáltam, elolvasta
a könyvet.
1: [my dog read the book while I was dozing.]
```

1.1.2 Vonatkozó mellékmondatok

A vonatkozó mellékmondatok fordítása számos érdekes problémát vetett fel. Ezek közül a rendszerünk számára a legnagyobb technikai kihívást az igei vonzatkeretek vonzataira tett megkötések távoli érvényesítése volt. A vonatkozó mellékmondatban valamely összetevő helyét egy vonatkozó névmás foglalja el. Amennyiben a vonzatkeretnek az adott összetevőre vannak szemantikai, vagy lexikális megkötései, azok nyilvánvalóan nem érvényesíthetők közvetlenül a vonatkozó névmáson, viszont a lexikális fejet tartalmazó antecedensnek meg kell felelnie a VP követelményeinek. Ez azt jelenti, hogy az egyes vonzatkereteket leíró szabályok nem köthetik ki maguknak a vonzatoktulajdonságait, hanem az egész VP-t leíró kategóriában minden összetevő reprezentációja mellett a rá vonatkozó megkötéseket is jegycsoportokban kell tárolni, hogy a VP fölött működő „technikai" szabályok a konkrét megvalósulástól függően ellenőrizhessék az egyes vonzatok tulajdonságait, vagy – vonatkozó névmás esetén – továbbadják a megkötéseket. Ezeket a megkötéseket a vonatkozó mellékmondat legfelsőbb szintű ábrázolásáig örököltetni kell, hogy azokban a mondatszerkezeti szabályokban, amelyek egy távoli antecedenssel társítják a mellékmondatot, végül ellenőrizni lehessen őket.

Szintén fontos megoldandó probléma volt az, hogy a demonstratív determinánst, illetve az 'olyan' módosítót tartalmazó NP-k fordítása függ attól, hogy kapcsolódik-e hozzájuk – adott esetben távolról – vonatkozó mellékmondat. Más, a vonatkozó névmásokkal párba állítható névmási elemekre is igaz, hogy önálló fordításuk eltér attól, amit antecedensként való megjelenésük igényel. Az ilyen elemeket generáló szabályoknak több generálósora van, amelyek közül egy olyan string típusú jegy alapján választunk, amely úgymond „üzenetet hoz" a távoli összetevőtől. Ez a jegy akkor töltődik ki, amikor a vonatkozó mellékmondatot társítjuk a főmondat megfelelő antecedenst tartalmazó összetevőjével.

Néhány példa a vonatkozó mellékmondatok fordítására:

- nem kimozgatott vonatkozó mellékmondatot tartalmazó NP

```
Moose[Hu-En]>az a kutya,amelyik a ház előtt ugat, teg-
nap aludt.
1: [the dog that barks in front of the house slept
yesterday.]
```

- Összehasonlításképpen az „az a kutya" NP fordítása vonatkozó mellékmondatot nem tartalmazó mondatban

```
Moose[Hu-En]>az a kutya tegnap aludt.
1: [that dog slept yesterday.]
```

- pronominális antecedens különböző fordításai a vonatkozó névmástól függően

```
Moose[Hu-En]>azok, akikkel találkoztunk, nem szeretik a
kutyámat.
1: [those who we met do not like my dog.]

Moose[Hu-En]>azok, amelyek pirosak, nem szépek.
1: [the ones that are red are not beautiful.]
```

1.1.3 Névszói állítmányt tartalmazó mondatok

A névszói állítmányok kezelésénél két nagy problémát kellett megoldanunk. A legfőbb nehézséget az jelentette, hogy a jelen idejű, harmadik személyű, kopulaként használt létige a felszínen nem jelenik meg. Nyelvtani modellünkben a mondat elemzését az ige köré felépített vonzatkeretre alapozzuk. Hangzó ige hiányában a szokásos VP-építő mechanizmus nem működik. Annak érdekében, hogy mégis a már meglévő szabályhalmazt alkalmazhassuk az ilyen mondatokra is, kénytelenek voltunk a névszói állítmány hangzó névszói részét a VP kiindulópontjaként felhasználni. Néhány technikai szabály segítségével aztán úgy alakítjuk át a reprezentációt, hogy a névszói csoport már vonzatként jelenjen meg, és az igét leíró jegyhalmazt úgy töltjük ki, hogy az egy megfelelő létigét kódoljon. Természetesen két azonos esetben álló névszói csoportról nem lehet egymástól függetlenül eldönteni, hogy melyikük az állítmány, és melyikük az alany. Ez a fordítás szórendje szempontjából is érdekes kérdés, amire még visszatérünk, de azért is foglalkoznunk kell vele, mert ha bármelyik névszói csoportot tekinthetjük a VP kiindulópontjának, akkor végeredményben két egyenértékű elemzést fogunk kapni feleslegesen. A megoldás az volt, hogy a mondatösszetevőket összegyűjtő szabályokat olyan feltételekkel egészítettük ki, amelyek biztosítják, hogy a névszói állítmányos szerkezet magját alkotó összetevőnek a másik vonzattal való sorrendje rögzített legyen.

Mivel a vonzatkeret-leírások az egyes vonzatokat esetük, illetve névutójuk alapján azonosítják, a névszói állítmányos mondatok két azonos esetben álló igevonzatát a rendszerben semmi nem különböztette meg egymástól. Így a vonzatok tetszőleges permutációja megfelelt a VP-s minta megkötéseinek, és két fordítás keletkezett, melyek közül az egyikben az alanyi, illetve állítmányi szerep nem helyesen lett kiosztva. A megoldás Vancsa László kollégánk ötlete nyomán az ún. determináltsági fok bevezetése lett. Minden névszói vonzat felépítése közben a megfelelő ponton kitöltünk

egy jegyet, amely egy 10 fokú skálán vehet fel értékeket. Ez az érték más és más a tulajdonnevek, a különféle névmások és a különböző determinánsokat tartalmazó NP-k esetén. Az alanyi és állítmányi szerep kiosztása a determináltsági fok alapján történik. Az értékek úgy lettek meghatározva, hogy a magasabb determináltsági fokú vonzat kerül alanyi pozícióba. A vonzatok permutálása után beiktattunk egy szűrőként működő szabályhalmazt, amely csak azt a változatot engedi tovább, ahol a vonzatok determináltsági fok szerinti sorrendje megfelelő.

Itt említjük meg, hogy a létezést kifejező és birtoklásmondatok fordítását is megoldottuk.

- ugyanaz az NP más pozícióba kerülhet a másik vonzat determináltsági fokától függően

```
Moose[Hu-En]>piros a kutyám.
1: [my dog is red.]

Moose[Hu-En]>az a kutyám.
1: [that is my dog.]
```

- a segédige befolyásolhatja a névszói állítmányos keret vonzatainak esetét

```
Moose[Hu-En]>annak a kutyának pirosnak kellene lennie.
1: [that dog should be red.]
```

- létezést kifejező és birtoklásmondatok

```
Moose[Hu-En]>van egy kutya a ház előtt.
1: [there is a dog in front of the house.]

Moose[Hu-En]>van egy kutyám.
1: [I have a dog.]
```

1.1.4 Melléknévi igeneves szerkezetek

A melléknévi igeneves szerkezeteket az igei vonzatkeretek segítségével fordítjuk. Az igenév vonzatait és módosítóit a VPP-t felépítő szabályok segítségével elemezzük, majd szükség esetén a hiányzó vagy másképp megjelenő vonzatokat a reprezentáció manipulálásával „átmenetileg" pótoljuk vagy átalakítjuk, azért hogy az egységes VP-leírás alkalmazható legyen. A VP-s minta fölötti szabályok aztán elvégzik a szükséges visszaalakításokat, illetve utasítást adnak a fordítás transzformálására.

Ezt valamivel részletesebben bemutatjuk a következő tranzitív vonzatkeret példáján.

```
(1) [A-NOM] megver [B-ACC]
(2) az {([A-által]) (tegnap) megvert} ... [B]
```

(1) az igei vonzatkeret alapalakja. A (2)-ben megjelenő befejezett melléknévi igeneves szerkezet, melynek határait kapcsos zárójellel jelöltük többféleképpen is eltér (1)-től. Egyrészt az eredeti alany vagy nem jelenik meg, vagy *által* névutós vonzataként realizálódik; másrészt az eredeti tárgy kívül esik ennek a szerkezetnek a határain. Ahhoz hogy az (1)-ben ábrázolt megkötéseknek eleget tegyen az igeneves szerkezet reprezentációja, a tárgy helyét ki kell tölteni ahhoz hasonló módon, mint ahogy a zéró

névmási tárgyat kezeljük. Az alannyal hasonló módon járunk el, ha hiányzik. Ha *által* névutós vonzatként jelenik meg, akkor az alany esetére és névutójára vonatkozó megkötést kódoló jegy értékét írjuk át. A VP azonosítása után a hiányzó tárgyra tett szemantikai vagy lexikális megkötéseket továbbadjuk, hogy azok majd érvényesítve legyenek az igeneves kifejezést tartalmazó NP összeállításakor, valamint az *által* névutós vonzat jelenlététől függően egyszerű vagy *by* prepozíciós passzívra transzformáltatjuk az eredeti vonzatkeret fordítását:

```
Moose[Hu-En]>az egér megverte a kutyát.
1: [the mouse beat the dog.]

Moose[Hu-En]>az egér által megvert kutya
1: [the dog, which was beaten by the mouse]

Moose[Hu-En]>a megvert kutya
1: [the dog, which was beaten]
```

- az igeneves kifejezés fordításának akár a pozíciója is függhet egy módosító jelenlététől

```
Moose[Hu-En]>az ugató kutya
1: [the barking dog]

Moose[Hu-En]>a ház előtt ugató kutya
1: [the dog barking in front of the house]
```

2 Nyelvi adatbázisok

2.1 Áttekintés

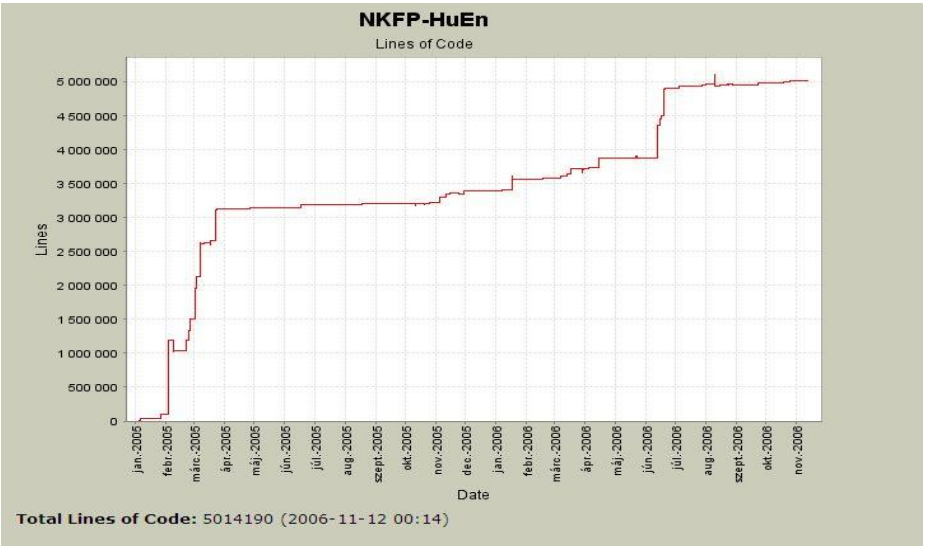
A fordítóprogram fejlesztése 2000-ben kezdődött, a munkálatokba a fejlesztés hat éve alatt csaknem ötven ember kapcsolódott be. A fordítóprogram magyar-angol nyelvi moduljának fejlesztése az NKFP támogatásával 2005 januárjában indult, és három helyszínen zajlik: a MorphoLogicban, a Nyelvtudományi Intézet Korpusznyelvészeti Osztályán és a Szegedi Tudományegyetem Informatikai Tanszékcsoportjában. A három intézményben jelenleg összesen kb. 15 fejlesztő dolgozik. A nagyméretű, több helyszínen is folyó projekt kezelésére CVS változáskövető rendszert használunk. A CVS rendszer adminisztrálja, hogy a fejlesztők mikor mit adtak hozzá, vagy hogyan módosították az adatbázisokat. Ebben a részben az ezekből készített kimutatásokat publikáljuk és értelmezzük, betekintést engedve a projekttel kapcsolatos anyagok fejlődésébe.

2.2 A fejlesztés

A magyar-angol nyelvi adatok a projekt kezdete óta, 2005 januárjától CVS-felügyelet alatt állnak, így fejlődésük jól nyomon követhető.

A CVS adatbázisok kiértékeléséhez többféle program is hozzáférhető, melyek jól áttekinthető grafikonokat is készítenek a tevékenységről. Az alábbi grafikon jellemző képet fest a lexikális erőforrások fejlődésének időbeli alakulásáról.

1. táblázat: A MetaMorpho magyar-angol CVS-forrássorainak növekedése (2006. november 12.)



Az ábrán látható, hogy a projekt kezdetén először összegyűjtöttük a felhasználható forrásokat, majd programokkal a megfelelő *mmd* formátumra konvertáltuk, és 2005 márciusában megosztottuk őket. Ezután egészen 2006 júliusáig tartott az anyagok fordítása és pontosítása, A kismértékű növekedés a magnyelvtan szabályok létrehozásának köszönhető. Idén júliusban a szótár alapú adatbázisunkat korpuszgyakorisági vizsgálatokból származó további mintákkal egészítettük ki.

A kiértékelő rendszer segítségével a szerzőkről is megtudható, hogy hány forrássorral járultak hozzá a létrejött adatokhoz és mennyit módosítottak ezeken. A kimutatásokból megismerhető továbbá a módosítások eloszlása a nap óráira és a hét napjaira vonatkozóan. Ugyancsak hasznos, hogy rangsorolja a különböző nyelvi modulokat méret, illetve a velük kapcsolatos munka aktivitása alapján.

2.3 Az eredmények

A magyar-angol projekt fejlettségi állapota jól lemérhető az elkészült szabályok számából.

2. táblázat. A MetaMorpho magyar-angol mintáinak száma. (2006. október 31.)

Szabálytípus	angol-magyar	magyar-angol
CORE	4120	4624
VP	22593	23884
NX	75135	73795
NX egyszavas	37492	60235
NX többszavas	37643	13560
ADJX	13255	12449
ADVX	2062	3270
ADVP	4940	0

Amint az a táblázatból látszik, a minták elérték az angol-magyar fordítóban lévő szabályok számát, de még alapvető határozói szerkezetek hiányoznak.

3 Szoftverfejlesztés, termékek és szolgáltatások

Az idei évben elsősorban a magyar-angol nyelvészeti munkálatokra összpontosítottunk, de a programok különböző változatai is új lehetőségekkel bővültek.

3.1 MorphoWord

Az év elején átalakítottuk a szerverprogramot úgy, hogy az addigi egy nyelvpár helyett tetszőleges számú nyelvi adatbázis egyidejű kiszolgálására is alkalmas legyen.

Javítottuk a sajátszótár-építési lehetőségeket a MorphoWord Pro alkalmazásban. A MetaMorpho fordítónak ebben a Microsoft Wordbe integrálódó változatában két lényeges fejlesztés is történt. Egyrészt megoldottuk, hogy a felhasználói szótár egy lépésben, akár egész szószedettel is bővíthető legyen. Az export/import funkcióval egyszerű Excel-táblázatból, vagy a fordítómemóriáknál használatos TMX formátumú forrásból lehet saját szószedeteket beolvasni, illetve kimenteni.

A hagyományos fordítási folyamat első lépése a fordítandó szöveg terminológiájának meghatározása. A fordítóprogram esetén is ugyanilyen hasznos lehet az az előfeldolgozási lépés, amely során az aktuális szöveg gyakori szavainak helyes jelentését meghatározzuk. A MorphoWordbe épített új terminológiakivonatoló modulunk nemcsak összegyűjti a szövegben lévő gyakori szavakat és kifejezéseket, hanem le is fordítja őket. Így a fordítás előtt előre meggyőződhetünk ezek helyességéről, és az aktuális jelentés megadásával javíthatjuk a fordítás minőségét. Az így beolvasott adatok a rendszer saját tudásába egyenértékű módon integrálódnak.

Elkészült a fordítóprogram internes változata a MorphoWord Net. Ennél a megoldásnál a felhasználó MS Wordjébe integrálódó kliens a fordítást a MorphoLogic szerverétől kapja. Az elszámolás a fordítandó szöveg hosszával arányosan történik.

3. 2 Webforditas.hu

A tavalyi év decemberében elindítottuk ingyenes angol-magyar webes mondatfordítási szolgáltatásunkat. Az ingyenes webes fordítófelületet több előnnyel is járt: miközben hasznos visszajelzéseket kaptunk, kapcsolatba kerültünk az igazi felhasználókkal és növeltük a program ismertségét.

Idén szeptemberben összevontuk a tíz éve www.mobidictionary.hu néven üzemelő szótárszolgáltatásunkat a MetaMorpho szövegfordítóval és kiegészítettük a weblapfordítás lehetőségével. A három szolgáltatás együtt a www.webforditas.hu cím alatt érhető el. A szolgáltatáscsomag ingyenesen, közreműködésünk nélkül is beépíthető más weboldalakba. Ezzel a lehetőséggel több webmester és weblapépítő élt, és ezzel tovább növelte a fordítóoldalunk látogatottságát.

Feldolgoztuk a www.webfordítás.hu első hónapjának látogatottsági adatait. 2006 októberében összesen 100 ezer látogatónk volt az oldalon. A látogatók 55%-a használt szótárat, 45% fordított szöveget és 8%-a vette igénybe az új weboldal fordító szolgáltatást. A látogatók 55%-a más oldalakon (index.hu, nol.hu, szotar.lap.hu, stb.) befűzött szolgáltatásokon keresztül, a többiek közvetlenül a webforditas.hu oldal megcímzésével jutottak el hozzánk. A látogatottság az első heti 22 ezerről a negyedik hétre heti 28 ezerre növekedett. Munkanapokon a látogatottság a hétvégi adatoknak közel kétszerese, és hét közben jórészt a munkaidőre esik. A látogatók átlagosan 260 ezer látogatást (félórás megkezdett tartózkodást) produkáltak és eközben átlagosan 10 percet töltöttek el. A látogatások során átlagosan 7,4 fordítást végeztek.

A látogatók ez alatt az egy hónap alatt 780 ezer mondatot vagy szövegrészletet fordítottak, 650 ezerszer kérdezték le a szótárakat és 73 ezer weboldalt fordítottak le. A tevékenyég kiszolgálásához a szervereinknek átlagosan napi 3,5 GByte adatforgalmat kell bonyolítania. A szolgáltatást a T-Online Adatparkban elhelyezett szervergépeink végzik, melyek a fizetős szolgáltatások előnyben részesítése mellett terhelésmegosztó üzemmódban működnek.

A látogatók földrajzi megoszlásának vizsgálatakor kiderült, hogy a látogatók 11%-a külföldi, ezen belül a listavezetők: Németország 18%, Egyesült Államok 15%, Egyesült Királyság 12%, Románia 12%, Ausztria 7,5%, Szlovákia 6,5%. Ez azt jelenti, hogy például az októberi százezer Magyarországi látogatóra ezer romániai és ötszáz szlovákiai látogató esett.

Érdekes a látogatók magyarországi megoszlása is: Budapest és Pest megye adta a látogatók 58%-át, ezután Szeged következik 11%-kal, majd ettől leszakadva Győr 2,5% Pécs 2,5%.

A kérdéseket tartalmilag is értékeltük. Itt most csak a MetaMorpho projekt szempontjából érdekes szöveg és weblapfordítás adataira térünk ki.

Gyakorisági sorrendbe állítottuk, és tematikusan kategorizáltuk a weblapfordítóval lefordított oldalakat. Ezzel fontos információhoz jutottunk a jövőbeni fejlesztésekhez. Kellemesen csalódtunk, mert a listát a szándékunknak megfelelően egy idegen nyelvű hírportál (www.cnn.com) vezette, de a www.bbc.com is ott volt az élmezőnyben. A második helyre az angol nyelvű Wikipedia-oldalak kerültek. Ennek is csak örülhetünk, hiszen tapasztalataink szerint a fordítás gyenge minősége nagyon gyakran a hibás forrásszövegnek köszönhető, a Wikipedia lexikonba azonban jellemzően jól átgondolt szabatos, érthető megfogalmazások kerülnek, ezért a program ebben a környezetben feltűnően jó minőségű fordítást ad. Harmadik helyet az internetes já-

tékoldalak foglalják el. A .hu doménű oldalak 5%-os aránya először ijesztőnek tűnt, de mint kiderült, ezek is alapvetően angol nyelvű oldalak voltak. Az erotikus tartalmú oldalak részaránya 1% alatt maradt.

A fordítási igények közvetlen kiértékelésénél is nagy segítséget jelentenek a felhasználói visszajelzések. A felhasználók a problémákat a webforditas.hu oldalról egyetlen gombnyomással, e-mail címük megadása nélkül is elküldhetik. A szükséges információt – a gépi fordítást, a fordító alkalmazást, és a környezetet – a visszajelzési oldal automatikusan előállítja. Sokan csak a fordítást küldik el, de vannak akik minősítést adnak, vagy javaslatokat is írnak.

A fordításokkal kapcsolatos visszajelzéseket külön értékeljük aszerint, hogy a webforditas.hu szöveg- vagy weblapfordítója, illetve a MoBiCAT vagy MorphoWord Net fizetős fordítószolgáltatás használata közben születtek. Ezen a helyen csak az idén bevezetett szolgáltatások visszajelzéseit értékeljük.

3.2.1 A weblapfordító visszajelzéseinek értékelése

Az első hónap pozitív eredményei egyértelműen igazolják az erőfeszítéseket. A negatív visszajelzések elsősorban technikai jellegűek voltak. Eleinte több népszerű oldalt technikai okokból nem tudott lefordítani a webforditas.hu, amikkel egyenként kellett foglalkozni. Sok felhasználó nem érti egyébként azt se, hogy a kép formájában megjelenő szövegek a program számára miért elérhetetlenek.

3.2.2 A szövegfordító értékelése

Tizenegy hónapnyi szolgáltatás adataival rendelkezünk, 2615 visszajelzést kaptunk 617 levélírótól. A problémalistán a különböző hibás fordítások jelzése mellett sokan vannak a magyar-angol fordító hiányát jelző levelek. A szövegfordító különösen sok helytelen forrásanyagot kap, valószínűleg azért, mert a programot elsősorban az angolul nem tudóknak ajánljuk, és ők gyakran saját maguk próbálnak tesztmondatokat írni. A programba számos hibatoleráló elemet építettünk be: megengedtük, hogy a tulajdonneveket kis kezdőbetűvel is lehessen írni, a rövidített angol létigéknél pedig nem kötelező az aposztróf használata. A mondatvégi pont után ki nem tett szóköz kezelését is meg kellett oldanunk, mert az összeolvadó szavak nem fordultak le, és annak ellenére folyamatosak voltak a panaszok, hogy erre a helytelenségre a beíró ablak alatt felhívtuk a figyelmet. Javult a fordítás megítélése akkor is, amikor a sortörések értelmezését megváltoztattuk. A programot egy dalszövegeket gyűjtő internetes oldalról is elérhetővé tették, ez is jelentős forgalmat generál. A változás hatására a verssorok egyenként fordulnak, és így általában érthetőbb eredmény születik.

4 Tervek 2007-re

A magyar-angol fordítóprogram adatbázis nyelvészeti munkálatai a végéhez közelednek. Az utolsó pályázati munkaszakasz feladata (2007. január 1.–május 1.) a magyar-angol fordító tesztelése. Minden jel arra mutat, hogy a teszteket az év elején valóban el fogjuk tudni kezdeni, és a magyar-angol fordító már a májusi határidő előtt nyilvánosan is elérhető lesz valamennyi MetaMorpho-alapú fordítóprogramban.

A magyar-angol fordítók publikálása után a hangsúlyt a fordítási minőség javítására fogjuk helyezni, új nyelvpárok fejlesztésébe csak a megfelelő nyelvi minőség elérése után kezdünk.

Bibliográfia

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. Merényi Csaba. A MetaMorpho magyar−angol gépi fordító rendszer igei vonzatkereteit működtető nyelvtan. In: Alexin Zoltán; Csendes Dóra (szerk.) *A 3. Magyar Számítógépes Nyelvészeti Konferencia előadásai*, 108−115. SZTE, Szeged (2005)

Szótárazási dilemmák a MetaMorpho magyar-angol fordítóprogram névszói adatbázisának építésében

Vincze Veronika[1], Lucza Mónika[1], Csendes Dóra[1], Kiss Gabriella[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Nyelvtechnológiai Csoport
H-6720 Szeged, Árpád tér 2.
{vinczev, lucza, dcsendes}@inf.u-szeged.hu
[2] MorphoLogic Kft. Budapest
H-1126 Budapest, Orbánhegyi út 5.
gkiss@morphologic.hu

Kivonat: Jelen cikk a MetaMorpho magyar-angol fordítóprogram kétnyelvű szótárának előállítása során tapasztalt gyakorlati nehézségekről és azok nyelvészeti hátteréről számol be. A szótár fejlesztésében a Szegedi Tudományegyetem a magyar névszói kifejezések – azaz főnévi kifejezések (NX), melléknévi (ADJX) és határozószói szótári bejegyzések (ADVX) – angolra fordításával vette ki részét. Az így létrejövő szótári adatbázis jelenleg közel 90 ezer bejegyzést tartalmaz. A névszói kifejezések fordítása során a célkitűzés az volt, hogy a gyakorisági elemzések alapján a magyar nyelvhasználatnak leginkább megfelelő jelentés kerüljön első helyre az adatbázisban, ugyanakkor igyekeztünk a lehető legtöbb jelentést felvenni.

1 Bevezetés

A gépi fordítórendszerek teljesítményét nagyban befolyásolja a rendszerben lévő szótári adatbázis minősége. A szótárak jelentőségét növeli még az a tény is, hogy a felhasználók a szótár bővítésével interaktív módon javítani tudják a rendszert. A létező fordítórendszerek szótárai a lexikai elemeket igen eltérő formátumban, tartalmi lefedettséggel, részletességgel, és eltérő precizitású formalizmussal tartalmazzák [2]. A szótári adatbázisok megvalósítása mindig az adott rendszer sajátosságainak függvénye. Az interlingva rendszerek szótárainak például természetükből fakadóan nem szükséges fordítási információkat tartalmazniuk, míg a sok nyelvre készülő fordítóprogramok esetében gyakori, hogy az alkalmazott nyelvekre részletes egynyelvű szótárakkal is rendelkeznek a kétnyelvű, a transzfer során használt szótárak mellett [1].

A MetaMorpho fordítóprogram-család a különféle fordítási feladatokhoz kíván eszközöket biztosítani [5,6,7]. A rendszer elemzési szabályokon alapszik, amelyekhez fordítások vannak közvetlenül hozzárendelve. Ezek a szabályok kódolják a különböző nyelvtani szabályszerűségeket, lexikális elemeket, vonzatkereteket, szemantikai jegyeket és egyéb mintákat.

A jelen tanulmány a MetaMorpho magyar-angol fordítóprogram kétnyelvű szótári adatbázisának előállítása során tapasztalt gyakorlati nehézségekről és azok nyelvészeti hátteréről számol be. A szótári adatbázis kialakítása két fő fázisban történt: az automatikus előfeldolgozási, illetve előfordítási fázist egy manuális ellenőrzési és javítási fázis követte. Az előfeldolgozás során sor került a kifejezések gépi felhasználásra alkalmas egységes szabály formátumra hozására, valamint potenciális fordítási javaslatok automatikus módon történő előállítására. Ezt követően a kézi ellenőrzés során arra törekedtünk, hogy minél szélesebb körű információval lássunk el egy-egy bejegyzést, így azok nem csupán a kifejezések fordítását tartalmazzák, hanem morfoszintaktikai, illetve szemantikai információkat is. Igyekeztünk továbbá figyelmet fordítani arra is, hogy jellemzően csak az alapszókincsbe tartozó szavak kerüljenek be az adatbázisba, így az elavult, tájnyelvi vagy szakzsargonba tartozó szavakat kiszűrtük.

2 Automatikus előfeldolgozás

A szótári adatbázis szócikkei három különböző forrásból származnak: (1) a MetaMorpho angol-magyar fordítóprogram anyagának kifordításával és feldolgozásával keletkezett szótári egységek; (2) a MorphoLogic által korpusz-gyakorisági vizsgálat alapján létrehozott szótári egységek; (3) a Nyelvtudományi Intézet és a Szegedi Tudományegyetem által létrehozott és jegyekkel ellátott szótári egységek. Az előfeldolgozás célja a fordítandó nyelvi forrás összegyűjtése, mmd-formátumra[68] [3] hozása, ill. emberi fordítási folyamat lerövidítése volt.

Az (1) csoport szótári egységeinek létrehozása volt a legnagyobb körültekintéssel végzendő feladat, mivel itt az angol-magyar szótári adatbázis visszafordításból származó azonos magyar fordításokat kellett egységesíteni, a nekik megfelelő angol fordításokat pedig alternatív megoldásként felsorolni. Ezekben az esetekben az angol-magyar fordítóprogramban szereplő szemantikai jegyeknek a kifordított szabályba való automatikus átemelése sokszor nem volt egyértelmű. Bizonyos szemantikai jegyek (pl. *földrajzi nevek, tulajdonneveket leíró szemantikai jegyek*) átvétele viszont kevésbé volt problémás.

A (2) csoport szótári egységeinek a feldolgozását a Magyar Nemzeti Szövegtár[69] alapján végeztük el. Egy Perl-script segítségével határoztuk meg azt, hogy az alapszótár alapján javasolt fordítás vagy fordítások milyen arányban fordulnak elő a Hunglish[70] kétnyelvű korpusz angol részeiben. Ennek alapján a lehetséges fordításokra egy sorrendet állítottunk fel és a legvalószínűbb fordítással kitöltöttük a szabály angol fordítását.

A (3) csoport szótári egységei egynyelvűek voltak, a szótári kifejezések el voltak látva szemantikai jegyekkel is. Itt a kitöltetlen generáló oldal (angol oldal) automatikus feltöltése az (2) csoportban ismertetett automatikus módszerrel zajlott.

Az automatikusan előállított szabályokat átnéző nyelvészek feladata az volt, hogy a fordítás helyességét ellenőrizzék és megjegyzésben jelöljék észrevételeiket a sza-

[68] Az mmd-formátum (lásd 3. fejezet példái) a MetaMorpho keretrendszer nyelvi adatainak és szabályainak formátuma.

[69] http://corpus.nytud.hu/mnsz/

[70] http://szotar.mokk.bme.hu/hunglish/search/corpus

bályra vonatkozóan. A szabályokhoz tartozó megjegyzéseket osztályokba soroltuk, ezekhez egységes jelölésrendszert alkottunk, amely alapján szükség szerint pl. adott szabályok egységesen kivonhatók lettek a rendszerből. Az (1), (2) és a (3)-mal jelzett kifejezésekhez tartozó szabályok százalékos aránya sorrendben a jelenlegi rendszerben a következőképpen alakult *45,3:34,4: 20,3:%*. Jelenleg összesen 73.795 főnévi, 12.449 melléknévi, valamint 3270 határozói kifejezést tartalmaz a szótári adatbázis. A fordítórendszer hatékonyságának növelése érdekében ezen felül közel 4500 kollokáció is szerepel az adatbázisban.

A magyar-angol források bővítése természetesen nem zárult le, később a szótárt a felhasználók visszajelzései alapján tovább bővítjük. A MorphoLogicban fejlesztés alatt van egy terminológia-kivonatoló, amivel a fordítóprogram felhasználója interaktív módon módosíthatja, ill. bővítheti a szótári adatbázist.

3 A fordításszabályok felépítése

Az automatikus előkészítés eredményeképpen a következő, ún. mmd-formátumú szabályok álltak elő. Egy mmd-szabály kötelezően áll egy fejlécből, egy elemző sorból és egy vagy több generáló sorból. Az mmd-szabály tartalmazhat megjegyzés sorokat is. Az mmd-szabály elemző ill. generáló sora funkcionálisan két részre bontható: feltétel- és értékadást leíró részre. A feltétel mindig kerek zárójelek között van megfogalmazva, a feltételt leíró elemek formailag jegy és érték párok, ahol az értékek sztring és szimbólum típusúak lehetnek. A szabály értékadás része kapcsos zárójelek között található. Az értékadás formai leírása is jegy-érték párokon alapul, annak a használata a feltételek elemeivel megegyező.

Példa:
```
*ADJX=alkoholmentes:304
HU.ADJX = ADJ(lex="alkoholmentes")
EN.ADJX = N[lex="alcohol"] + PUNCT[lex="gluehyphen"] + ADJ[lex="free"]
;cmt: adj.mmd
;lexs: |non gluehyphen alcoholic,alcohol gluehyphen free
;tr_A: non-alcoholic
```

A szabály első sorából kiolvasható az adott kifejezés nyelvtani kategóriája (jelen példában ez ADJX-szel jelölt melléknév). A második sor a magyar (elemzési oldal), a harmadik sor az angol (generálási oldal) kifejezést adja meg részeire bontva. A magyar oldalon levő melléknévnek az angol oldalon egy főnév és egy melléknév kötőjellel összekapcsolt kombinációja felel meg. A szabály tartalmazza még az előfeldolgozás során előállított fordítási javaslatokat is (jelölésük: ;tr_A). A javasolt fordítások azonban nem mindig helytállóak, így ezekben az esetekben az ellenőrzést végzőknek kellett keresni megfelelő fordítást az adott kifejezésre. A szabályokban továbbá szerepelhetnek még morfológiai, szemantikai, szórendre, illetve szóhasználatra vonatkozó információk is. Jelölni lehet például, ha a magyar és angol kifejezés száma eltér, illetve a főnevek esetét is feltüntetjük, amennyiben nem alanyesetben állnak (morfológiai információ):

HU.NX[ennum=PL] = N(lex="tutyi")
EN.NX = N#1[lex="carpet"] + N[lex="slipper"]

HU.NX = N(lex="tehéntej")
EN.NX = N#1[lex="cow", case=GEN] + N[lex="milk"]

A melléknévi kifejezések fordításakor megjelöljük, ha a melléknév angol megfelelője – az általános szabálytól eltérően – a főnév mögött helyezkedik el (szórendi információ):

HU.ADJX[enpos=POST] = ADJ(lex="mustáros")
EN.ADJX(:head="PREP") = PREP[lex="with"] + N[lex="mustard"]

Utóbbi szabály egyben a szintaktikai fej jelölésére is példa. A fejet csak akkor jelöljük, ha nem esik egybe a névszói csoport alapértelmezett fejével

4 A névszói csoportok fordítása során szerzett tapasztalatok

A MetaMorpho fordítóprogram szótári adatbázisának fejlesztésében a Szegedi Tudományegyetem a magyar névszói kifejezések – azaz főnévi és melléknévi kifejezések – és határozószói szótári bejegyzések angolra fordításával vette ki részét. A fejlesztés során nem csupán az egyes bejegyzések angol megfelelőjének megadása volt a feladat, hanem szükség esetén további nyelvészeti, használatbeli jellegzetességekre vonatkozó információt is tárolni lehetett a szabályokban. Fontosnak bizonyult, hogy a gyakorisági vizsgálatok alapján a magyar nyelvhasználatnak leginkább megfelelő jelentés kerüljön első helyre a szótárban, mivel a jelenlegi rendszer egy jelentést kezel. A későbbi fejlesztésekre való tekintettel lehetőség szerint igyekeztünk egy adott bejegyzés minél több jelentését felvenni.

4.1 A főnévi csoportok fordítása során szerzett tapasztalatok

A főnévi kifejezések fordítása kapcsán felmerült problémákat három fő típusba lehet sorolni. Az első problémakörbe a többjelentésű szavak tartoznak, hiszen bizonyos esetekben nehezen lehetett eldönteni, hogy az adott bejegyzésnek melyik jelentése a (leg)gyakoribb (pl.: *ráhajtás*, *partnercsere*, *fazon*). A második csoportba azok a szavak tartoznak, ahol a szótári információk nem teljesen helytállóak, például hibás fordítás szerepel a magyar–angol szótárban (*vattapamacs*, *szülőszék*). A harmadik tipikus probléma a kultúraspecifikus szavakhoz köthető: e szavak fordítása igen nehézkes, hiszen sokszor a másik nyelvterület nem is ismeri az adott dolgot a maga valójában, vagy pedig teljesen más képzeteket társít hozzá (*máglyarakás*, *tanyarendszer*). Az egyes típusok különböző altípusokra oszthatók, amelyekre az alábbiakban mutatunk be egy-egy jellegzetes példát:

I. típus: többjelentésű szavak esete

(a) A magyar szónak két angol megfelelője van

Ez akkor okoz problémát, amikor nem lehet a gyakoriságukat vagy fontosságukat rangsorolni, mivel egyenrangúak, egyforma gyakorisággal fordulnak elő, de eltérő

esetekben használjuk őket. Az emberi fordítóknak nem okoz problémát eldönteni, hogy melyik angol szót használják az adott környezetben. Annak érdekében, hogy a gépi rendszer is hasonló megbízhatósággal tudjon dönteni a MorphoLogic fejlesztői egy szemantikai modul integrálásán dolgoznak.

pl:
magyar: *körte*
angol: *light bulb / pear*

(b) A magyar szónak több jelentése van, de a rendelkezésre álló kétnyelvű szótárban a (leg)ritkábban használatos van megadva

Nagyon sokszor találkoztunk olyan szavakkal, amelyeknek több eltérő jelentésű használata van, és igencsak nagy meglepetést okozott a kétnyelvű szótár által adott jelentés. Ilyen pl. a *ráhajtás*, amelynek az angol szótári megfelelője az „*over*" szó (kötésben ráhajtás). A javítást végző csapatban 5-6 egyéb jelentést gyűjtöttünk, melyek között a szótári megoldás nem szerepelt.

(c) Nehéz eldönteni a magyar jelentések esetén a használati sorrendet

Ide azok a jól körülhatárolható, egyértelmű magyar jelentéssel és angol fordítással rendelkező szavak tartoznak, amelyek használatának gyakorisági sorrendje nem állapítható meg egyértelműen, az egyes jelentések szubjektív döntés alapján kerülnek besorolásra. Ide tartoznak pl. a *sitt, szivornya, fazon*, stb.

II. típus: Szótári információk nem helytállóak

(a) Kifordított szótárból fakadó problémák

Az automatikus előfordítás során, részben a MetaMorpho angol-magyar fordítórendszerének szótári adatbázisa került visszafordításra. Emiatt előfordult olyan eset, amikor a bejegyzés magyar fordítása hibás volt, így a visszafordítás során ugyancsak félrevezető eredményt adott. Ilyen például:

angol eredeti: *pharaoh's serpent / pharaoh's rat*
magyar visszafordított: *fáraókígyó / fáraópatkány*
helyes magyar kifejezés: *a fáraó kígyója / egyiptomi mongúz*

(b) A magyar szó jelentése jól körülhatárolható, de hibásan szerepel a szótárban

Ez volt a leggyakrabban előforduló probléma. Ilyenkor hosszabb-rövidebb internetes kutatómunka alapján kerestük meg az alkalmas fordítást. Ilyen volt pl. a *vattapamacs,* amelyhez a szótárban megadott fordítások közül a „*swab*" (fültisztító) jelentése állt legközelebb. Ebben a konkrét esetben pl. angol drogériák internetes katalógusait használtuk. Hasonló példa a *magassági botkormány* is, amikor is egy repülési lexikont böngészve sikerült rábukkanni a megfelelő kifejezésre.

(c) A magyar jelentés jól körülhatárolható, de nem szerepel a szótárban

Általában a speciális szakkifejezések sorolhatók ebbe a kategóriába, mint például a *rázószekrény*. Itt is a jól bevált internetes keresés hozta meg a megfelelő eredményt: kombájnok magyar és angol részletes műszaki leírásait tanulmányozva sikerült megtalálni az angol megfelelőt.

(d) A magyar jelentés nem egyértelműen körülhatárolható és nem szerepel a szótárban, vagy a szótárban szerepel ugyan fordítás, de az egyik magyar jelentéssel sem f ér össze

Ez egy elég nehezen kezelhető csoport, hiszen ha a magyar jelentést sem tudjuk megragadni, akkor a legritkább esetben tudjuk csak megtalálni a helyes angol megfelelőt. Ilyen volt például a *számolócédula*, ami elsőre a mellékszámítások elvégzésére szolgáló kis papírfecninek tűnt. Az internetes keresés viszont az aprónyomtatványok között ismerte fel, amelyből kiderült, hogy a pincérek által használt, reklámgrafikával díszített írótömb is ezen a néven ismert. A szótárban szereplő fordítása, a „*bill slip*" viszont a legjobb esetben is csak a pénztárblokk megfelelője. Ilyen esetekben először a lehetséges magyar jelentéseket gyűjtöttük össze, majd azoknak külön-külön megkerestük a fordítását. Első helyre a leggyakrabban előforduló magyar jelentés megfelelőjét írtuk.

(e). Egyértelmű magyar jelentés, de a kétnyelvű szótárban megadott jelenést nem ismeri az egynyelvű szótár

Az egyik legmeglepőbb példa a *szülőszék* esete, amelyre az EISZ[71] a „*lasanum*" fordítást adja. Ezt azonban nem ismeri egyetlen általunk használt egynyelvű szótár sem. Internetes keresésnél a Google 16 találatot ad rá, amelyek jelentését „*chamber pot*"-ként (bili) határozzák meg. Ez egy különösen sajnálatos eset, hiszen az angol kifejezés is éppoly beszédes, mint magyar megfelelője: „*birthing chair*".

III. típus: Kultúraspecifikus szavak

A kultúraspecifikus szavak esetében sem használható jól a kétnyelvű szótár, legalábbis az első jelentés meghatározásakor. Itt az okozza a problémát, hogy igen nagy eséllyel ezeknek a szavaknak természetüknél fogva nem létezik az angol megfelelője. Ilyenek például a *máglyarakás, sárarany*, vagy éppen az *aranykorona*. Ebben az esetben az ilyenkor szokásos fordítói eljárást alkalmaztuk: a lehető legrövidebb és legpontosabb körülírást adtuk meg hozzájuk.

IV. Egyéb, atipikus problémák

(a) Hangutánzó, hangfestő szavak

Általában ezek a szavak nem, vagy esetleg helytelenül szerepelnek a kétnyelvű szótárakban; pl.: *rotyogás*. Ha a szabály javítója nem ismeri anyanyelvi környezetben szerzett tapasztalatai alapján az ilyen jellegű szavakat, akkor támpont nélkül elég nehezen talál hozzájuk elfogadható megfelelőt. Ilyen esetekben a leírt jelenséghez kapcsolódó körülírásokat kerestük, és azok előfordulási gyakoriságát vizsgáltuk. A „*rotyogás*" például jól körülírható a „*boiling sound*" kifejezéssel, és ennek brit angol használata alkalmasnak és elegendően gyakorinak tűnt.

(b) Alapszókincs megítélése

Sok esetben nehéz eldönteni, hogy mi tartozzon bele az alapszókincsbe. A szavak ismerete, használati gyakorisága az anyanyelvi nyelvhasználóknál is szubjektív, életmód-, lakhely vagy családi háttér-függő. Jól illusztrálja ezt az a példa, hogy az egyik javító számára a *vejsze* szó ismert, sőt általánosan használt volt, míg a *vasgyúró* nem – a többség számára pedig pont fordított volt a helyzet.

(c) Szavak, amelyek fordítása egyik irányban magától értetődő, visszafelé pedig megtévesztő lehet

[71] http://www.eisz.hu/

Akadnak olyan szavak, amelyek "becsapósak" lehetnek a fordító és/vagy a szabályt ellenőrzők számára; ilyen például a magyar *winchester* (angol: *HDD / Winchester*). Míg angolból semmi problémát nem okoz a fordítás, visszafelé nem biztos, hogy annyira egyértelmű.

Értelemszerűen külön zavaró tényező, amikor a fent felsorolt esetek nem tisztán, hanem vegyesen fordulnak elő. Az esetek többségében ez volt a jellemző.

4.2 A melléknévi csoportok fordítása során szerzett tapasztalatok

A melléknévi kifejezések fordítása időigényesebbnek bizonyult, mint a főnévi csoportoké. Ennek oka a szóanyagban kereshető: egyfelől sokkal több az elvont kifejezés (*árválkodó*, *kuláns*, *fellengző*), másfelől igen sok a speciális szakterülethez tartozó szó (*kápolnakoszorús*, *jogdíjköteles*). Ugyanakkor sokkal több szó került kiszűrésre, mivel a nyelvtan képes kezelni a produktív képzéseket (például melléknevek fokozása (*legkisebb*) vagy ható képzős alakok (*bírálható*)), ezért ezeket az alakokat nem vesszük fel a szótárba, hiszen csak feleslegesen növelné a szótár méretét. Kivételt képeznek a rendhagyó alakok, amelyeket természetesen szerepeltettünk a szótárban (pl.: *látható – visible*).

A melléknévi csoportok fordítása során felmerültek bizonyos problémák, amelyek nehézséget jelenthetnek a fordítóprogram számára. Hat csoportba osztottuk a tipikus problémákat. Igyekeztünk mindegyik problémára olyan megoldást találni, ami képes elősegíteni a fordítóprogram fejlesztését, későbbi tökéletesítését.

Az első problémát azok a melléknevek jelentik, amelyek önmagukban sohasem fordulnak elő (*nevű*, *színű*), kötelezően kíséri őket egy másik melléknév vagy számnév: **a nevű fiú* vagy **a színű pulóver* kifejezések agrammatikusak, szemben az *a Feri nevű fiú* vagy *a sárga színű pulóver* kifejezésekkel. Ezek a melléknevek ;HALFLEX megjegyzést kaptak:

```
HU.ADJX = ADJ(lex="színű")
EN.ADJX = ADJ[lex="coloured"]
;HALFLEX
```

A második problémakörbe azok a melléknevek sorolandók, amelyek egyik jelentésükben ;HALFLEX megjegyzést kapnának (l. előző probléma), de más jelentésben szerepelhetnek önállóan is (*éves*): *tizenöt éves háború*, ellenben *az éves jelentés*. Ezeknek a mellékneveknek megadtuk az önállóan használható fordítását, és ;ADJHALFLEX kommenttel láttuk el, így jelölve, hogy más használata is lehetséges az adott szónak:

```
HU.ADJX = ADJ(lex="éves")
EN.ADJX = ADJ[lex="annual"]
;ADJHALFLEX
```

Harmadszor, bizonyos melléknevek angol fordítása függ attól is, hogy a melléknév éppen attributív vagy predikatív pozícióban fordul elő (*kérdéses*). A jelzői funkcióra egy példa: *a kérdéses termék* (*the product in question*), illetve a névszói-igei állítmány névszói részének betöltésére is egy illusztráció: *A játéka még kérdéses.* (*His play is still doubtful.*) Ilyen esetben mindkét jelentést megadtuk, a jelzői jelentést

írtuk fel az angol oldalon, a predikatív jelentést pedig a szabály megjegyzés részében tüntettük fel:

HU.ADJX[enpos=POST] = ADJ(lex="kérdéses")
EN.ADJX(:head=”PREP”) = PREP[lex="in"] + N[lex=”question”]
;PREDIC: doubtful

A negyedik csoportba azok a szavak kerültek, amelyek főnevek és melléknevek is lehetnek, tipikusan ilyenek a népnevek (*angol, holland*), de más típusú szavaknál is előfordult ez a jelenség (*százlábú*, *csuhás*). Melléknévként fordítottuk őket, majd ;ASNOUN kommenttel láttuk el őket, és a főnévi jelentésüket is megadtuk:

HU.ADJX = ADJ(lex="holland")
EN.ADJX = ADJ[lex="Dutch"]
;ASNOUN: Dutchman

Az ötödik esetben a melléknév leggyakoribb használatában olyan szókapcsolatban fordul elő, amelynek angol megfelelője egy szó (*nyolcvanas* (*évek*) vs. *eighties*), ugyanakkor más jelentései is léteznek: *nyolcvanas férfi* (azaz nyolcvan évesnél idősebb, vagy 1980-ban született). Ilyen esetekben a leggyakoribb használatnak megfelelő fordítást adtuk meg, feltüntettük a második és harmadik jelentést is, és ;NUM kommenttel jeleztük, hogy (év)számot tartalmazó kifejezésről van szó:

HU.ADJX = ADJ(lex="nyolcvanas")
EN.ADJX(:head="N") = N[lex="eighty", num=PL]
;sense2: EN.ADJX(:head="PREP") = PREP[lex="in"] + PRON[lex="his"] + N[lex="eighty", num=PL]
;sense3: EN.ADJX = ADJ[lex="born"] + PREP[lex="in"] + NUM[lex="1980”]
;NUM

A hatodik problémakör legfőképpen a leggyakoribb mellékneveket érinti, vagyis hogy a melléknév pontos fordítása igen gyakran az utána következő főnév függvénye (*nagy*, *szép*): *nagy üzlet* (***big*** *business*) vs. *nagy szerep* (***large*** *role*) vs. *nagy siker* (***great*** *success*) vs. *nagy felbontás* (***high*** *resolution*). A példában a melléknévnek legalább négyféle helytálló fordítása is lehet, azonban az adott főnév mellett (többnyire) csak egy adott fordítás a megfelelő. Mivel azonban a szabályban csak a melléknév szerepel, ezekben az esetekben igyekeztünk elsőként a legáltalánosabb jelentést megadni, de a specifikusabb jelentéseket is felsoroltuk:

HU.ADJX = ADJ(lex="nagy")
EN.ADJX = ADJ[lex="big"]
;sense2: EN.ADJX = ADJ[lex="large"]
;sense3: EN.ADJX = ADJ[lex="great"]
;sense4: EN.ADJX = ADJ[lex=”high”]

Vonzatos melléknevek esetében a kötelező vonzatot (ragos vagy névutós főnevet) ;CASE kommenttel adtuk meg. Többjelentésű melléknevek esetében külön gondot jelentett, hogy nem mindegyik jelentésre volt érvényes minden, a szórendre vagy szóhasználatra vonatkozó információ. Ilyenkor ;ADJVAL kommentben megjegyeztük, hogy melyik jelentésre melyik információ vonatkozik.

4.3 A határozószói csoportok fordítása során szerzett tapasztalatok

A határozószói csoportok fordítása bizonyult a legkevésbé problémásnak. Ennek két fő oka volt: egyfelől nagyságrendileg kevesebb határozószó szerepel a szótárban a többi szófajhoz képest, a névszói kifejezések mindössze 4%-a volt ADVX. (Összehasonlításképpen: a névszói csoportok 72%-a főnévi csoport, 24%-a melléknévi csoport volt.) Másfelől pedig a határozószavak – a főnevekkel és melléknevekkel összevetve – igen ritkán többértelműek, így fordításuk is jóval könnyebbnek bizonyult. Egy példa a határozószói szabályokra:

```
HU.ADVX = ADV(lex="baloldalt")
EN.ADVX(:head="PREP") = PREP[lex="on"] + DET[dettype=DEF] +
ADJ[lex="left"] + N[lex="side"]
```

5 Kollokációk

A fordítóprogram minél hatékonyabb működése érdekében a szótári adatbázisba kollokációk is bekerültek. Kollokációnak tekintettünk minden olyan többtagú kifejezést, amelynek tagjai viszonylag gyakran szerepelnek együtt, és formájuk többékevésbé rögzített [4]. Néhány példa: *gyáva nyúl*, *hatos lottó* (NX-k), *gyengén látó*, *kreol bőrű* (ADJX-k), *ízig-vérig* (ADVX) és *eb ura fakó* (idióma). Ezek fordítása a legnehezebb, hiszen a kollokációk nem teljes mértékben kompozicionálisak (vagyis jelentésük nem számítható ki alkotórészeik jelentéséből és azok összekapcsolódási módjából), így a kifejezés részeinek lefordításából előállt szókapcsolat a legtöbb esetben nem tekinthető a kifejezés angol megfelelőjének [8]. Ebből következően az automatikusan generált fordítás igen kevés esetben volt elfogadható, magunknak kellett megtalálni a kollokáció pontos angol megfelelőjét. Nehezítette a munkát az is, hogy igen sok szaknyelvi – különösen jogi és gazdasági – terminus szerepelt a kollokációk között (például *járulékfizetési hajlandóság*, *külkereskedelmi mérleghiány*), amelyek fordítását sokszor még a kétnyelvű szakszótárak sem adták meg.

6 További fejlesztések

Jelenleg a szótárban szereplő főneveket különféle szemantikai jegyekkel látjuk el – többek között *abstract*, *human*, *animate*, *currency*, *bodypart*, *mass* jegyeket használunk –, továbbá feltüntetjük a nyelvtani nemet és a megszámlálhatóságot is. Megfelelő nyelvtani szabályokkal kiegészítve a fordítóprogram így könnyebben tud kezelni bizonyos nyelvtani jelenségeket (például névmási referencia), ezáltal pontosabbá válik a létrejövő fordítás.

A szemantikai jegyek bejelölésének ugyanakkor fontos szerepe van az igei vonzatkeret kitöltésében is: például a *kifizet* ige tárgya csak valamilyen pénznem lehet, azaz kötelezően rendelkezik currency=YES jeggyel:

```
HU.VP = SUBJ + TV(:lex=”kifizet”) + OBJ(pos=N, case=ACC, currency=YES)
Az üzletember kifizetett százezer forintot.
```

Az igei vonzatkeret meghatározása és a főnevek szemantikai jegyeinek megadása a többjelentésű szavak fordítását is megkönnyíti. Például az *aláír* ige mellett a *perjel* szó csak 'egyházi elöljáró' (angolul *prior*) jelentésben fordulhat elő, mivel a *perjel* kétféle jelentése közül csak a *prior* rendelkezik a human=YES értékkel, a *slash* természetesen human=NO értékű.

HU.VP = SUBJ(human=YES) + TV(:lex="alá|ír") + OBJ(pos=N, case=ACC)
*A perjel aláírta az iratokat. – The prior / *slash signalled the documents.*

7 Összegzés

A cikkben összefoglaltuk a MetaMorpho magyar-angol fordítóprogram kétnyelvű szótári adatbázisának előállítása során szerzett tapasztalatokat. Röviden vázoltuk azokat a tipikus problémákat, amelyek a szavak többértelműségéből, illetve a szótárak pontatlanságából adódtak. Említést tettünk a kollokációk fordításával kapcsolatos nehézségekről is. A jelenleg zajló fejlesztések – a szemantikai jegyek bejelölése – a fordítóprogram további tökéletesítését teszik lehetővé.

Bibliográfia

1. Arnold, D. J., Balkan, L., Meijer, S., Humphreys, R. L., Sadler, L.: Machine Translation: An Introductory Guide. Oxford: NCC Blackwell (1994)
2. Isabelle, P.: Electronic Dictionaries and Machine Translation Systems. In: Proceedings of the International Symposium on Electronic Dictionaries (ISED-88). Tokyo (1988)
3. Prószéky, Gábor; László Tihanyi: MetaMorpho: A Pattern-Based Machine Translation System. In: Proceedings of the 24th 'Translating and the Computer' Conference, 19–24. ASLIB, London, United Kingdom (2002.)
4. Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.): Proceedings of CICLING-2002. Mexico City (2002)
5. Tihanyi, L.: A MetaMorpho projekt története. In: Alexin, Z., Csendes, D. (eds.): MSzNy 2003 – I. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem (2003) 247–252
6. Tihanyi, L.: A MetaMorpho projekt 2004-ben. In: Alexin, Z., Csendes, D. (eds.): MSzNy 2004 – II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem (2004) 85–87
7. Tihanyi, L.: A MetaMorpho fordítóprogram projekt 2005-ben. In: Alexin, Z., Csendes, D. (eds.): MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem (2005) 99–107
8. Váradi, T.: Többszavas kifejezések kezelése MT szótárban. In: Alexin, Z., Csendes, D. (eds.): MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem (2005) 233–244

A MorphoTM főnévicsoport-szinkronizáló módszereinek továbbfejlesztése és vizsgálata

Pohl Gábor

Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar
1083 Budapest, Práter utca 50/A
pohl@itk.ppke.hu

Kivonat: A MorphoTM (korábbi nevén MetaMorpho TM) olyan fordítómemória, amely nemcsak teljes mondatpárokat, hanem főnévi csoportokat (NP), illetve a mondatban ezeket szimbolikus NP-helyekre cserélve kapott mondatvázakat tárol fordításaikkal együtt adatbázisában. A főnévi csoportokat automatikusan határozzuk meg és szinkronizáljuk fordításaikkal. Cikkünk első felében összehasonlítjuk a magyar főnévi csoportok meghatározására kínálkozó módszereket, a MetaMorpho magyar szintaktikai elemzőt, illetve a főnévi csoportokat fordításaik alapján szótári megfeleltetésekkel és sekély nyelvtannal meghatározó, korábban kidolgozott módszerünket. Cikkünk második felében bemutatjuk, hogy hogyan váltottuk le a főnévi csoportok hasonlóságának meghatározására eddig használt heurisztikus képletet gépi tanulással meghatározott osztályozóra.

1 Bevezetés

A MorphoLogicnál fejlesztett MorphoTM (korábbi nevén MetaMorpho TM[72]) egy olyan EBMT-alapú[73], nyelvi tudásra is építő fordítómemória (TM), amely nem csak egész mondatpárokat, hanem a mondatnál kisebb részek párjait is tárolja. Jelenleg a teljes mondatpárokon kívül főnévi csoportokat (NP), illetve a mondatban ezeket szimbolikus NP helyekre cserélve kapott mondatvázakat tárolunk fordításaikkal együtt a fordítómemória adatbázisában. A keresés során, amennyiben a keresett mondathoz kellőképp hasonló forrásoldallal rendelkező teljes mondatpár nem található a

[72] A névváltoztatást a MetaMorpho fordítórendszer és a MorphoTM fordítómemória jobb megkülönböztethetősége indokolta. A MorphoTM fordítómemóriában továbbra is felhasználjuk a MetaMorpho fordítórendszerhez kifejlesztett szintaktikai elemezőt és nyelvtanokat, azonban a rendszer jelentős része független a MetaMorpho gépi fordítórendszertől, amely egyébként tartalmaz fordítómemória jellegű bővíthetőséget, ezért is szükség volt a névváltoztatásra.

[73] EBMT (Example Based Machine Translation) – minta alapú gépi fordítás. A fordítórendszer emberi fordításokból származó mintákból készít fordításokat, a tisztán statisztikai fordítórendszerektől abban különbözik, hogy tárolt nyelvi tudásra is épít.

memóriában, viszont megtalálhatók a keresett mondat vázához és főnévi csoportjaihoz kellőképp hasonló forrásoldallal rendelkező mondatrész párok, akkor az utóbbiak fordításoldalából – a megfelelő morfológiai alakok generálásával – építünk javasolt fordítást [1][2][3].

A főnévi csoportok fordításaikkal való összerendelését (szinkronizációját, párhuzamosítását) nem bízhatjuk a fordítómemóriát használó fordítóra, mert a főnévi csoportok megjelölésére és összerendelésére fordított munkaidő nem feltétlenül térülne meg a későbbiekben a fordítómemória fedésének növekedése révén. Ezen kívül a fordítómemória motor más fordítómemória rendszerekbe (pl. MemoQ) való beépítését is nehezítené, ha a felhasználói felületen a szokásos funkcióktól eltérőeket is megkövetelne. A főnévi csoportokat tehát automatikus módszerekkel határozzuk meg és szinkronizáljuk fordításaikkal.

Az automatikus főnévicsoport-szinkronizáció – teljes pontosságot biztosító módszer hiányában – alapvető hibaforrásként jelenik meg a rendszerben, a hibásan tárolt párokat később plusz munkával kell eltávolítani a memóriából, ezért az alkalmazott módszerekkel szembeni legfontosabb elvárás, hogy magas pontossággal határozzák meg a főnévicsoportpárokat. Emellett a magas fedés annyiban fontos, hogy ha csak kevés főnévi csoporthoz tudunk párt rendelni, akkor a fordítómemória fedése nem lesz sokkal nagyobb egy csak teljes mondatok kezelésére képes fordítómemóriáénál. A páratlan főnévi csoportokat a mondatváz részeként tároljuk (1. példa), hiszen ezeket a mondatvázból kiemelve nem tudnánk automatikusan meghatározni a maradék fordítását. Ez azt eredményezi, hogy alacsony fedésű módszerek alkalmazása esetén a mondatvázak a bennük tárolt párnélküli főnévi csoportoktól speciálisabbak lesznek, így kevésbé remélhetjük, hogy egy későbbi fordítási feladat során felhasználhatók lesznek.

> [I] have read [his new book on bread baking] and [I] am going to try [one of his recipes].
>
> Elolvastam [a kenyérsütésről szóló új könyvét] és ki fogom próbálni [egy receptjét]. (1. példa)

1. példa: Az angol *I* személyes névmáshoz nem található a magyar fordításban neki megfeleltethető főnévi csoport, illetve tegyük fel, hogy az automatikus módszerrel nem sikerült egymáshoz rendelni a mondatok utolsó főnévi csoportjait. (A példákban a maximális méretű főnévi csoportokat szögletes zárójelek határolják.) Ekkor a memóriába a teljes mondatpáron kívül az *NP1 := his new book on bread baking = a kenyérsütésről szóló új könyv* főnévicsoportpár, illetve az *I have read [NP1] and now I'm going to try one of his recipes = Elolvastam [NP1] és ki fogom próbálni [egy receptjét].* mondatvázpár kerül. A szimbolikus NP helyek megőrzik az eredeti főnévi csoportok morfológiai tulajdonságait, jelen esetben a *book* egyes számú voltát, illetve hogy a *könyvét* egyes számú, tárgy esetű.

A pontosság és a fedés mellett nagyon fontos az alkalmazott főnévicsoport-meghatározó és szinkronizáló módszerek sebessége, hiszen a fordító jogosan várhatja el, hogy a tárolt mondatpárok főnévi csoportjainak fordításai akár már a következő mondat fordításakor is megjelenjenek a javaslatok között, hiszen a hagyományos fordítómemóriák is gyorsan tárolják, és azonnal elérhetővé is teszik a fordításokat.

A MorphoTM rendszerben a korábbiakban két módszert javasoltunk a főnévi csoportok meghatározására [4][5]. Első ötletként a tárolt mondatpár angol és magyar oldalán is mondatelemzővel határoztuk volna meg a főnévi csoportokat, ezt a mód-

szert azonban a MetaMorpho magyar nyelvtanának [6][7] akkori alkalmatlansága miatt el kellett vetnünk, és kidolgoztunk egy módszert a magyar főnévi csoportok angol párjaik alapján, szótári és szófaji megfeleltetésekkel valamint sekély nyelvtani szabályokkal történő meghatározására. A MetaMorpho magyar elemző fejlődését figyelembe véve ebben a cikkünkben a korábbi ismertetésnél bővebben is összehasonlítjuk a két módszert, részletesebben elemezve az egyes módszerek előnyeit és hátrányait, illetve bemutatjuk a két módszer ötvözésének lehetőségeit, rámutatva arra, hogy a főnévi csoportok fordítás alapján történő meghatározása akkor is hasznos lehet, ha mindkét nyelvhez jó, de sajnos nem tökéletes, illetve lassú mondatelemzővel rendelkezünk.

Cikkünk második felében bemutatjuk, hogy a főnévicsoportpárok hasonlóságának meghatározására kidolgozott szótáralapú módszerünkben [4][5] a heurisztikus súlyozást hogyan váltottuk ki gépi tanulás alkalmazásával.

Végül a mérési eredmények ismertetése után a MorphoTM továbbfejlesztésével kapcsolatos terveinkről is szót ejtünk majd.

2 Főnévi csoportok automatikus meghatározása

A főnévi csoportok pontos és gyors meghatározása alapvető fontosságú feladat a MorphoTM rendszerben. A főnévi csoportok jó minőségű szinkronizálásához pontos főnévicsoport-meghatározó módszer szükséges azért, hogy a szinkronizáló algoritmusnak már lehetőleg csak párokat kelljen egymáshoz rendelni, és ne kelljen a hibásan meghatározott főnévi csoportokból adódó hibák szűrésével foglalkoznia. Tökéletes főnévicsoport-meghatározó módszert feltételezve elég lenne néhány egymásnak megfeleltethető szó, hogy egymáshoz rendeljük az egyes főnévi csoportokat, enélkül azonban a szinkronizáló módszernek kell elkerülnie az adatbázis hibás párokkal való feltöltését. Kicsit szabatosabban úgy fogalmazhatnánk, hogy minél pontosabban tudjuk meghatározni a főnévi csoportokat, annál nagyobb fedésűre hangolhatjuk a szinkronizáló algoritmusunkat.

A pontosság mellett a fordítómemória egésze (és nem a főnévicsoport-szinkronizáló algoritmus) szempontjából a fedés is fontos, hiszen önmagában nem sokat érünk egy tökéletesen pontos, módszerrel, ha az a főnévi csoportoknak csak kis részét képes felismerni.

A pontosság és fedés mellett azonban sajnos nem feledkezhetünk meg arról sem, hogy gyors módszerre van szükség. Tökéletes módszer híján egyszerre mindhárom cél sajnos nehezen közelíthető (nincs új a nap alatt).

A következőkben a MorphoTM rendszerben alkalmazott főnévicsoport-meghatározó módszereket fogjuk bemutatni és összehasonlítani, majd a módszerek ötvözésére vonatkozó javaslatot fogunk bemutatni.

2.2 Főnévi csoportok meghatározása szintaktikai elemzővel

A főnévi csoportok meghatározásának legalapvetőbb módszere szintaktikai elemző alkalmazása. A MorphoTM rendszerben az angol mondatokat a MetaMorpho angol nyelvtanát használva elemezzük. Amennyiben több elemzés is születik, jelenleg csak az elsőt vizsgáljuk. Amennyiben nem áll össze a teljes mondat elemzése akkor a

lehetséges részfákból a MetaMorpho heurisztikus gyökérválogató[74] módszereivel választunk egy elemzést, így a szinkronizáció során már csak egyetlen elemzés főnévi csoportjaihoz keresünk párt.

A MetaMorpho magyar nyelvtanának fejlődése lehetővé teszi, hogy a magyar mondatok főnévi csoportjait is szintaktikai elemzővel válasszuk ki, azonban egyelőre a következő pontban bemutatott módszerünket használjuk.

2.2 Főnévi csoportok meghatározása fordításaik ismeretében

Tavaly módszert mutattunk [4] főnévi csoportok fordításaik alapján szótárral, illetve sekély nyelvtannal történő meghatározására, jelenleg a MorphoTM rendszerben ezt a módszert használjuk a magyar főnévi csoportok meghatározására. A módszert most részletesen nem ismertetjük újra, csak a leglényegesebb elemeit foglaljuk össze.

Az angol elemzővel automatikusan meghatározott főnévi csoportokhoz rendelhető magyar főnévi csoportokat az angol főnévi csoportok szavait és kifejezéseit a magyar szövegre leképezve próbáljuk meghatározni. Az angol főnévi csoport nem csak grammatikai funkciót betöltő szavainak lehetséges fordításait tövesített szótári kereséssel (többszavas kifejezéseket is keresve), illetve hasonló alakú szavakat (*cognate*) keresve [8] próbáljuk a magyar mondatban megtalálni. Mivel egy angol szó több megfelelője és akár többször is előfordulhat a magyar mondatban, a lehetséges találatok közül azt választjuk ki, amelynek szavai a lehető legrövidebben illeszkednek a magyar mondatra. Természetesen a találatok között más szavakat is tartalmazhat a kijelölt illeszkedés. Az illeszkedést ezek után egyszerű szabályok szerint, az angol főnévi csoport le nem fedett szavainak szófaját is figyelembe véve teljes magyar főnévi csoporttá bővítjük.

A módszer előnye, hogy más nyelvekhez is könnyen elkészíthető, mindössze egy morfológiai elemző, egy szótár és egy főnévi csoportok meghatározására használható sekély nyelvtani szabályrendszer szükséges hozzá (természetesen a nyelvpár másik oldalán továbbra is szükség lenne szintaktikai elemzőre).

2.3 Magyar főnévicsoport-meghatározó módszerek összehasonlítása

Az összehasonlítás főbb szempontjairól (pontosság, fedés, sebesség) már írtunk a 2. szakasz elején, most lássuk, hogyan felelnek meg az egyes feltételeknek a fenti módszerek.

Legkönnyebben a módszerek sebessége illetve erőforrásigénye mérhető. A szintaktikai elemzőt nem használó módszer néhány ezredmásodperc alatt lefut hosszabb mondatpárokon is, ezzel szemben a MetaMorpho magyar szintaktikai elemző sajnos egyelőre nem mondható gyorsnak. Hosszú, összetett mondatoknál az elemzés egy átlagos személyi számítógépen akár percekig is eltarthat (közben a viszonylag kötetlen szórendből és a nyelv egyéb sajátosságaiból adódóan sok esetben 10 milliónál is több lehetséges csomópontot azonosít az elemző). Érthető, hogy a nyelvtan fejleszté-

[74] A MetaMorpho gépi fordítórendszerben a gyökérválogató módszereket mozaikfordítások készítésére használják, így akkor is képes a rendszer valamiféle fordítást adni, ha a teljes forrásmondatot nem tudta egyetlen fával lefedni.

sénél jelenleg elsődleges szempont a pontosság és fedés növelése, de a későbbiekben a fejlesztőknek az erőforrásigénnyel is szembe kell majd nézniük.

A fedés terén a szinkronizációs alkalmazásunk esetében egyelőre szintén a szintaktikai elemzőt nem igénylő megoldás tűnik jobbnak. 100 tesztmondatunkból 1-re valamiért nem futott le a magyar elemző, 17 mondat esetében pedig semmilyen elemzés nem született. A 82 elemzéssel rendelkező tesztmondatból (csak az első elemzéseket nézve) 16 esetében a mondat ismeretlen szavai következtében egyáltalán nem talált főnévi csoportot az elemző (a valóságban minden mondatban volt legalább egy főnévi csoport). A maradék 66 mondat esetében sem talált meg minden főnévi csoportot az elemző, illetve sokszor nem találta meg a maximális méretű főnévi csoportokat, csupán részeiket. Az MetaMorpho angol elemzőhöz viszonyítva ezek az eredmények még javulhatnak (feltehetően fognak is, hiszen a magyar elemző fejlesztése folyamatosan tart, az egy évvel ezelőtti állapothoz képest hatalmas javulást tapasztaltunk).

Az elemzőt nem igénylő megoldás a szinkronizáció szempontjából előnyös módon keresi a magyar főnévi csoportokat, azokat próbálja meg kijelölni, amelyeket a szintén szótári és szófaji megfeleltetéseket figyelő szinkronizációs módszer feltehetően egymáshoz rendel majd. A fedést azonban a módszer esetleges pontatlansága ronthatja, hiszen nem az a kérdés, hogy hány főnévicsoport-jelöltet talál a módszer, hanem, hogy hány valódi főnévi csoportot.

A pontosság terén egyértelműen jobbnak bizonyult a MetaMorpho magyar nyelvtana. A 2. példán látható, hogy az elemzőt nem használó módszer sok esetben rosszul határozza meg a főnévi csoport határait.

EN: *The Ombudsman has wide powers of investigation.*
HU: *Az Ombudsman széleskörű vizsgálati jogkörrel rendelkezik.*

```
NP_EN1>  The Ombudsman
NP_HU1>  Az Ombudsman

NP_EN2>  wide powers of investigation
NP_HU2>  Az Ombudsman széleskörű vizsgálati jog-
körrel
```

(2. példa)

2. példa: Az angol mondatban a MetaMorpho angol elemzővel talált főnévi csoportok és a magyar oldalon szintaktikai elemzőt nem használó módszerrel hozzájuk rendelt főnévi csoportok. Látható, hogy a második magyar főnévi csoport határát nem tudta pontosan meghatározni a módszer. (Ez a probléma orvosolható lenne, ha – balról jobbra haladva – a szükséges hasonlósági értéket elérő főnévi csoportok által lefedett szavakat foglaltnak jelölnénk, és nem használnánk őket más főnévi csoportban, ez a módszer azonban más problémákat is felvet, például, ha egy főnévi csoport egy másik részeként és önállóan is szerepel a mondatban, akkor a hosszabbat esetleg nem tudjuk így azonosítani.) A MetaMorpho magyar elemző helyesen azonosítja mindkét főnévi csoportot.

Az elemzőt nem használó módszert a 100 tesztmondatunkban 1 alkalommal megzavarta, hogy a tövesített szótári megfeleltetésnél eddig nem vizsgáltuk a szavak szófaját, így egy mondatpárban a *the tag* (= *a dögcédula*) főnévi csoporthoz a *jelölték* magyar szót rendelte az algoritmus (a *tag=jelöl* pár is szerepelt a szótárban).

Olyan esetek is előfordultak, ahol segített az elemzőt nem használó módszer nagyobb flexibilitása. Néhány esetben az angol elemző a mondat szabad bővítményeit helytelenül a főnévi csoporthoz csapta, ekkor egy szabályos főnévi csoportokat kere-

ső magyar elemzővel nem tudtunk volna párt találni ezekhez a „főnévi csoportokhoz”, azonban egy kis csalással a magyar főnévi csoporthoz hozzácsapva a szabad bővítményt már jó mondatvázpárt tudtunk az adatbázisba helyezni. Ezt a „kis csalást” az elemzőt nem használó módszerünk automatikusan elvégezte.

2.4 Javaslat a magyar főnévicsoport-meghatározó módszerek ötvözésére

A főnévicsoport-meghatározó módszereinket kétféleképp is ötvözhetjük. Egyrészt felhasználhatjuk az angol mondat főnévi csoportjait arra, hogy a lehetséges magyar elemzések közül olyat válasszunk, amelynek főnévi csoportjai tartalmazzák az angol főnévi csoportokhoz tövesített szótári megfeleltetéssel, illetve hasonló szavakat keresve hozzárendelt magyar szavakat. Ezzel az elemzési időt nem tudnánk csökkenteni, az elemzések pontossága azonban tovább nőhetne.

Másrészt leválthatjuk a magyar mondatban kijelölt főnévicsoport-vázat teljes főnévi csoporttá bővítő egyszerű szabályrendszert a MetaMorpho elemzőre. Ennek a megoldásnak előnye, hogy pontosabb eredményt tudunk majd elérni vele az eddigieknél, ugyanakkor remélhetjük, hogy az elemzési idővel se lesznek gondjaink, mivel az elemző csak hosszú mondatokra lassú, néhány szavas főnévi csoportokat gyorsan elemez. Így a néhány lehetséges főnévi csoport ellenőrzése se tartana sokáig. Kérdés viszont, hogy mennyire fogja pontosan meghatározni a főnévi csoportok határait a teljes mondat ismerete nélkül a MetaMorpho magyar nyelvtana.

3 A főnévi csoportok hasonlóságának meghatározása

A főnévi csoportok meghatározása után el kell dönteni, hogy a lehetséges párjelöltek közül melyeket rögzítsük párként az adatbázisban. Ha a mondatpár mindkét oldalán szintaktikai elemzővel jelöltük ki a főnévi csoportokat, akkor az összes lehetséges párosítást meg kell vizsgálnunk; ha az egyik nyelv esetében – szintaktikai elemzőt nem használva – a mondat fordításának főnévi csoportjai alapján határoztuk meg a főnévicsoport-jelölteket, akkor csak azt kell megvizsgálnunk, hogy ezek tényleg eléggé hasonlítanak-e párjaikra.

A feladat mindkét esetben az, hogy egy párjelölthöz egyetlen, a hasonlóságot jól jellemző skalár értéket rendeljünk. (A mindkét oldalon elemzőt használó módszer esetében lehetséges, hogy egy főnévi csoport több másikhoz is hasonlít, az ilyen helyzetek feloldását tavalyi cikkünkben [4] ismertettük.)

A MorphoTM rendszerben a hasonlóság mérésére a tavaly kifejlesztett, szótári és szófaji megfeleléseket kereső módszerünket [4] és ennek most bemutatott újabb változatát használjuk. A következőkben ezeket a módszereket fogjuk bemutatni és értékelni.

3.1 Szótáralapú és szófaji megfeleltetés

A hasonlósági vizsgálat során az összehasonlított két főnévi csoport szavait egymás után többféle módon is megpróbáljuk egymásnak megfeleltetni, majd az egyes mód-

szerek által lefedett tokenek számából számítjuk ki a hasonlóságot jellemző skalár értéket.

Először tövesített szótári keresést alkalmazunk: a forrásnyelvi főnévi csoport szavainak lehetséges töveit keressük egy speciális, tövesített indexet és találatlistát tartalmazó szótárban, majd a találatok közül csak azokat hagyjuk meg, amelyek a forrásoldalra illeszthetők és fordításuk minden szavának legalább egy lehetséges töve megtalálható a fordításbeli főnévi csoportban. A szótár segítségével többszavas kifejezéseket is keresünk. A főnévicsoportpárban így lefedett tokenek számát ettől kezdve **D**-vel jelöljük.

A szótári megfeleltetés után, a főnévicsoportpár le nem fedett, nagybetűt vagy számot tartalmazó szavai között hasonló alakúakat (*cognate,* [8]) keresünk. A főnévicsoportpárban így lefedett tokenek számát ettől kezdve **C**-vel jelöljük.

A korábban le nem fedett szavakat ezután szófajaik alapján próbáljuk egymáshoz rendelni. A szófajuk alapján megfeleltetett tokenek számát ettől kezdve **P**-vel jelöljük.

Végül a lefedetlen szavak közül kiválogatjuk a pusztán grammatikai funkciót betöltőket, ezeket az összehasonlítás során kisebb súllyal vehetjük majd figyelembe, hiszen nem jelentenek lényegi különbséget a két főnévi csoport között. A pusztán grammatikai funkciót betöltő szavak számát **F**-fel jelöljük.

A hasonlóság mértékét kiszámító módszerekben szükségünk lesz még a két főnévi csoport szavainak (most tokenjeinek) számára, ezt **W**-vel jelöljük.

(Az egyes megfeleltetési lépéseket korábbi cikkünkben [4] bővebben ismertettük, itt most ezért nem tértünk ki a részletekre.)

3.2 Heurisztikus hasonlósági érték

Eddig a MorphoTM rendszerben az előzőekben ismertetett tulajdonságjegyekből az 1. képletben meghatározott heurisztikus hasonlósági értéket alkalmaztuk, azokat a főnévicsoportpárokat tekintve tárolandó párnak, amelyeknél a hasonlósági érték meghaladott egy küszöbértéket.

$$h = \frac{1 \cdot D + 0{,}9 \cdot C + 0{,}3 \cdot P - 0{,}1 \cdot F}{W - F} \qquad \text{(1. képlet)}$$

1. képlet: Az eddigiekben alkalmazott *h* hasonlósági érték számításának módja. Azokat a frázispárokat tekintettük tárolandó párnak, ahol a *h* hasonlósági érték meghaladta a 0,67 küszöbértéket. A képletbeli együtthatókat néhány mondatpáron, kísérletezéssel állapítottuk meg.

3.3 Gépi tanulással meghatározott osztályozó

Bár a korábbi heurisztikus képlet (1. képlet) a gyakorlatban jónak tűnt, úgy döntöttünk, hogy kis korpuszt építve megvizsgáljuk, hogyan válthatnánk ki egy gépi tanulással meghatározott (azaz empirikus alapokon nyugvó) osztályozóval.

Az eredmények mérhetőségén kívül a gépi tanulás mellett szólt az is, hogy így a későbbiekben lehetőségünk lesz az eddigiek mellett újabb hasonlóságvizsgálati módszerek kipróbálására is, az egyes módszerek összevetése egyszerűen megoldható lesz.

Az alábbiakban bemutatjuk, hogyan építettünk egy kis tanító illetve tesztkorpuszt, hogyan normalizáltuk a 3.2 pontban felsorolt tulajdonságjegyeket (*feature*), majd bemutatjuk a WEKA gépitanuló-rendszerben [9] különböző osztályozókkal elért eredményeket.

3.3.1 Korpuszkészítés

A tanító illetve tesztkorpusz építésekor elsődleges célunk az volt, hogy a tényleges osztályozási feladatot szimuláljuk. A tesztkorpuszhoz mindenféle szempontok nélkül 100 angol-magyar mondatpárt választottunk ki, a mondatok átlagos hossza 14 szó volt. A mondatpárokból úgy építettünk tesztkorpuszt, hogy az angol oldalon a MetaMorpho angol elemzővel kijelöltük a főnévi csoportokat, majd lehetséges párjaikat a 2.2 pontban ismertetett elemzőt nem igénylő algoritmussal határoztuk meg, mivel jelenleg ezt a módszert használjuk a magyar főnévi csoportok kijelölésére. Ezek után minden egyes párt megvizsgáltunk, és kézzel megjelöltük, hogy helyesnek találjuk-e, majd automatikusan meghatároztuk a 3.1 pontban ismertetett tulajdonságjegyeket. Így egy olyan korpuszt kaptunk, amely a kinyert tulajdonságjegyek mellett tartalmazta, hogy tárolandó párnak tekintjük-e a tulajdonságjegyekkel jellemzett frázispárt.

Elvetettük azokat a párokat, ahol a főnévi csoportok közül az egyik csak a másik egy részének fordítását tartalmazta, azaz csak teljesen megfeleltethető párok egymáshoz rendelését fogadtuk el.

Azokban az esetekben, amikor a mondat fordítása más főnévi csoportokkal esetleg nem lett volna jó, de a főnévi csoportok összerendelése helyes volt, a főnévicsoportpárokat elfogadtuk, egy ilyen mondatpárt találunk a 3. példában is.

EN: *There were pictures of castles and lakes and pretty girls on the walls.*
HU: *A falakra kastélyok és tavak és szép lányok képeit ragasztották.*

```
NP_EN1>  pictures of castles and lakes and pretty girls
NP_HU1>  kastélyok és tavak és szép lányok képeit

NP_EN2>  the walls
NP_HU2>  a falakra
```

(3. példa)

2. példa: Ha a főnévi csoportok összerendelése helyes, akkor is elfogadjuk őket, ha mondatváz fordítása más főnévi csoportokkal esetleg rossz lenne. A mondatvázak jelen esetben az angol oldalon *There were [NP1] on [NP2]*, illetve a magyar oldalon *[NP2] [NP1] ragasztották*, amelyek csak ritkán feleltethetők meg egymásnak. A későbbiekben a mondatvázfordítások megfelelőségét is érdemes lehet vizsgálni.

Abban az esetben is elfogadtuk a főnévicsoportpárt, amikor a főnévi csoportok a mondatban egymás fordításai voltak, de a fordító apró, a lényeges szavakat nem érintő változtatást ejtett, például az angol *this family* főnévi csoportot magyarul *a családnak* fordította. Ezt azért engedtük meg, hogy jobb mondatvázpárt tárolhassunk az adatbázisban. Ehelyett a későbbiekben a mondatvázpárok megfelelőségét is érdemes lehet vizsgálni.

A korpuszt a WEKA rendszer által használt ARFF formátumban készítettük el, az olvashatóság és módosíthatóság érdekében kommentben rögzítve az egyes mondatpárokat és a belőlük kinyert főnévi csoportokat.

Az elkészült korpusz 186 helyes és 42 helytelen összerendelést tartalmazott.

3.3.2 Az adatok normalizálása

Az előző pontban ismertetett mintahalmazunkat vizsgálva azt találtuk, hogy a szótári megfeleltetéssel (D), a hasonló alakú szavakat keresve (C) és a hasonló szófajú szavakat keresve(P) megfeleltetett szavak száma – a várakozásnak megfelelően – szinte lineáris korrelációban áll a szavak számával (W). Ez után megvizsgáltuk, hogy az 1. (heurisztikus) képletben alkalmazott normalizálást is (vagyis a pusztán grammatikai funkciót betöltő, meg nem feleltetett szavak kihagyását), amely kis mértékben jobbnak bizonyult a kis korpuszunkon, a később bemutatott osztályozók megbízhatóságát 1-3%-kal növelte. Az így kapott normalizált tulajdonságjegyeket a 2. képletben rögzítettük.

$$\frac{D}{W-F}, \frac{C}{W-F}, \frac{P}{W-F}, \frac{F}{W-F} \qquad \text{(2. képlet)}$$

2. képlet: A főnévi csoportok hasonlóságának meghatározására használt tulajdonságjegyek normalizálása. Az empirikus képlet egybevág az elméleti meggondolásainkkal, a pusztán grammatikai funkciót betöltő szavak elhagyhatók az összehasonlítás során, illetve hosszabb főnévicsoport-párokban a szavak számával egyenesen arányosan több szót tudunk lexikai módszerekkel egymásnak megfeleltetni.

3.3.4 Tanulóalgoritmusok

A WEKA rendszerben elérhető osztályozó algoritmusok közül többet is kipróbáltunk. Próbálkoztunk többrétegű perceptronokból épített neurális hálóval (MLP), amelyet a szokásos *back-propagation* eljárással tanítottunk; próbálkoztunk RBF hálóval, ahol a radiális bázisfüggvényeket *K-means* klaszterezéssel állapítottuk meg, illetve kipróbáltuk a C4.5 döntési fa WEKA rendszerbeli J48 nevű implementációját. A komolynak tekinthető tanulóalgoritmusok mellett két egyszerűbb osztályozót is kipróbáltunk, egy lineáris regressziós modellre épülőt illetve egy logisztikus regressziós osztályozót.

A lineáris regressziós modell, az előzőekben ismertetett normalizálással gyakorlatilag az 1. képlet együtthatóinak korpuszalapú meghatározását jelenti.

A logisztikus regressziós modell hasonlít a lineáris regressziósra, ugyanakkor a paraméterek beállításakor a helyes döntések illetve a helytelen döntések arányát maximalizálja (3. képlet), illetve előnye, hogy kimenete valószínűség, azaz közvetlenül használható egy főnévicsoportpár hasonlóságának értékelésekor (4. képlet).

$$\ln\left(\frac{P(\text{fordítás})}{1-P(\text{fordítás})}\right) = \alpha\frac{D}{W-F} + \beta\frac{C}{W-F} + \gamma\frac{P}{W-F} + \delta\frac{F}{W-F} + \varepsilon \qquad \text{(3. képlet)}$$

3. képlet: A logisztikus regressziós osztályozó a tanulás során a helyes döntések meghozásának valószínűségét maximalizálja a görög betűkkel jelölt együtthatók beállításával.

$$P(\text{fordítás}) = \frac{1}{1 + \exp(-\alpha \frac{D}{W-F} - \beta \frac{C}{W-F} - \gamma \frac{P}{W-F} - \delta \frac{F}{W-F} - \varepsilon)} \qquad \text{(4. képlet)}$$

4. képlet: A logisztikus regressziós osztályozó kimenete annak a valószínűsége, hogy a vizsgált frázispár egymás fordítása.

3.3.5 Eredmények

Az egyes osztályozókat tízszeres keresztkiértékeléssel (*10-fold cross-validation*) tanítottuk és teszteltük. Azt vizsgáltuk, hogy a 228 mintából hány esetben döntenek helyesen. Viszonyítási alapként (*baseline*) tekinthetjük azt, hogy minden párjelöltet párnak tekintünk, így az osztályozónk a korpusz adottságainak megfelelően 186 esetben hozna helyes döntést.

Ha csak az osztályozók döntési pontosságát vizsgáljuk, akkor egy helytelen pár rögzítését egy helyes pár fel nem vételével azonos hibának tekintjük. A C4.5 döntési fa kivételével ez nem okoz gondot, a többi osztályozó kimenete egy 0 és 1 közötti valós érték, amelyet nem feltétlenül a 0,5 értéknél kell elvágnunk, például a logisztikus regressziós osztályozó valószínűségi kimenetét tekintve dönthetünk úgy is, hogy csak a 80% valószínűséggel fordításnak tekinthető párokat rögzítjük az adatbázisban.

A viszonylag bonyolult MLP, RBF és C4.5 osztályozóknál jobban szerepeltek a regressziós osztályozók, amelyek nagyjából azonos eredményeket értek el. A tanított osztályozók mellett a teljes tesztanyagon (azaz nem keresztkiértékeléssel) megvizsgáltuk a tavaly meghatározott képlettel elérhető osztályozási pontosságot is. Az eredményeket az 1. táblázatban foglaltuk össze.

osztályozó	**# helyes döntés**	**# helyes pár el nem fogadása**	**# helytelen pár elfogadása**
baseline	186	0	42
MLP + back-propagation	191	14	23
RBF + K-means	188	12	28
C4.5 (J48)	193	4	31
lineáris regresszió	**196**	8	24
logisztikus regresszió	**196**	7	25
régi heurisztikus képlet	184	14	30

1. táblázat: Az egyes osztályozókkal elért eredmények.

Látható, hogy a tavaly meghatározott heurisztikus képlet sajnos a baseline-nál is rosszabbul teljesített; érdekes azonban, hogy a tőle csak együtthatóiban különböző lineáris regressziós osztályozó lett a legjobb, nagyjából azonos eredményt érve el a logisztikus regressziós osztályozóval. A tanulással meghatározott együtthatók esetében a legfőbb különbség az volt, hogy a szófaji megfeleltetés jóval nagyobb pontszámot kapott.

Érdekes volt, hogy a korpusznak csak az első 50 mondatpárján tanítva és tesztelve az osztályozókat sokkal jobb eredményt értek el a regressziós modellek, a bonyolultabb osztályozók viszont a teljes korpuszon mérténél is jobban lemaradtak, feltehetően a tanítóminták kis számából adódóan. Ez azt is jelentheti, hogy a jelenleginél nagyobb korpuszon ezek az osztályozók behozhatják mostani lemaradásukat. A nagyobb korpuszból véletlenszerűen választva 50 mondatpárt a regressziós osztályozók eredményei már jobban hasonlítottak a teljes korpusznál mértekre, a bonyolultabb osztályozók viszont hasonlóan rosszul teljesítettek.

A korpusz növelésén túl azt is érdekes lesz megvizsgálni, hogy ha mindkét nyelv esetében elemzővel kiválasztott főnévi csoportokat vetünk össze, milyen eredményeket tudunk majd elérni az egyes módszerekkel.

4 Összefoglalás

Cikkünkben bemutattuk az elmúlt egy év a MorphoTM rendszer főnévicsoport-szinkronizáló moduljával kapcsolatos legfőbb eredményeit. A MetaMorpho magyar nyelvtanával pontosan tudnánk főnévi csoportokat keresni, de az elemzési sebesség sajnos ezt még nem teszi lehetővé, és a nyelvtan fedése is sokat javulhat még. A főnévi csoportok meghatározásának kérdése után bemutattuk, hogyan váltottuk le gépi tanuló módszerek alkalmazásával az eddig használt heurisztikus főnévicsoport-hasonlóságot mérő képletünket (pontosabban csak együtthatóit, hiszen egy hasonló struktúrájú képlet bizonyult a legjobbnak).

5 További tervek

A bemutatott módszereink további vizsgálata érdekében tervezzük egy nagyobb főnévicsoport-szinten párhuzamosított korpusz építését.

Főnévi csoportokon kívül lehetőségünk lenne más mondatrészeket is tárolni a fordítómemória adatbázisában (pl. melléknévi csoportok, határozói szerkezetek), ezekkel eddig nem foglalkoztunk, de úgy gondoljuk, hogy kezelésüket a főnévi csoportokéhoz hasonlóan tudnánk megoldani.

A 3. példában rámutattunk arra, hogy esetleg érdemes lehetne a mondatvázak hasonlóságát is vizsgálni, hogy csak feltehetően helyes fordításokat tároljunk az adatbázisban. Felvetődik azonban a kérdés, hogy a helyességvizsgálattal hány valójában megfelelő párt vetnénk el.

A maximális méretű főnévi csoportokon belüli kisebb főnévi csoportok szinkronizációját is hasznosnak tartjuk. Az 1. példa főnévi csoportjaira visszatekintve láthatjuk, hogy viszonylag kis többletmunkával megsokszorozhatnánk az adatbázisba felvett főnévi csoportok számát.

A MetaMorpho magyar nyelvtanának fejlődésével lehetőségünk lesz a jelenlegi szómegfeleltetéseket használó módszereken túl elemzésifa-szinkronizáló módszerek vizsgálatára is.

Bibliográfia

[1] Hodász G., Grőbler T., Kis B.: Translation Memory as a Robust Example-based Translation System. In Proceedings of the Ninth EAMT workshop, University of Malta, Valletta, pp. 82-89, 2004.

[2] Hodász G.: Nyelvi hasonlóságon alapuló intelligens keresés fordítómemóriában. In *II. Magyar Számítógépes Nyelvészeti Konferencia* (szerk.: Alexin Z, Csendes D.), Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, pp. 108-116, 2004.

[3] 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.

[4] Pohl G.: Angol–magyar szótáralapú főnévicsoport-szinkronizáció és fordításalapú főnévicsoport-meghatározás. In *III. Magyar Számítógépes Nyelvészeti Konferencia*, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, pp. 125-133, 2005.

[5] Pohl G.: English-Hungarian NP Alignment in MetaMorpho TM. In *EAMT 11th Annual Conference, 2006,* (CD-ROM)

[6] Tihanyi L.: A MetaMorpho fordítóprogram projekt 2005-ben. In *III. Magyar Számítógépes Nyelvészeti Konferencia* (szerk.: Alexin Z, Csendes D.), Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, pp. 99-107, 2005.

[7] Merényi Cs.: A MetaMorpho magyar-angol gépi fordító rendszer ige vonzatkereteit működtető nyelvtan. In *III. Magyar Számítógépes Nyelvészeti Konferencia*, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, pp. 108-115, 2005.

[8] 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

[9] Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Részleges gépi fordítás a NooJ rendszerben

Váradi Tamás

MTA Nyelvtudományi Intézet
1068 Budapest Benczúr u. 33

Kivonat Az előadás ismerteti azokat az eredményeket, amelyeket a NooJ nyelvelemző fejlesztőrendszer[8][5] gépi fordításra való alkalmazása terén elértünk. Bemutatja a rendszer azon új képességeit, amelyek alkalmassá teszik a lokális grammatikákat kétnyelvű felhasználásra. A lokális grammatikák kiválóan alkalmasak az egyedi lexikai szabálytól a szóosztályokra érvényes általános szabályok megfogalmazására. A dolgozat fő tézise, hogy a rendkívül könnyen kezelhető, gyors és interaktív NooJ rendszer jól alkalmazható részleges gépi fordítást igénylő feladatokra.

Kulcsszavak: gépi fordítás, véges állapotú nyelvelemezés, lokális grammatika, NooJ

1. Bevezetés

A jelen dolgozat célja, hogy bemutassa azokat a lehetőségeket, amelyeket a NooJ keretfejlesztő rendszer gépi fordításhoz kínál. Egy gépifordító rendszer létrehozása természetesen rendkívül összetett folyamat, amely kihívást jelent a számítógépes nyelvészet egésze számára, és nagy szerep jut benne a szoftver technológiának is. Ebben a dolgozatban az alulról építkezés jegyében áttekintjük azokat az elveket, amelyek szerint a mondatok gépi fordítása legalább részlegesen elvégezhető, és bemutatjuk ezek megvalósítását a NooJ rendszerben. A 2. részben ismertetjük a részleges, minta alapú gépi fordítás elveit, a a 3. részben bemutatjuk a NooJ eszköztárát, amellyel lehetővé válik gépi fordítás megvalósítása, a 4. részben példákat mutatunk be a főnévi csoportok fordítására.

2. Elméleti háttér és motiváció

A nyelvi modell, amely a jelen munkálatot motiválja Morris Gross lokális grammatika fogalmára[3] épül. Ez az elmélet a végesállapotú technológiával jól kezelhető lokális grammatikai viszonyokra helyezi a hangsúlyt, sok tekintetben a generatív grammatika uralkodó ágával szemben, ahol a távoli függőségek vizsgálata a domináns. Gross a ma már igen elterjedt lexikális nyelvtanok, sőt mondhatni a konstrukciós grammatika előfutárának tekinthető. További fontos jellemzője munkásságának nemcsak a lexikon és szintaxis határának, de a szintaxis és szemantika közötti viszony újrafogalmazása. A nyelvi szerkezetek szemantikai alapú megközelítésével a szellemi elődének tekintett Harris[4] munkásságát követi.

2.1. Minta alapú fordítás

A gépi fordítás technológiáját tekintve eljárásunk alulról felfelé haladó (bottom up) *minta alapú fordításnak* tekinthető, amely az alábbi elvekben megegyezik a Metamorpho rendszer[6][7] megközelítésével. A minta fogalma a lokális grammatikában is folytonos átmenetet jelent az egyedi szavakat tartalmazó lexikon valamint a csak szóosztályokra vonatkozó szintaktikai szabályok között. Ez a rugalmasság az implementáció síkján is megvalósul, mert a NooJ lexikai komponense ugyanúgy véges állapotú transzdúszerekből áll, mint a szintaktikai rész, sőt a lexikai redundancia kezelésére a lexikonban is szabályokat alkalmazunk.

A lokális grammatikák végesállapotú transzdúszerek, amelyeket lépcsőzetesen alkalmazunk (*cascaded finite state transducers*) oly módon, hogy először a legszűkebb hatókörű (azaz a legtöbb egyedi lexikai elemet tartalmazó) szabályokat futtatjuk, majd az egyedi lexikai elemeket fokozatosan szóosztályok váltják fel, míg végül a legáltalánosabb, tehát csak szóosztályokra vagy szintaktikai csoportokra hivatkozó szabályok következnek. Ez az eljárás biztosítja azt, hogy a lehető leghamarabb megtaláljuk az adektvát fordítási megfelelőt. Az egész eljárás robosztusságát az adja, hogy a grammatikák lexikai és egyéb megkötéseit végsőkig feloldva eljutunk a szó szerinti megfeleltetéshez. Vagyis, ha semmi egyéb lokális grammatika nem volt illeszthető egy kifejezésre, végső soron megkapjuk az alkotóelemeknek a szótárban felsorolt célnyelvi megfelelőit. Ha több van belőlük, akkor mindegyiket.

2.2. Részleges fordítás

A harmadik rokon vonás a Metamorpho rendszerrel, hogy a lokális grammatikák kimenetét a célnyelvi megfelelők adják, azaz a mintaillesztéssel egyben előáll a hozzá tartozó célnyelvi megfelelő. Naivitás azzal áltatnunk magunkat, hogy egyenes út vezet a lépcsőzetes eljárás első szintjein előállt célnyelvi megfelelőktől a koherens célnyelvi mondat megformálásáig. Természetesen további kutatások szükségesek annak az eljárásnak a kidolgozására, amelynek során a sok-sok egyedi megfelelés egyetlen mondattá áll össze.

Bár Gross a lokális grammatikát végső soron az egész mondat struktúrájának leírására alkalmasnak tartotta, mi szerényebb célt követünk. Először a felszini szerkezetből szótár és minta-alapú eljárások segítségével megragadható mondatrészeket igyekszünk előállítani. Ez az eljárás elveiben hasonlít a részleges felszíni elemzéshez (chunking)[1][2], de nem áll meg az elemi szerkezeteknél, hanem a végesállapotú eljárást a maximális kiterjesztésű főnévi csoportokig viszi. A dolgozat egyik központi tétele, hogy a mondatban a maximális kiterjesztésű főnévi csoportok gépi fordítása elérhető célt jelent, amely kétnyelvi gyakorlati alkalmazásokban (pl. információkinyerés) önmagában is gyakorlati értékkel bír. A főnévi csoportok sikeres gépi fordításának elvi esélyét egy szintaktikai és egy szemantikai tényező is növeli.

Szintaktikailag kedvező az a tény, hogy a főnévi csoporton belüli szórend még a magyar esetében is kötött, ami azt jelenti, hogy a főnévi csoport végesállapotú

technológiával jól kezelhető. Ennek a szempontnak az érvényesülését csorbítják magyarban az igeneves szerkezetek, amelyek szinte teljesen nyitottá teszik a módosító szerkezetet. Angolban a prepozíciós szerkezetek főnévi csoporthoz sorolásának [*PP attachement* jól ismert problémája okoz gondokat. Mindezekkel együtt döntően lokális függőségekkel kell számolnunk.

A szemantikai érvet talán helyesebb inkább hipotézisként megfogalmaznunk. E szerint a maximális kiterjesztésű főnévi csoportok olyan szematikailag funkcionális (referáló) egységek, amelyek „megőrződnek" a fordítás során. Jogos az elvárás, hogy az alábbi mondatban

> [A biztonsági szolgálat emberei] [pontban hatkor] megnyitották [a Magyarországra települt legújabb áruházlánc első üzletének a kapujait].

a szögletes zárójelek között szereplő részek pontos, tényszerű fordítása tartalmazza ezek önálló célnyelvi megfelelőit a célnyelvi mondatban. Azt nyilvánvalóan nem várhatjuk el, hogy ezek belső szerkezete is azonos legyen (pl. az igeneves módosító szerkezetnek valószínűleg vonatkozó mellékmondat felelne meg az angolban fordításban) de azt igen, hogy az egész egységnek legyen hasonló szintaktikai státuszú megfelelője. Jegyezzük meg, hogy itt most nem a „megnyitotta kapuját" kifejezésről van szó, amely többszavas lexikai egységként funkcionál, és a kifejezés *egészének* van fordítási megfelelője, ami viszont nem feltétlenül tartalmazza a „kapu" szó megfelelőjét (pl. lehet „launched its operations, set up shop, started trading" stb.)

Végezetül felmerül a kérdés a vállalkozás motivációjáról: nem eleve nagyfokú naivitás egy ilyen komplex műveletet mint amilyen a gépi fordítás, egyetlen szoftver eszközzel megkísérelni? Nem hamis illúziókat keltünk, amikor a NooJ-t mint gépi fordítórendszert próbáljuk beállítani? Úgy vélem, hogy bár ezek a kérdések sok tekintetben jogosak, nyomós érvek szolgálnak a NooJ gépi fordításra való alkalmazása mellett. Először is, amint a 3. részben látni fogjuk, a NooJ igen komplex rendszer, ami ugyanakkor ingyenesen elérhető és sokoldalúan fejleszthető: valójában tetszőleges nyelvre akár teljesen az alapokról a kívánt nyelvi kódokkal felépíthető egy működő nyelvelemző rendszer. A NooJ tehát készen szállítja a nyelvészeknek a szükséges szoftvertechnológiai infrastruktúrát. Amint a fentiekben rámutattunk, a vele kifejlesztett nyelvelemző alkalmazások, esetünkben a részleges gépi fordítás, akár rapid alkalmazásfejlesztő munkaeszközként, akár önálló alkalmazásként hasznos tud lenni.

3. A NooJ eszköztára gépi fordításra

Ebben a részben összefoglaljuk a NooJ rendszer azon új funkcióit, amelyek lehetővé teszik a gépi fordításra történő alkalmazását. Mint ismeretes, a NooJ elődje, az INTEX rendszer, teljesen egynyelvű alkalmazás volt, egyszerre csak egy nyelvvel lehetett dolgozni, amelyet a munkamenet elején meg kellett adni. A NooJ kezdettől fogva törekedett a többnyelvűség támogatására. Már eddig is nemcsak számos file formátumot, de nyelvet is kezelt a rendszer. Mégis valójában csak a legutóbbi időkben valósult meg a nyelvek közötti átjárhatóság. Ennek alapja a kétnyelvű lexikon.

3.1. A lexikon

Minden fordítórendszer legfontosabb eleme a kétnyelvű szótár. A NooJ valójában a létező szótárakat kapcsolja egybe. A forrásnyelvi szótárban az idegennyelvi megfelelőket a szemantikai jegyekhez hasonló módon adhatjuk meg. Az 1. ábra részletet mutat be egy magyar-angol szótárból: a címszó és a szófaj mellett a ragozási útmutató (`+FLX` jegy) valamint az angol megfelelő (`+en` jegy) található. Figyeljük meg, hogy a többjelentésű szavak minden külön jelentése önálló bejegyzést kap a szótárban. Ennek megfelelően a forrásnyelvi szöveg kezdetben több értelmezést kap. Ez a notáció azt is lehetővé teszi, hogy egyéb célnyelveken is megadjuk a megfelelőket, íly módon tetszőleges n-nyelvű szótárat készíthetünk.

Nagyon fontos elv, hogy egy szótári elem több szóból is állhat. Minden olyan többszavas kifejezést, amelynek minden eleme és azok sorrendje rögzített, a szótárban a legcélszerűbb szerepeltetni. A leghosszabb megfelelő (*longest match*) elve gondoskodik arról, hogy ilyen esetben a többszavas kifejezés illeszkedik először a szövegre. A fordító rendszer hatékonyságát és robosztusságát döntő mértékben megszabja a szótárban található többszavas kifejezések mennyisége. A többszavas kifejezések mellett a NooJ rendszer természetesen megtalálja az azokat alkotó egyedi szavakat is, így tehát a `házi feladat` kifejezés mellett a rendszer egyenként is elemzi a `házi` és a `feladat` szavakat is, ami adott esetben feleslegesen növeli a szöveg többértelműségét. Ha le akarjuk tiltani az ilyen típusú többes elemzéseket, használhatjuk a `+UNAMB` jegyet, ami letiltja ugyanennek a szónak vagy szókapcsolatnak egyéb értelmezéseit.

```
ernyő,N+FLX=3AD1+en=umbrella
autó,N+FLX=2A4+en=car
iskola,N+FLX=1A3+en=school
kormány,N+FLX=1A5b+en=government
kormány,N+FLX=1A5b+en=steering wheel
házi feladat,N+FLX=1A7b+en=homework+UNAMB
```

1. ábra. Az idegennyelvi megfelelő megadása a NooJ szótárban

Két szótár összekapcsolása természetesen legfeljebb egy szótárprogram számára elégséges. Egy fordítóprogram számára minimális követelmény, hogy mindkét szótár segítségével szóalakok morfológiai elemzése és generálása is lehetséges legyen. A morfológiai elemzés mindig is része volt az INTEX/NooJ rendszernek. Több módon is megoldható volt, kezdetben a szóalakok elemzésükkel együtt szerepeltek a szótárban, amelyet egy szótőlista és paradigmák segítségével maga a rendszer állított elő. Ez az út nem adott lehetőséget az alakok generálására. A NooJ újabb változatai inflexiós vagy derivációs file-okat használnak, amelyek reguláris kifejezésekkel definiált transzdúszerek. Ez a megoldás nemcsak elemezni, hanem generálni is képes alakokat. Az 5. ábrán láthatjuk azt is, hogyan képes a célnyelvi megfelelők helyes alakjait előállítani.

3.2. Lokális grammatika

A NooJ rendszer egyik vonzó jegye a véges állapotú transzdúszerek könnyű kezelhetősége, amelyet egy jól kezelhető grafikai felület biztosít. E rövid áttekintésben a lokális grammatikák NooJ-beli implementációjának olyan új jegyeit említem, amelyek különösen hasznosak a gépi fordítás számára.

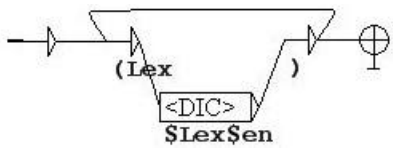
- változók használata
- hivatkozás lexikai jegyekre
- jegyek hozzárendelése az annotált egységekhez
- lexikai jegyek örökítése
- lexikai megkötések
- komplex változók használata lexikai megkötésekben

A változók használata elengedhetetlen ahhoz, hogy a lokális grammatika ne csak teljes egészében a bennük meghatározott egyedi szavakra legyen érvényes. A lexikai jegyekre való hivatkozás teremti meg a lehetőségét, hogy a transzdúszer kimenetében a szótári egység jegyeként számontartott célnyelvi megfelelőt alkalmazzuk, ráadásul a megfelelő toldalékolt alakban. Mivel a fordítási megfelelések tipikusan nem szavak hanem szintagmák között állnak fel, fontos, hogy a szintagmát is jegyekkel láthassuk el valamint, hogy a fej jegyeit is be tudjuk emelni a szintagma jegyei közé.

Terjedelmi korlátok miatt nem térhetünk ki az egyes funkciók részletesebb ismertetésére, a 4. részben bemutatott gráfok illusztrálják használatukat.

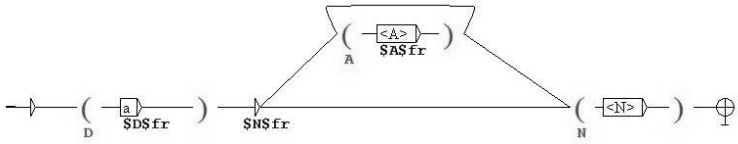
4. Előzetes eredmények

Miután áttekintettük az eszköztárat, tekintsük az alapesetet, hogy miként lehet a NooJ rendszerben egy direkt fordítórendszert megvalósítani?



2. ábra. Szó szerinti fordítás angolra

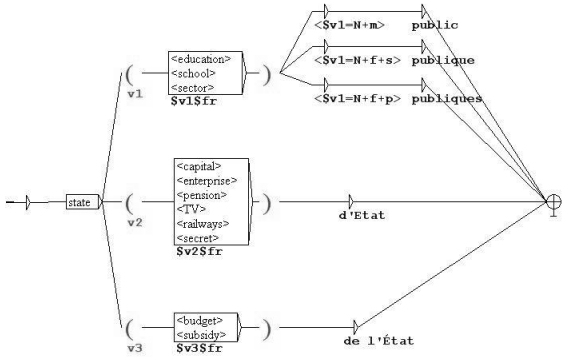
A 3. ábrán látható transzdúszer a Lex változóban tárolt tetszőleges szótári egység (`<DIC>`) angol célnyelvi megfelelőjét adja kimenetként, melyre a `$LEX$en` komplex változó alakjában hivatkozunk (feltételezve az 1. ábrán található szótári kódolást). Erre a megoldásra természetesen csak végső soron, minden egyéb lehetőség kimerítése után hagyatkozunk. Jellemzőbb azonban a szintaktikai szerkezet függvényében meghatározni a fordítási megfelelőt. Noha a 3 ábrán taláható minta akár pontos angol megfelelőjét tudná nyújtani *„a magyar állami oktatás”*



3. ábra. Egy egyszerű magyar főnévi csoport fordítása franciára

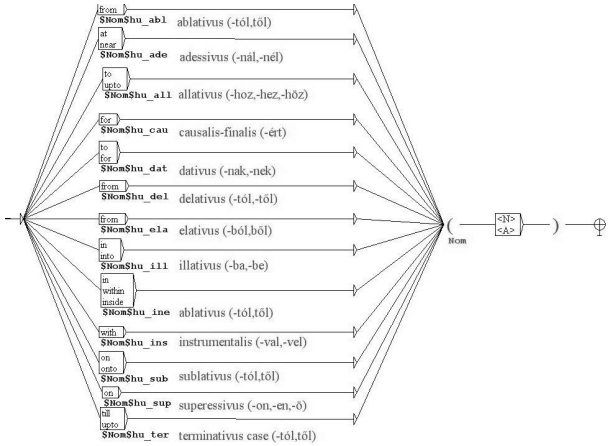
kifejezésnek, ugyanez franciára az eltérő sorrend miatt már csak a 3 ábrán látható lokális grammatikával leheetséges.

Ez a megfeleltetés is hamar túl elnagyoltnak bizonyul több okból. Nem tartalmaz például semmilyen finomabb disztribúciós megkötést a jelzők sorrendjére vonatkozóan. Ha ez mindkét nyelvben azonos, akkor természetesen nem is kell vele foglalkoznunk, hiszen egy fordítórendszer nem önmagáért való célként elemzi a forrásnyelvet, hanem csak olyan szintig, ameddig az releváns a két nyelv eltérései szempontjából. A 3 azért is fogyatékos volt, mert nem tartalmazta a franciában létező nembeli egyeztetést sem. Ennek megvalósítására a 4 ábra jobb fölső sarkában találunk példát. Az ábra nagyobb része a *„state”* angol szó francia megfelelőit adja három listával definiált kontextus függvényében. Ebben az eggyszerű esetben a szótárban is listába vehettük volna a kéttagú kifejezéseket. Könnyen elképzelhetünk azonban olyan gráfot, ahol általánosabb, szintaktikai `<N>` vagy szemantikai `<+bodypart>` jegyekkel határozzuk meg az odaillő lexikai elemeket.



4. ábra. Lexikai és morfológiai szelekciós megszorítások

Magyar-angol, magyar francia vonatkozásában gyakori eset, hogy egy szótári egységnek, tipikusan prepoziciónak, egy kötött morféma, azaz esetrag felel meg. Ilyenkor a szótári megfeleltetés nehézkes, és nehéz általánosan érvényes grammatikát találni. Az 5 ábrán található durva megoldást is csak mindegy egyéb lehetőség kimerítése után alkalmazzuk.



5. ábra. Angol prepozició magyar esetrag megfelelés alapesetei

Az eddigi ábrákon bemutatott esetek túl általánosak és elnagyoltak voltak, inkább csak a NooJ lehetőségeit illusztráló példaként szolgáltak. A lokális grammatikák alkalmazásának legfontosabb terepe ott van, ahol a minták részben lexikalizáltak, részben nyitott lexikai osztályokat tartalmaznak, és ezért szótári listázásuk vagy nem célszerű vagy lehetetlen is. Ilyen lehet a névkifejezések zöme (már ahol egyáltalán szükség van bármilyen fordításra) különösen például a dátumok, földrajzi helymeghatározások, időpont kifejezések.

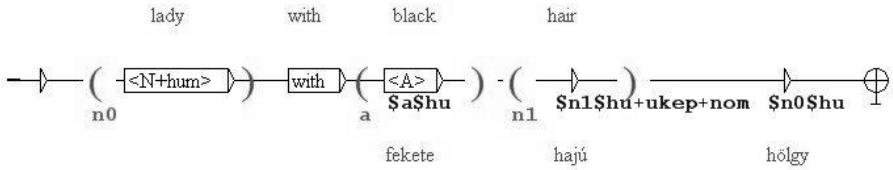
Az 1 táblázat csak utalásszerűen tartalmaz további olyan megfeleléseket magyar és angol főnévi szerkezetek között, amelyekben tipikusan lexikai szelekciós szabályok uralkodnak és a NooJ rendszer fent ismertetett eszközeivel jól kezelhetők.

`N with A N`	*girl with black hair*	`A N-Ú N`	*fekete hajú lány*
`A-speaking N`	*Spanish-speaking students*	`A nyelvű N`	*spanyol nyelvű diákok*
`N of N`	*freedom of assembly*	`A N`	*gyülekezési szabadság*
`N (Adv) Adv`	*house immediately opposite*	`A N`	*közvetlen szemközti ház*
`N P N`	*people at the reception*	`A N`	*a fogadáson lévő emberek*

1. táblázat. lokális szerkezeti megfelelések magyar-angol főnévi csoportokban

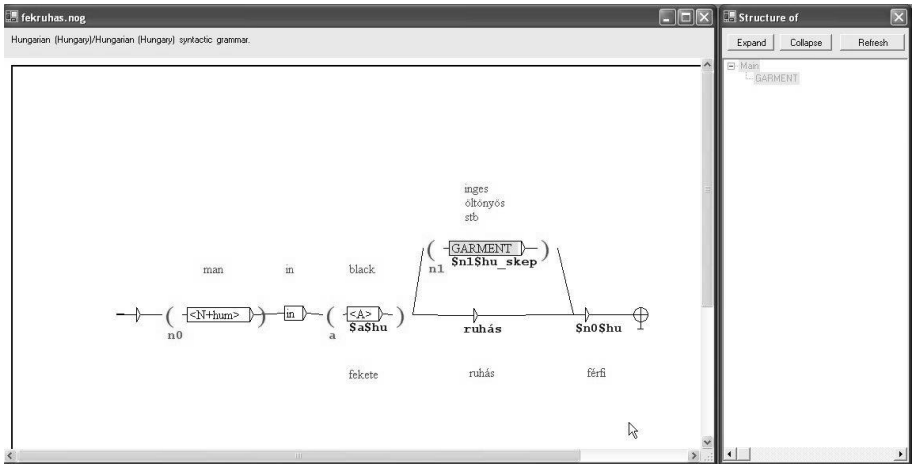
Példaként tekintsük a táblázat első sorában látható megfelelést. A 6. ábrán található az a vázlatos lokális grammatika, amely megvalósítja az angol kifejezés magyarra fordítását. Könnyedén meg tudjuk oldani a változókban tárolt angol lexikai egységek magyar megfelelőinek, beleértve az *-Ú* képzős alaknak az előállítását és helyes sorrendbe rendezését. Figyeljük meg, hogy az egyszerűség kedvéért csak azzal az egy szemantikai megkötéssel éltünk, hogy a *n0* elem `+hum` jeggyel rendelkezzen, az *n1* elemre nézve nem írtunk elő semmilyen megkötést. Márpedig ez helytelen alakhoz vezet akkor, ha *n1* nem testrészt vagy egyéb elidegeníthetetlen birtokot illetve inherens tulajdonságot jelent. Ez esetben ugyanis

az *-Ú* képző helyett a *-Vs* képzőt kell alkalmaznunk (*man with a black umbrella*:*fekete esernyős férfi*).



6. ábra. Angol-magyar módosító szerkezetek

Szemantikai megszorításokat többféle módon tehetünk: a lexikai elemek felsorolásával, beágyazott gráfok illetve szemantikai jegyek segítségével. A 7. ábrán azt mutatjuk be, hogyan lehet beágyazott gráfokkal szemantikai alosztályokat képezni. A GARMENT nevű algráf jelen esetben nem más, mint a ruhadarabok diszjunktív halmaza, amelyet ebben a szélsősegesen egyszerű esetben természetesen a főgráfban is megadhattunk volna. De még ebben az esetben is célszerű volt így eljárnunk, mivel az algráfok újrafelhasználása útján nagyobb modularitást érhetünk el, az algráf nevének alkalmas megválasztásával pedig a főgráf áttekinthetőségét és a rendszer karbantarthatóságát növelhetjük. Nem is beszélve arról az esetről, amikor a beágyazott gráf szerkezete lényegesen bonyolultabb, melyet érdemes is rejtve tartani a magasabb szintek elől.



7. ábra. Megfelelések szemantikai osztályokkal kifejezett megkötésekkel

5. Összegzés és további feladatok

A NooJ nyelvi fejlesztő rendszer mára kialakult eszköztára komoly szoftertechnológiai segítséget ad komplex nyelvi elemző rendszerek kifejlesztéséhez. Ezek között reális vállalkozásnak tűnik a részleges gépi fordítás megvalósítása. A maximális kiterjesztésű főnévi csoport több szempontból alkalmas célpont a fordításhoz. Az igeneves illetve prepozíciós módosító szerkezetektől eltekintve belső szerkezetében döntően lokális függőségi viszonyok dominálnak, amelyek véges állapotú grammatikával jól kezelhetőek. Nem idiomatikus, nem metaforikus stb. használatban szemantikailag is olyan önálló egységet képviselnek, amelyek belső szerkezetük mégoly eltérő volta ellenére is várhatóan megfelelésben állnak egymással. Kétnyelvi alkalmazásokban elengedhetetlenül fontos például a névkifejezések, dátumok, időkifejezések stb. fordítása. Egyéb gyakorlati alkalmazásként megemlíthetjük például a nyelvoktatást, ahol bizonyos nyelvi jelenségek gyakoroltatásához szintén hasznos lehet egy ilyen részlegesen fordító rendszer.

A részleges fordítás célként tételezése nem feltétenül jelenti azt, hogy a vázolt eljárást inherensen alkalmatlannak tartanánk a mondat teljes szerkezetének a megragadására. Ezt a célt inkább egy kutatási program első állomásának tekintük. A további feladatok a mondat szintaktikai vázának megragadására irányulnak, melyekben az ige és vonzatkeretének az integrálása kapja a központi szerepet.

Hivatkozások

1. Steven Abney: Partial parsing via finite-state cascades. 2. évf. (1996) 4. sz., *Journal of Natural Language Engineering*, 337–344. p.
2. Steven P. Adney: Parsing by chunks. In Carol Tenny (szerk.): *The MIT Parsing Volume, 1988-89*. http://www.vinartus.net/spa/89d.pdf, 1989, MIT Press.
3. Maurice Gross: The construction of local grammars. In E. Roche – Y. Schabes (szerk.): *Fininte State Language Processing*. Cambridge, Mass., 1997, The MIT Press, 329–352. p.
4. Zellig S. Harris: *Papers on Syntax*. Synthese Language Library sorozat, 14. köt. Dordrecht:Holland, 1981, D. Reidel Publishing Co.
5. Svetla Koeva – Denis Maurel – Max Silberztein (szerk.). *Nooj pour la Linguistique et le Traitement Automatique des Langues* (konferenciaanyag). Presses Universitaires de Franche-Comté, 2006.
6. Gábor Prószéky: Machine translation and the rule-to-rule hypothesis. In Krisztina Károly – Ágota Fóris (szerk.): *New Trends in Translation Studies. In Honour of Kinga Klaudy*. Budapest, 2005, Akadémiai Kiadó.
7. Gábor Prószéky – László Tihanyi: Metamorpho: A pattern-based machine translation project. In *24th Translating and the Computer Conference, 19-24* (konferenciaanyag). London, 2002, 19–24. p.
8. Max Silberztein: *NooJ Manual*. http://www.nooj4nlp.net/NooJ

VI. Beszédtechnológia, kommunikáció

Internetes beszédadatbázis a magyar mássalhangzó-kapcsolódások akusztikai szerkezetének bemutatására

Abari Kálmán[1], Olaszy Gábor[2]

[1] Debreceni Egyetem,
Pszichológia Intézet és Matematikai és Számítástudományi Doktori Iskola
abarik@delfin.unideb.hu
[2] MTA Nyelvtudományi Intézet olaszy@nytud.hu

Kivonat: A magyar hangtani kutatások bővelkednek leíró jellegű összefoglalásokban, azonban fonetikai-akusztikai szinten nagyon kevés munka található. A hangokból felépülő hangsor elemei közül talán legkevesebbet tudunk a mássalhangzók egymáshoz való kapcsolódásáról, az itt fellépő koartikulációs folyamatokról és azok akusztikai vetületéről. Az itt ismertetett interaktív, multimédiás beszédadatbázissal ezen a hiányon kívánunk segíteni. Az adatbázis közel 1200 féle két és több elemű mássalhangzó kapcsolatról szolgáltat adatokat (hullámforma, spektrogram, intenzitásgörbe, meghallgatás, belső mérési, összehasonlítási lehetőségek). Az adatbázis hátterében egy komplex kutatás húzódik meg, amelynek teljes anyaga könyv formájában hozzáférhető [4]. Ez a könyv szorosan kapcsolódik a jelen adatbázishoz, onnan veszi a példákat és viszont. A két anyag tehát együttesen használható jó hatásfokkal. Az adatbázis a http://fonetika.nytud.hu honlapon lesz hozzáférhető 2007-től.

1 Bevezetés

A beszéd hangsorépítésében a magánhangzók (V) és mássalhangzók (C) egyfajta váltakozása teszi lehetővé, hogy a percepció számára változatos, egymástól megkülönböztethető hangsorokat ejtsünk. A hangsorokban szereplő elemi alapegységek (fonémák) sorrendiségét a nyelv határozza meg. A fonémasor felépítési szabályaival a nyelvi szabályok szintjén a fonotaktika és a fonológia foglalkozik [7]. Az itt ismertetett adatbázisban alapvetően a kettő és több elemű magyar mássalhangzó kapcsolódások hangformáját adjuk közre, abból a célból, hogy az érdeklődő tanulmányozhassa a koartikulációból adódó akusztikai vetületek szerkezeti részleteit. Azt mutatjuk be, hogy a mássalhangzók torlódásakor azok hogyan hatnak egymásra, mely esetekben történnek változások a torlódásban résztvevő hangok szerkezetében. A beszédhangok kapcsolódásának lényeges eleme az úgynevezett átmeneti fázis, amelyik tulajdonképpen a hangkapcsolódási szakasz, a két hang összekapcsolásának akusztikai megvalósulása. Az átmeneti fázisban a hang(ok) akusztikai szerkezetet folyamatosan változik. A változás az artikulációs mozgások folyamatosságának következménye. A beszéd során minden artikulációs pozíciónak megvan a maga

akusztikai vetülete. Ha az artikulációs szervek elmozdulnak, akkor ez az akusztikai vetület is azonnal változni fog (spektrálisan más tartalom jelenik meg a hangban). Az átmeneti fázisra jellemző akusztikai tartalom változása széles skálán mozog. A változásban a hangjelenségekre alapvetően jellemző három fizikai tényező vesz részt: a frekvenciaszerkezet változása, az időszerkezet változása és végül a hangkapcsolatra jellemző intenzitásmenetben történő módosulás. E három elem valamelyike (vagy kombinációjuk) a legtöbb esetben változik a hangkapcsolatban. A legegyszerűbb hangkapcsolatnak tekinthető például, amikor ugyanazon hangok találkoznak egymással. Ilyenkor spektrális változás nincs, többnyire csak az időszerkezet módosul, a hangra jellemző időtartam megnyúlik (*r**áá**ll, legjo**bb** **b**arátom*). A hangátmenetekre jellemző akusztikai szerkezeti változások függnek a két hangra jellemző artikulációs képzési helytől és módtól is. A két hang közötti artikulációs mozgást a beszélőnek végre kell hajtani, ez időt vesz igénybe. Ha az artikulációs átmenet nem igényel bonyolult mozgássort, akkor rövidebb idő alatt jöhet létre az átmenet, ha igen, akkor hosszabb idő alatt. Ha a magánhangzók közötti átmeneteket vizsgáljuk, akkor például a nyelv, az ajkak, az állkapocs mozgása egymástól független és folyamatos. Mindhárom is változhat akadálytalanul, ezért az átmenet képzése sima (*k**iá**ltás, b**ei**ndul*). A CV, VC kapcsolatoknál az átmenetre jellemző mozgássor bonyolultabb, mivel a mássalhangzókra jellemző képzési helyet és módot, valamint a két hang közötti esetleges gerjesztésváltást is meg kell valósítani, de itt sem lép fel különösebb akadályozó tényező. Belátható, hogy a legbonyolultabb helyzet a két- és többelemű mássalhangzó kapcsolódásoknál lép fel, amikor az egyik C képzési helyéből, módjából és gerjesztéséből kell a másik (esetleg a harmadik) kapcsolódó C képzésére jellemző artikulációs helyzetbe vezérelni a beszédszerveket (*fe**lspr**icceli, la**jstr**om*). Ez bizonyos esetekben a mássalhangzó kapcsolatra jellemző akusztikai módosulásokkal is jár. Célunk, hogy feltárjuk és bemutassuk ezeket a módosulásokat, és folyamatában vizsgáljuk a kettős, hármas, és négyes mássalhangzó kapcsolódásokban létrejövő artikulációs és akusztikai változásokat.

A hangsorokban alapvetően négyféle hangkapcsolódási forma jöhet létre: VV, CV, VC és CC. Ezek további kombinációi is lehetségesek [6]. Az egyes kapcsolódások részvételi aránya nyelvfüggő, a magyarban a leggyakoribb hangsorépítő elem a CV és VC kapcsolat [3] a VV és CC elemek részvétele kisebb. Ennek ellenére a vizsgálatuk fontos, hiszen szerves részei a hangsorépítésnek. A hármas és négyes mássalhangzó kapcsolódások előfordulása még ritkább. Ha azonban teljessé kívánjuk tenni a mássalhangzó kapcsolódások akusztikai jellemzésének leírását, akkor ezek vizsgálatát is el kell végezni.

2 Az adatbázis nyelvi anyaga

A mássalhangzó kapcsolódásokat szó szintű lexikai egységeken mutatjuk be hangos formában női és férfi hangon. A szavak kettős, hármas, négyes és ötelemű mássalhangzó kapcsolatokat tartalmaznak. A mintaszavakkal mintegy 1240-féle mássalhangzó kapcsolatot mutatunk be, minden mássalhangzó kapcsolat, jellemzően egy adott mintaszóban szerepel, kizárólag szó belseji helyzetben. Az adatbázis minden CC kapcsolatot bemutat (több mint 600 féle). A CC kapcsolatok vonatkozásában teljességre törekedtünk, mivel ezek képezik a mássalhangzó kapcsolódások alapelemeit. A CCC kapcsolatokból a magyarban leggyakoribban előfordulók szerepelnek

az adatbázisban (mintegy 550 kapcsolat). A CCC kapcsolatokban gyakran tükröződnek a belőlük levezethető CC kapcsolatok akusztikai tulajdonságai. Mindezeken felül közel száz CCCC kapcsolat és ötféle CCCCC is lekérhető az adatbázisból. A komplett vizsgálati anyag összeállítása két lépcsős munka volt a következő fázisokkal. Első fázis: statisztikai alapú gyűjtéssel olyan szavakat kerestünk, amelyekben előfordul az adott mássalhangzó kapcsolat. Kiválasztottuk az adott mintaszót, azaz minden vizsgálandó kapcsolatra egyetlen szót jelöltünk ki (ezt mintaszónak nevezzük). Összeállítottuk a felolvasandó mintaszavak listáját a hangfelvételhez, elkészítettük a hangfelvételeket. A digitálisan tárolt hanghullámon bejelöltük a hanghatárokat (manuális címkézés). A második lépcsőben történt az anyag előkészítése az adatbázisban való bemutatásra. A hanghatár-címkéket vizuális és auditív módon kontroll ellenőrzésnek vetettük alá. Átírtuk a mintaszavak szöveges formáját hangszimbólumokra. A beszédadatbázis tehát minden mintaszóra a következő adatokat tartalmazza: a szó szöveges alakja és a szó hangalakja, a szó hullámformája meghallgatáshoz, a hanghatárok címkéi, hogy megjeleníthetők legyenek, a hangok időtartama és az akusztikai regisztrátumok. A beszédadatbázisban a mintaszavak kiejtésére átlagosan 10,5 hang/s-os artikulációs sebesség a jellemző.

2.1 Hangjelölések

Az adatbázisban a beszédhangok jelölésére saját jelrendszert használunk. Olyan hangjelölési formát dolgoztunk ki, ami géppel kezelhető (1 hang = 1 karakter), továbbá viszonylag könnyen olvasható (kivétel a dz és dZ hangjel). A számítógépes hangszimbólumok és a betűjelek közötti megfeleltetéseket a 1. táblázat mutatja (csak azokat a jeleket mutatjuk be, amelyek eltérnek a betűképtől, a hosszú hangokat a kettőspont hozzáadásával jelöljük).

1. Táblázat: a számítógépes hangjelek a betűkép szerint

magánhanzók	betűkép	hangjel	mintaszó és hangjeles átírása
	á	A:	át = A:t
	ü	U	üt = Ut
	ö	O	öt = Ot
	é	E:	én = E:n
mássalhangzók			
	gy	G	gyár = GA:r
	ty	T	tyúk = Tu:k
	ny	N	nyár = NA:r
	zs	Z	zseb = Zeb
	s	S	sok = Sok
	sz	s	szék = se:k
	cs	C	csak = Cak

3 A hanghatárok kijelölése

A két és több elemű mássalhangzó kapcsolatok vizsgálatában a hanghatárok megállapítása az egyik legkritikusabb feladat, hiszen erre épülnek a hangidőtartam adatok, valamint a spektrális kapcsolódási jellemzők mérései is. A számítógépes technika

sokat segít a hangelhatárolás egyre pontosabb megállapításában, de az alapvető nehézségek ugyanúgy fennállnak, mint régen. A hanghatár megállapítása szempontjából három hangkategóriát kell megkülönböztetnünk: azokat, amelyeknél a hanghatárt a rezgésképből szinte egyértelműen ki lehet jelölni; azokat, amelyeknél ez nehezebb, itt bonyolultabb vizsgálatokat kell végezni; és végül azokat, amelyeknél nincs egyértelmű hanghatár a két hang között, a határpont helye minden esetben jórészt a mérést végző személy döntésétől függ. A hanghatárok kijelölésénél alapvetően a hangok képzési módjából adódó szerkezeti komponensekből indulunk ki (például a zár-rés hangoknál keressük a zárszakaszt és a réselemet, mint szekvenciálisan következő hangelemet). Mint látni fogjuk a későbbi példákban a mássalhangzó kapcsolatoknál vannak olyan szerkezeti változások a kapcsolódó hangoknál, hogy az egymásra hatások következtében néha nem lehet egyértelműen eldönteni a hanghatár helyét.

Egyértelmű hanghatár-kijelölés

A hanghatár viszonylag pontosan kijelölhető a hangsor azon pontjain, ahol az egyes hangokat alapvetően a gerjesztési jel megváltozása különíti el. Ilyenkor akár gépileg, akár vizuális ítélet alapján már meghatározható a két hang határpontja (*a**lsz**ik, át**n**éz*).

Hanghatár-kijelölés több paraméter alapján

Azoknál a hangkapcsolatoknál, amelyeknél nincs gerjesztésváltás a két hang kapcsolódási pontján, a hanghatár kijelölése nehezebb (*re**pc**e, a**lm**a*). Az időfüggvény vizsgálatán túl figyelni kell az intenzitásváltozást, amely jelzésként szolgálhat a hanghatár pontos megállapításához (pontosság alatt értjük a maximum 10-15 ms-os hibahatárt). Ezen túlmenően az auditív meghallgatáshoz célszerű biztosítani egy úgynevezett lépésenként (periódus szintű időosztásos) tágítható kapuzásos megszólaltatást is. Ezzel akár jobbról balra, akár fordítva fokozatosan hallgathatjuk a hullámforma hangját és meghatározhatjuk a mért mássalhangzók csatlakozási pontját. Ilyen meghallgatási lehetőséget biztosít a Profivox beszédszintetizáló rendszer vizsgálati szoftvere a Profidev [5], amelyet a hanghatárok kijelölésénél fel is használtunk.

Nincs egyértelmű hanghatár

Azoknál a hangkapcsolatoknál, ahol a gerjesztés nem változik, az artikulációs mozgások időben lassan változnak, a hanghatár még a dinamikus hangspektrogram felrajzolásával is csak közelítőleg határozható meg. Ilyenek például az [m]+[n] kombinációja (*ro**mn**ak*). Ezekben az esetekben a kutató egyéni döntésétől függ, hogy melyik pontot nevezi ki hanghatárnak. Ebből következik, hogy az ilyen hangkapcsolatoknál megállapított hangidőtartam-adatok nagyobb szórást mutathatnak, mint a korábbi kategóriákéi.

4 Az adatbázis szerkezeti felépítése

A magyar mássalhangzó kapcsolódások akusztikai szerkezetét bemutató adatbázist a kezdetektől webes alkalmazásként képzeltük el, majd fejlesztettük ki, hiszen ez a forma biztosítja a tárolt információ lehető legszélesebb körű felhasználását. Egyszerű

kezelőfelületen, a keresőgépeknél megszokott módon kérdezhetjük le az adatbázist, a használathoz pedig csak egy böngészőprogram és Java plug-in szükséges.

A webes architektúra jelentősen meghatározta az adatbázis szerkezeti felépítését. Számításigényes algoritmusok megvalósítására itt nincs lehetőség, a megjelenítésre szánt információt az előkészítettség minél nagyobb fokán kell letárolni. Ennek megfelelően adatbázisunk a szervezetten tárolt adatokon túl, számos, időigényes előfeldolgozáson átesett állományt is magába foglal.

A gyors visszakeresés érdekében relációs adatbázisban tároljuk az 1240 db mintaszó szöveges formáját, a hangsorát, a hangsorban az egyes hangok időtartamadatait, illetve azt a konkrét mássalhangzó kapcsolatot, amelyre az illető mintaszó a példa.

Minden mintaszóhoz négy képet is tároltunk (összesen 4960 db -ot), melyek rendre a szó hullámformáját, spektogramját, intenzitásmenetét és spektogram+intenzitás képét jelenítik meg. Az akusztikai szerkezetek tanulmányozása mellet, a képek mindegyikéről leolvasható a szó hangsora és jelezzük a hanghatárokat is. A mintaszavak hangfelvételei és a manuálisan bevitt címkézési információk alapján a képek a Praat 4.0 fonetikai elemző segítségével készültek.

Tárolásra kerültek továbbá a szó meghallgatásához, valamint az interaktív hullámforma megjelenítéséhez (pl. lassított/gyorsított lejátszáshoz) szükséges egyéb állományok is.

5 Az adatbázis szolgáltatásai

A mássalhangzó kapcsolatokat alapvetően négy csoportban mutatjuk be: CC, CCC, CCCC és CCCCC kapcsolódások. A lekérdezésnek két módja van.

A) **Hang alapú** lekérdezés (bal ablak), amikor a keresett mássalhangzó kapcsolódás hangjait adjuk meg, valamint a négy vizsgálati csoport aktuális adathalmazát állítjuk be. A hangmegadáshoz segítségre van a hangtáblázat, amiből kiválaszthatjuk a kívánt hangokat. Például a bd hangkapcsolat megadásánál a CC vizsgálati csoportot kell beállítani, hiszen két mássalhangzóból áll a keresett kapcsolat.

B) **Betű alapú** lekérdezés (jobb ablak), amikor betűkapcsolatot adunk meg (ez állhat több betűből is és nincs korlátozva arra, hogy csak mássalhangzó kapcsolat adható meg). Ilyenkor nem kell beállítani a bal oldali ablakban vizsgálati csoportot, mivel a kereső minden szó betűképén végigmegy. Például az *anda* betűsorozat megadásakor minden olyan szót megkeres, amelyikben ezek a betűk ilyen sorrendben előfordulnak (ha talál ilyent), és azokat választja ki.

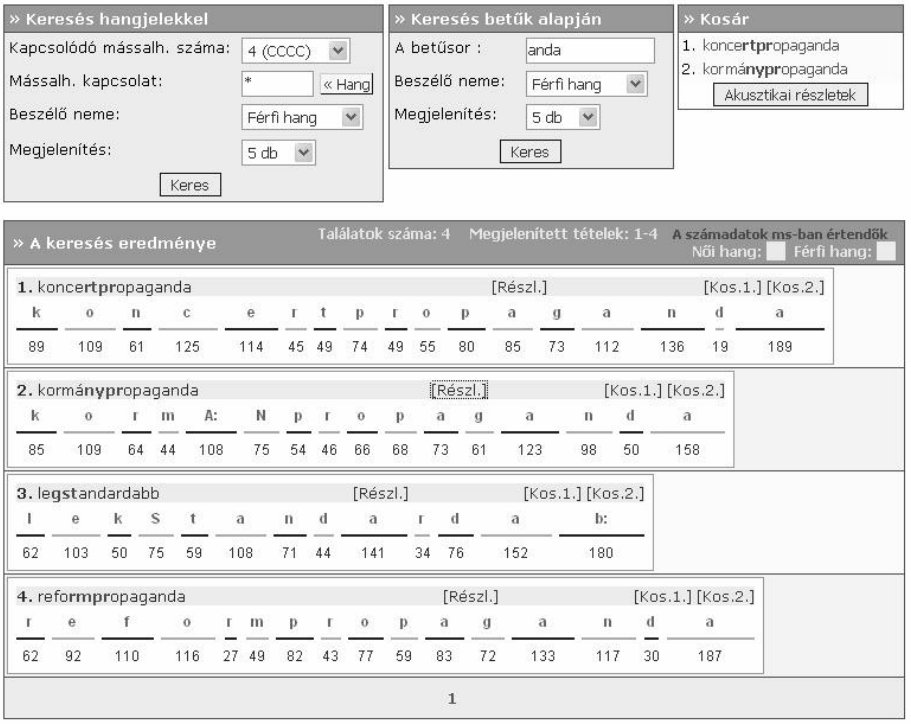


Fig. 1. Az *anda* betűsorozat lekérdezése a Keresés betűk alapján ablakból

A lekérdezés eredménye egy lista amely a megtalált szavakat sorszámozva tartalmazza. A lista egy-egy eleme a mintaszó szöveges és hangátírásos formája, valamint a hangok időtartama szám formájában a hangok alatt (a szám ms értéket jelent). A találati lista minden szavával a továbbiakban a felhasználó rendelkezik. Két műveletet végezhet el: a részletek jelzésre kattintva egy másik ablakot jeleníthet meg a hullámforma részletekkel, illetve a jobb felső sarokban lévő kosárba teheti bele a szót további akusztikai vizsgálatra. Mindkét esetben új ablakokban kell tovább dolgozni.

Kosár: a kosárba tett szó akusztikai paramétereit megjeleníthetjük, ha az Akusztikai részletek-re kattintunk. Ekkor megjeleníthető a spektrális szerkezet (formánskép), az intenzitásmenet, egy rövid szöveges magyarázat a mintaszóban bemutatott mássalhangzó kapcsolat jellemzőiről. A mintaszót meg is lehet hallgatni. A kosár két mintaszó elhelyezésére van felkészítve, hogy közvetlen összehasonlításokat is lehessen tenni (egymás alatt jelennek meg a diagramok).

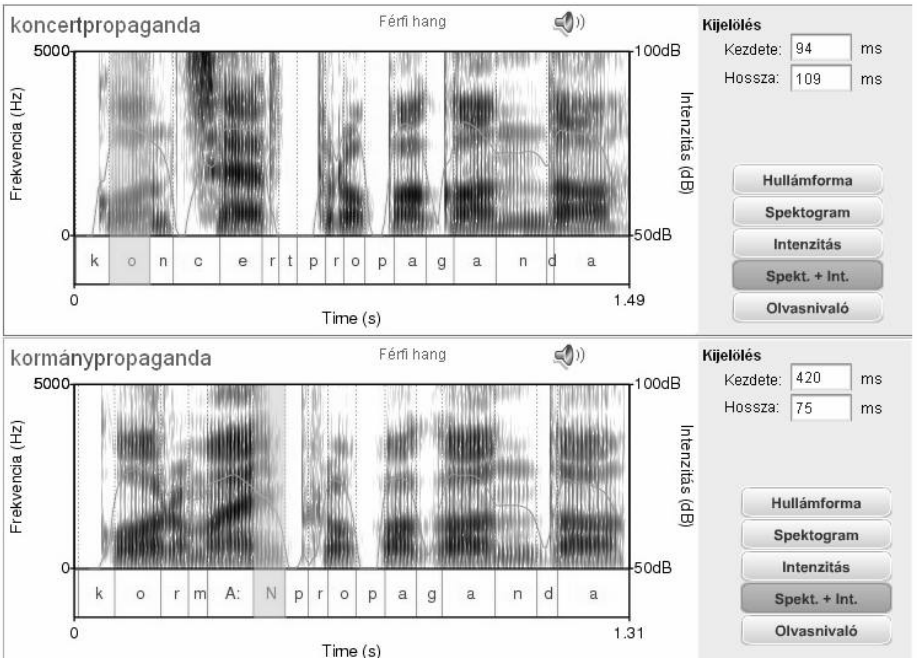


Fig. 2. A kosárba tett két mintaszó akusztikai részleteinek megjelenítése összehasonlításra

Interaktív hullámforma megjelenítés: ebben az ablakban a felhasználó interaktív módon is tanulmányozhatja a hullámforma részleteket kép (ablak tágítás) és hang formájában is. Időtartam méréseket is végezhet. Többféle hanglejátszási forma segíti az elemzést (lassított, gyorsított, egy periódus ismételve stb.).

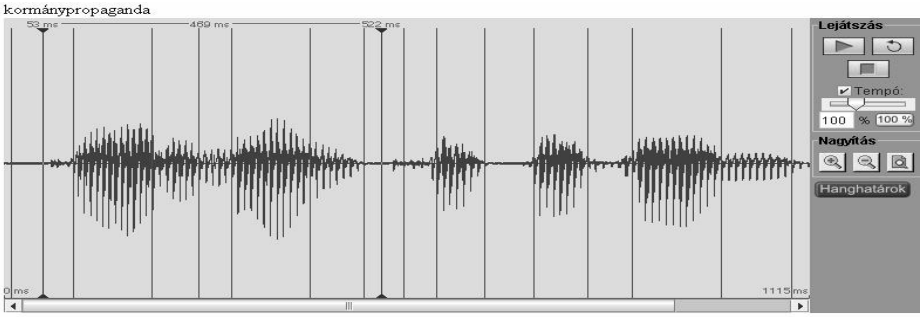


Fig. 3. Interaktív hullámforma megjelenítés a ’kormánypropaganda’ szóra

6 A mássalhangzó kapcsolódások szerkezetéről

A következőkben néhány speciális szerkezeti sajátosságot mutatunk be a mássalhangzó kapcsolódások méréseiből.

Bizonyos CC kapcsolatoknál a két csatlakozó hang képzési konfigurációjától, valamint a képzési módtól függően az összekapcsolási ponton svá-szerű hangelem jelenik meg. Ilyen például, amikor zöngés zárhangok találkoznak (*la**bd**a*). A zöngés zárhangok zárfelpattanása például egy CV kapcsolatban egy 10-15 ms-os felpattanási hangelem, amely beleolvad a magánhangzó kezdetébe, tehát semmi köze nincs egy svá-hoz. A svá a CC kapcsolat artikulációs mozgásainak az eredménye és keletkezése hangfüggő [1]. Anélkül, hogy belemennénk a svá képződés részleteibe egy speciális szerkezeti formát mutatunk be, amely akkor áll elő, amikor a mássalhangzó kapcsolat egyik eleme az [r]. Ez meghatározott esetben hanghatár-jelölési gondot is okoz, ezért választottuk ezt a példát (amikor az [r] zöngés zárhanghoz (*abrak*) kapcsolódik). Ilyenkor a hanghatárt a svá elem közepére jelöltük. Miért? A példánál maradva, a [b]+[r] kapcsolatban speciális helyzet áll elő, mivel az [r] indításához is szükség van egy svá elemre [1], [2] és a [b] zárfelpattanásának helyén is egy svá elem keletkezik a CC kapcsolatban. Elméletileg tehát két svá elemnek kell követnie egymást. Az ejtés során ezek eggyé olvadnak össze, ezért kellett a hanghatárt a svá belsejében elhelyezni

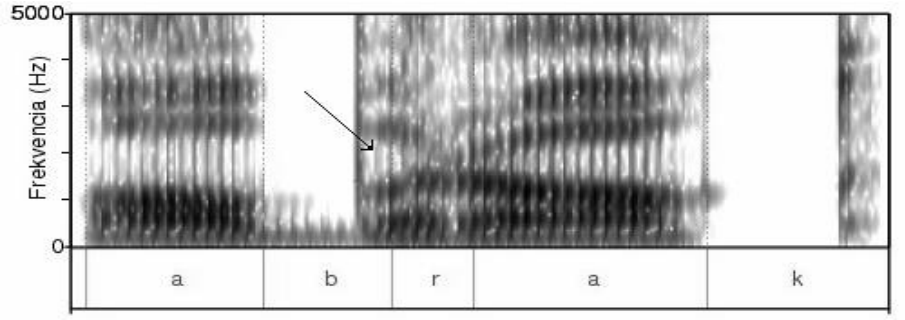


Fig. 4. Az összeolvadt két svá az *abrak* szóban. A hanghatárt a svá közepére kell jelölni.

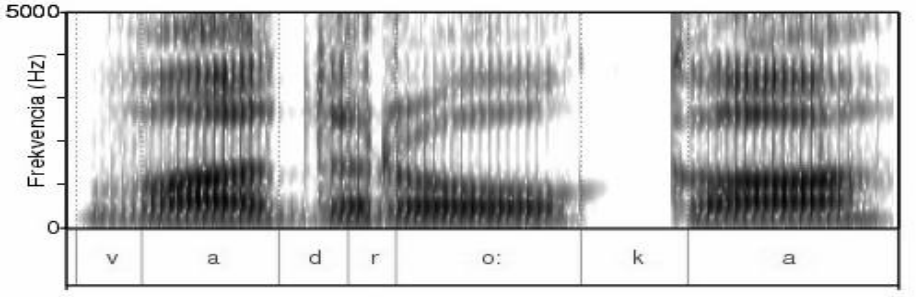


Fig. 5. Az összeolvadt két svá a *vadróka* szóban. A hanghatárt a svá közepére kell jelölni.

mintegy jelölve, hogy ez a hangrész megoszlik a két hang között (Fig. 4.). Hasonló hanghatárjelölést láthatunk a *vadróka* mintaszó [d]+[r] kapcsolatában is (Fig. 5.)

Egy másik példában a nazálisok okozta egyik furcsa szerkezeti változást mutatjuk be. Ha a nazális hang a mássalhangzó kapcsolat második eleme, az első hangja pedig zöngétlen réshang (*kismadár*), akkor az ilyen kapcsolatokban kimutatható egy 30-40 ms tartamú, gyakorlatilag néma fázisnak tekinthető hangelem, amely a hangkapcsolódási ponton jön létre a nazális mássalhangzó zöngés rezgésének beindulása előtt (Fig. 6.).

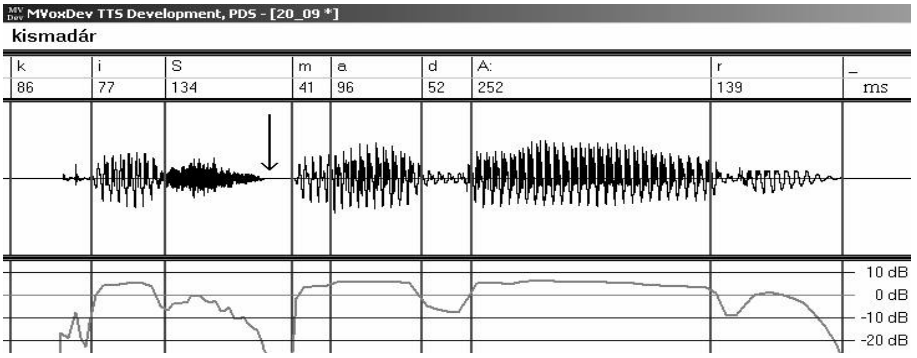


Fig. 6. A másodlagos néma fázis (nyíl) a réshang és a nazális között.

Ezt a szakaszt elneveztük **másodlagos néma fázis**nak. A másodlagos néma fázisok jelenléte a hangsorban több kérdést is felvet. Először is kérdéses, hogy ez a hangszakasz melyik hanghoz tartozik, hogyan kell a hanghatárt ilyen esetekben bejelölni? Másodszor, hogy a rezgésképek alapján úgy látható, mintha a hangsorban egy zár-rés hang tükörképe szerepelne (tükör affrikáta?). Még szokatlanabb hangszerkezeti kép alakulhat ki, ugyanezen hatás eredményeként a zár-rés hangoknál is (Fig. 7.). Ekkor akár két néma fázis is lehet a zár-rés hangban, azzal kezdődik és azzal is fejeződik be.

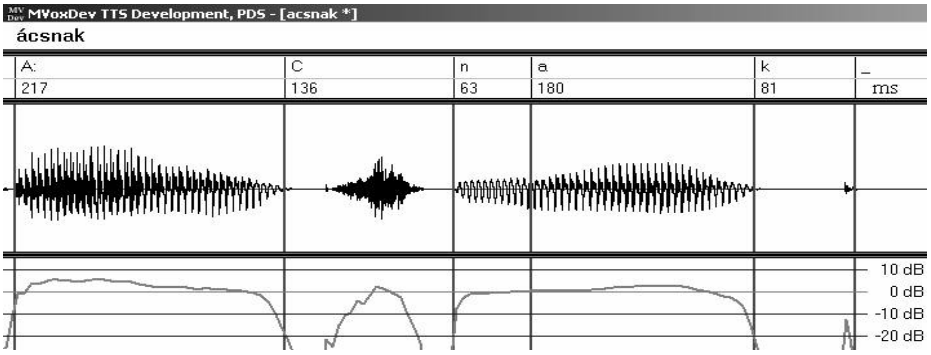


Fig. 7. A zár-rés hang néma fázissal kezdődik, utána következik a réselem, majd a másodlagos néma fázissal záródik

Mi a magyarázata a jelenségnek? A nazális hangok képzésénél az orrüreg nyitott állapotban van, a levegő ezen keresztül áramlik, a hangszalagok rezegnek. Az orrüreg nyitását/zárását lényegileg a nyelvcsap végzi. Az üregváltáshoz (a nyelvcsap

mozgatásához) idő kell. Ha a megelőző mássalhangzó zörejes gerjesztésű, akkor a gerjesztésváltáshoz is idő kell. A két időkomponens összeadódhat a zöngétlen mássalhangzók és a nazálisok találkozásánál és ez okozza a másodlagos néma fázis kialakulását. Ennek a hangelemnek a tanulmányozása még további kutatásokat igényel.

7 Összefoglalás

Bemutattuk a magyar mássalhangzó kapcsolódások, mint hangsorépítő elemek koartikulációs folyamatainak vizsgálatára készített interaktív, multimédiás beszédadatbázist amely Internetes alkalmazásként használható 2007-től a http://fonetika.nytud.hu honlapon. Az adatbázis mintegy 1240 féle mássalhangzó kapcsolódásra tartalmaz egy-egy mintaszót férfi és női ejtésben. A kért mássalhangzó kapcsolat mintaszava hangdefinícióval, illetve betűképpel kereshető az adatbázisban. Lekérdezhetők a szó akusztikus diagramjai, az adott mássalhangzó kapcsolat szerkezetének szöveges magyarázata. A mintaszavak hangban is megszólaltathatók. Az adatbázis hátterében egy komplex kutatás húzódik meg, amelynek teljes anyaga könyv formájában hozzáférhető [4]. Ez a könyv szorosan kapcsolódik a jelen adatbázishoz, onnan veszi a példákat és viszont. A két anyag tehát együttesen használható jó hatásfokkal.

Bibliográfia

1. Gósy Mária: A semleges magánhangzó nyelvi funkciói. In: Gósy M. (szerk.) Beszédkutatás 2006. MTA Nyelvtudományi Intézet 2006. 8-21.
2. Olaszy Gábor: A magyar beszéd leggyakoribb hangsorépítő elemeinek szerkezete és szintézise. Nyelvtudományi Értekezések 121. Akadémiai Kiadó 1985.
3. Olaszy G.: Model to predict Hungarian sound durations for continuous speech. Acta Linguistica Hungarica, Vol.49 (3-4), 2002. 321-345.
4. Olaszy Gábor: Mássalhangzó kapcsolódások a magyar beszédben. Tinta Kiadó (2007. megjelenés alatt)
5. Olaszy G., Németh G., Kiss G.: Hungarian audiovisual prosody composer and TTS development tool. In: Prosody 2000. Editors: Puppel Stanislaw, Grazina Demenko. Poznan, 2001. 167-178.
6. Szende Tamás: A beszédfolyamat alaptényezői. Akadémiai Kiadó (1976)
7. Siptár P., Törkenczy M.: The phonology of Hungarian. The Phonology of World's Languages. Oxford/New York, Oxford University Press. 2000.

Ezt a kutatást az OTKA TO4829 is támogatta.

Magyar kiejtési szótár az Interneten

Abari Kálmán[1], Olaszy Gábor[2],
Zainkó Csaba[3], és Kiss Géza[3]

[1] Debreceni Egyetem,
Pszichológia Intézet és Matematikai és Számítástudományi Doktori Iskola
abarik@delfin.unideb.hu
[2] Magyar Tudományos Akadémia, Nyelvtudományi Intézet
olaszy@nytud.hu
[3] Budapesti Műszaki Egyetem, Távközlési és Médiainformatikai Tanszék
{zainko, kgeza}@tmit.bme.hu

Kivonat: Internetes magyar kiejtési szótár mindezidáig nem készült magyar nyelvre. Az igény viszont világszerte nagy. Ezt a hiányt kívánjuk pótolni fejlesztésünkkel. Szótárunk megvalósításának terveit alapvetően az elektronikus lehetőségek maximális támogatására alapoztuk. Ez sok különbséget jelent egy hagyományos szótárhoz képest. Az egyik ilyen a szóállomány. Szótárunk nemcsak szótöveket tartalmaz, hanem azok ragozott, toldalékolt formáit is, mindösszesen 1,8 millió lexikai egységet. Ezért ebben a kiejtési szótárban a szótárelemeket **szóalak**nak hívjuk. A szótár minden lexikai eleme tehát egy-egy szóalak, amelynek a kiejtését nemzetközi fonetikai hangjelekkel (IPA) adjuk meg. Külön lexikai csoportot alkotnak a leggyakoribb magyar vezetéknevek, amelyeknek a kiejtését szintén megadjuk. A szótár különlegessége a hangos szótárrész, amelyből 60.000 szóalak hangban is meghallgatható. A szótár a http://fonetika.nytud.hu honlapon lesz hozzáférhető 2007-től.

1 Bevezetés

A kiejtési szótárak jól használhatók a kutatásban, az oktatásban (egyetemi oktatás, nyelvtanítás), gyakorlati alkalmazásokban és még számos területen. Az ilyen szótárak elektronikus, publikus formában való közreadása (Interneten) tovább tágítja a használat terét. Tudomásunk szerint magyar nyelvre ilyen nyilvánosan hozzáférhető nyelvtechnológiai adattár nem áll rendelkezésre az Interneten. Az igény viszont világszerte nagy. Ezt a hiányt kívánjuk pótolni.

Miért van szükség kiejtési szótárakra? Mert a nyelv írott és ejtett formája különbözik. A kettő közötti kapcsolat nyelvfüggő. Ahhoz, hogy egy nyelv szavait ki is tudjuk ejteni, tudnunk kell, hogy milyen hangsort kell megvalósítanunk az artikuláció során. Az úgynevezett fonetikusabb nyelveknél a kiejtés és az írás között szoros a kapcsolat, míg a kevésbé fonetikus nyelveknél nehezebb az írott alakból származtatni a kiejtést. A magyart közepesen fonetikus nyelvnek tarthatjuk. Ez azt jelenti, hogy az írásképnek megfelelő hangsorozat meghatározása nem túl bonyolult, bár vannak jócskán furcsa, nem várt kiejtési formák is.

Mit várunk el egy elektronikus kiejtési szótártól. Azt, hogy a keresett szó begépelése után a program adja meg annak kiejtését, lehetőleg nemzetközi fonetikai hangjelekkel. Így a szótár használatához nincs szükség speciális fonetikus szimbólumkészlet megismerésére, hanem ezeknek a szabványos, jól dokumentált jeleknek az ismerete elégséges, így még az idegen nyelvű, a témában járatos látogató számára is használható. Az általunk megvalósított kiejtési szótár mind szerkezeti felépítésében, mind szolgáltatásaiban lényeges különbségeket mutat egy hagyományos szótárral szemben. A magyarra 1992-ben adtak ki kiejtési szótárt [2]. Az ilyen szótárakban a szerzőknek nem alapvető célja, hogy a magyar szóállomány kiejtési formáit (a lehető legtöbb szóra) megadja, inkább a különleges kifejezéseket, az idegen szavak kiejtését teszi közzé. A szerző [2] így összegzi szótárának célkitűzését „A Magyar kiejtési szótár régi hiányt pótol a könyvpiacon. Segítséget nyújt a legkülönfélébb hasonulások, egyes idegen szavak és rövidítések helyes kiejtésében, sőt azon szavak esetében is, amelyeket éppen hogy úgy kell kiejteni, ahogyan írva vannak, de a mindennapi beszéd ettől eltérő, helytelen alakokat alkalmaz. A 10880 szót és szókapcsolatot tartalmazó szótár hasznos segítőtársa lesz mindazoknak, akik nem csupán írni, de beszélni is helyesen szeretnének magyarul”. Egy újabb szerző legújabb kiadású ilyen szótára [6] már 40.000 elemet tartalmaz.

Az Internetes kiejtési szótár kialakításánál maximálisan támaszkodtunk a számítástechnika, az Internet adta lehetőségekre. Ebből következik, hogy a szótár szóállományának kialakítása gyökeresen más elveken nyugszik, mint amilyeneket a fenti hagyományos szótárak alkalmaztak. Esetünkben nagy, elektronikusan rögzített szövegkorpusz képezi a szóállomány kialakításának az alapját. A leendő szótár elemeit automatikus módszerrel válogattuk ki a szövegkorpuszból. Ennek következménye, hogy nemcsak szótöveket tartalmaz a szótár, hanem azok ragozott, toldalékolt formáit is, vagyis ez a szótár ténylegesen tartalmazza a magyar lexémák nagy többségét, nem téve különbséget eredetük, jelentésük között. Ezeket a szótárelemeket **szóalak**nak hívjuk. A szótár minden lexikai eleme tehát egy-egy szóalak. Az általunk használt szóalak pontos definíciója a következő: **olyan betűkből álló lexikai egység egy szövegben, amelyiket nem betű karakterek határolnak** (zömmel szóközök). A betűkarakter sorozat betűtartalma minden egyes szóalaknál legalább egy betűkarakterrel eltér a szótár más szóalakjától. Belátható, hogy elegendően nagy szövegkorpusz esetén az ilyen szóalak-állomány jól lefedi a magyar nyelv leggyakrabban használt szavainak szóállományát, tehát szótárként használható. A szótár készítésének fontos szoftver eleme a hangátíró algoritmus, amelynek segítségével a szóalakokat átírjuk hangalaki formába (ez a kiejtés formája). A hangtani szabályok a magyarra a nyelvészeti szakirodalomban jól definiáltak, bár leginkább leíró formában hozzáférhetők [1]. A jelen szótár elkészítéséhez saját hangátíró algoritmust készítettünk. Ennek fő gerincét a szakirodalomban megtalálható kiejtési szabályok alkotják, ezeket építettük be. Emellett alkalmazunk kivétel listákat, amelyek azokat a kiejtési formákat írják le szóalakokhoz kapcsolva, amelyeket a szabályok nem tudnak lekezelni. Ezek a kivétel szótárrészek szabadon bővíthetők. Az elektronikus feldolgozás (beszédtechnológia) ma már lehetővé teszi, hogy hangos szótárrészt is készítsük. Ez segíti a felhasználót a tényleges kiejtés megértésében, a hangidőtartamok érzékelésében, a szó ritmusának elsajátításában. Ezt is kihasználtuk és szótárunkban a leggyakoribb szótárelemek meg is hallgathatók.

2 A szóalakok állománya

A szóalakok állományának meghatározása elektronikus formában történt, az Internetről automatikusan gyűjtött adatokból [7]. A gyűjtést újságok internetes kiadásaiból és elektronikus könyvtárak anyagaiból végeztük. Az adatgyűjtéshez azért választottuk az újságok internetes kiadásait és elektronikus könyvtárakat, mert tapasztalatunk szerint azok az átlagos webes oldalakhoz képest jóval gondosabban megszerkesztettek és átnézettek, helyesírásilag korrekt szövegeket tartalmaznak. A forrás ilyen jellegű szelektív megválasztása ellenére ezek az oldalak is tartalmaztak például idegen nyelvű részeket is, amelyek kiszűrését meg kellett oldani. A nem magyar szövegek detektálását saját fejlesztésű nyelvdetekciós szoftverrel végeztük. A nyelvdetekció elsősorban mondat szinten történt, a magyartól eltérő nyelvű mondatokat kiemeltük a szövegtárból. Azoknál a dokumentumoknál, ahol az idegen nyelvű mondatok voltak többségben, ott a teljes dokumentumot kihagytuk a szövegkorpuszból.

A 80 millió szót tartalmazó szövegkorpuszban 1,8 millió különböző szóalakot számoltunk meg, ezek alkotják a kiejtési szótár kereshető szöveges állományát. Ezek a szóalakok adott gyakorisággal fordulnak elő a nagy szövegkorpuszban. Ha ezt a gyakoriságot vesszük figyelembe, akkor előállíthatunk egy fedési diagramot, amelyik megmutatja, hogy a szóalakok a teljes szövegkorpusz hány százalékát fedik le (Fig 1. ábra).

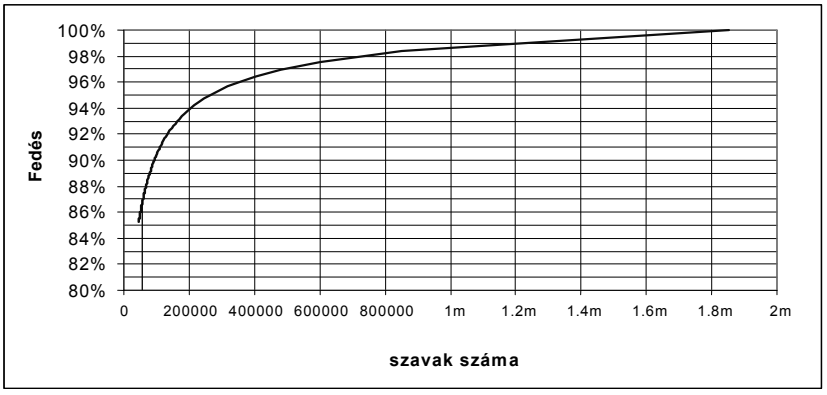


Fig. 1. Fedési diagram, maely megmutatja, hogy a szóalakok gyakoriság szerinti válogatása a teljes 80 millió szavas szövegkorpusz hány százalékát fedi le.

A szótár szóalakjaiból statisztikai gyűjtéseket végeztünk, amelynek két eredményét mutatjuk be, a szótagok szerinti eloszlást, valamint a hangok szerintit. Kiválogattuk a leggyakoribb szóalakokat a szótagszámuk szerint. Az eloszlást a Fig.2. ábra mutatja.

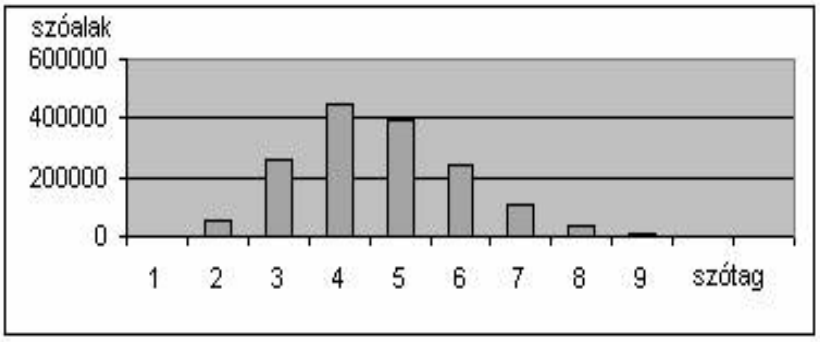


Fig. 2. A magyar elektronikus kiejtési szótár szóalakjainak eloszlása a szótagok száma szerint

Ezek szerint a magyarban a leggyakoribb szóalakok a 3, 4, 5 és 6 szótagú szavak. Megjegyezzük, hogy ez az eloszlás a szóalakokra vonatkozik, nem pedig a magyar szövegekben előforduló szavak általános eloszlására. Ez utóbbiról részletes adatokat [5]-ben találhatunk. Az eredményt a Fig. 3. ábra mutatja.

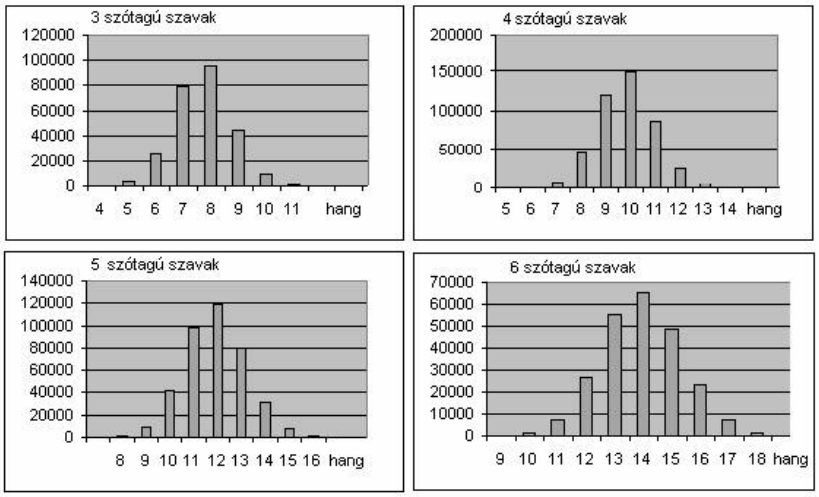


Fig. 3 A 3, 4, 5 és 6 szótagú magyar szóalakok hangszám szerinti eloszlása a kiejtési szótár szóalak állományában

Mindezek az adatok azt mutatják, hogy a kiejtési szótárunk jellemzően 7-15 hangot tartalmazó szóalakokat tartalmaz.

2.1 A hangjelölések és a hangátírás

A kiejtési szótár akkor használható jól, ha a hangjelölésekre nemzetközileg elfogadott jelrendszert alkalmaznak a fejlesztők. Mi is ezt az elvet követtük. Így bármely anyanyelvű felhasználó azonosítani tudja a kiválasztott magyar szó kiejtési formáját. A fonetikában használatos hangszimbólumokkal (IPA jelek) adjuk meg a szóalakok

hangalaki formáit. Az átíráshoz 9 magánhangzó és 25 mássalhangzó jelét használjuk, nem számítva a fonológiailag hosszú hangokat, amelyek hosszúságát kettősponttal jelöljük. A hangok szimbólumait a szótárban szögletes zárójelek közé tesszük.

1. Táblázat: A szótárban használt IPA jelek a magyar beszédhangok jelölésére

Betű	IPA jel	Betű	IPA jel	Betű	IPA jel	Betű	IPA jel
á	aː	b	b	n	n	zs	ʒ
a	ɔ	p	p	ny	ɲ	s	ʃ
o	o	d	d	j	j	cs	t͡ʃ
u	u	t	t	h	h	l	l
ü	y	g	g	v	v	r	r
i	i	k	k	f	f	dz	d͡z
é	eː	gy	ɟ	z	z	dzs	d͡ʒ
ö	ø	ty	c	sz	s		
e	ɛ	m	m	c	t͡s		

A szöveg-hang átalakítást egy nagy elemszámú szótár állományára csak automatikus feldolgozással lehet hatékonyan elvégezni. Hangátírási algoritmust kellett készíteni. Ennek fő elemeit a magyar szakirodalomban található kiejtési szabályok képezik. A megvalósított algoritmus alapfilozófiája a következő: meghatározunk szabályokat és alkalmazzuk azokat, majd felsoroljuk a kivételeket a szabályok alól. A kérdés minden esetben az volt, hogy mit nevezzünk szabálynak. Azt az elvet követtük, hogy egy hangátírási formánál felmértük annak hatókörét. A többségi előfordulást tekintettük szabálynak, és a kivétellistákba helyeztük el és oldottuk fel a kevesebb előfordulást. Így tudtuk helyes átírással lefedni a kiejtéssel kapcsolatos igen gazdag formációk közel teljes állományát.

A hangátírási algoritmusunk három szinten végzi a hangátírást: betű-hang szabályok, posztlexikális módosulások, kivételek kezelése (kivétel listák a szabályok mellett, illetve a kivétel szótár, amely önálló eleme a rendszernek).

Az alapot képező hangátírási szabályrendszer végzi általánosságban a betűképek hangszimbólumokká való átalakítását. Az eredmény sok esetben a végleges hangátírási forma (*ablak* = [ɔ b l ɔ k], azonban sok esetben nem. Ha nem, akkor a következő további feldolgozási formákat alkalmazzuk. Külön kivételként kezeljük a hangkivetés néhány esetét (*mondta* = [m o n t ɔ], *küldte* = [k y l t ɛ]ét (*mondta* = [m o n t ɔ], *küldte* = [k y l t ɛ]. A hangidőtartamok tekintetében két további alsóbb szinten módosulhat a kapott hangsorozat: hosszan írjuk, röviden mondjuk (*vá**ll**alat* = [v a: **l** ɔ l ɔ t], *ko**mm**unikál* = [k o **m** u n i k a: l]), *me**nny**ország* = [m ɛ ɲ o r s a: g], jo**bb**ra = [j o **b** r ɔ]. Ennek ellenkezője is előfordul, amikor röviden írjuk, hosszan ejtjük a szó valamelyik hangját ***USB*** = [**u:** e ʃ b e:], *NAT**O*** = [n a: t **o:**].

Posztlexikális szabályok

A magyar hangátírás legproblematikusabb része a hasonulások korrekt kezelése. Esetünkben azokkal foglalkoztunk, amelyeket a helyesírásunk nem jelöl. Ezek a szabályok a mássalhangzókat érintik. Ilyen beépített szabályok a részleges hasonulásból a zöngésedés, illetve zöngétlenedés, a képzés helye szerinti hasonulásból az [n]

hasonulása [m ɲ] hangokká (szí**n**pad = [s i **m** p ɔ d], po**n**ty = [p o ɲ c]). A teljes hasonulások, illetve az összeolvadás tekintetében hangsúlyozottan alkalmaztuk a kivétellistákat (például több esetben a kiejtés közeledik az írásképhez, főleg összetett szavak határán (*telj**e**s* = [t ɛ **j:** ɛ ʃ], de *feljavít* = [f ɛ l j ɔ v i: t]), kétséges = [k e: tʃ: e: g ɛ ʃ], de kétsávú = [k e: t ʃ a: v u]). A szótár végleges kialakítása során közel 50.000 szóalak kiejtése került kézi ellenőrzés formájában meghatározásra.

Kivétel szótár

A hangátírási szabályok harmadik fő modulja a kivétel szótár. Ide kerülnek elhelyezésre az idegen szavak, nevek, rövidítések és minden olyan kiejtési forma, amelyik az előző két modullal nem lefedhető (city = [s i t i], plasa = [p l a: z ɔ], Peugeot = [p ø ʒ o:], MTA = [ɛ m t e: ɔ:]). Esetünkben a magyar családnevek kiejtési meghatározásai is ebben a kivétel szótárban vannak (Kossuth = [k o ʃ u: t], Bernáthffy = [b ɛ r n a: t f i], Előrsy = [ő r ʃ i]), Unghváry = [u n g v a: r i]). A kivétel szótár közel 12.000 elemet tartalmaz.

3 A hangos szótár

A mai beszédtechnológiai eszközök lehetővé teszik, hogy egy kiejtési szótárban nagy számú hangzó példát is beépítsünk. Esetünkben 60.000 szóalak hangos formáját készítettük el (ennek nem az a formája, hogy felolvastuk a szavakat). Ezek mindegyike meghallgatható. Azért döntöttünk a hangos szótárrész elkészítése mellett, mert ezzel közelebbi támpontot tudunk nyújtani a nemzetközi közösségből kikerülő felhasználóknak a magyar hangsorépítés időszerkezeti viszonyairól is. A kiejtési hangjelekben ugyanis nincsenek információk a hangok időtartamairól, csak a fonológiai rövid-hosszú szembenállásról (a magyarban fontos a kiejtésben jó hosszan realizálni a hosszú hangokat). Ezért a tényleges hangzó forma meghallgatásával az alapvető magyar kiejtés időszerkezeti képét is megismerheti, érzékelheti a felhasználó.

A szóalakok kiválogatását a szótár teljes állományából végeztük, mégpedig a Fig. 2. ábrán megadott szóhosszúsági eloszlásnak megfelelően. Így minden hosszúságból arányos számú szó került a meghallgatható listába. A szavakat a Profivox magyar beszédszintetizátorral olvastattuk fel [4], így generáltunk 60.000 wav fájlt teljesen automatikusan. A szintetizátorban a hangidőtartamok kialakításánál a magyar kiejtési időmodell szabályai működtek [3]. A szóalakok kiejtési ritmusa megfelel a magyar köznyelvi kiejtés kritériumainak.

4 A kiejtési szótár szerkezeti felépítése

A kiejtési szótár az Interneten bárki számára hozzáférhető, használatához egy böngészőprogram szükséges. A szótárból a keresőgépekhez hasonló, könnyen kezelhető felületen keresztül kapjuk meg a kívánt szóalakok listáját. Az ilyen – webes környezetben megszokott – keresések során alapvető követelmény, hogy a keresőkérdésre illeszkedő találatok gyorsan jelenjenek meg a böngészőben.

A keresési idő csökkentése érdekében a kiejtési szótár 1,8 millió szóalakjának betűképét és hangsorát relációs adatbázisban (MySql) tároljuk. A 60.000 meghallgatható szóalak mindegyikéhez pedig egy-egy külön tárolt hangállomány (WAV) is kapcsolódik. A szótár helyigénye igen nagy, közel 3 GB, melynek nagy részét (kb. 95 %-át) a szintetizált szavak alkotják.

5 A szótár szolgáltatásai

Keresés a szótárban. A keresett szót magyar betűkarakterekkel kell megadni. Betűkapcsolatot is megadhatunk, ilyenkor minden olyan szót megkapunk, amelyikben az adott betűkapcsolat szerepel. A * karakter egyfajta joker szerepet tölt be a keresés megadásában. Például az „úszóedző*” megadására a szótár minden olyan szót megmutat, amelyik a * előtti karaktersorozattal íródik (*úszóedzővel, -nek, -ről, -mnek, -iket* stb.). A kikeresett szó magyar helyesírású betűképe mellett megjelenik a szó hangjainak sorozata IPA szimbólumokkal megadva. Ez a kiejtési forma. Amennyiben a szó mellett megjelenik egy hangszóró ikon az azt jelenti, hogy a szó meghallgatható.

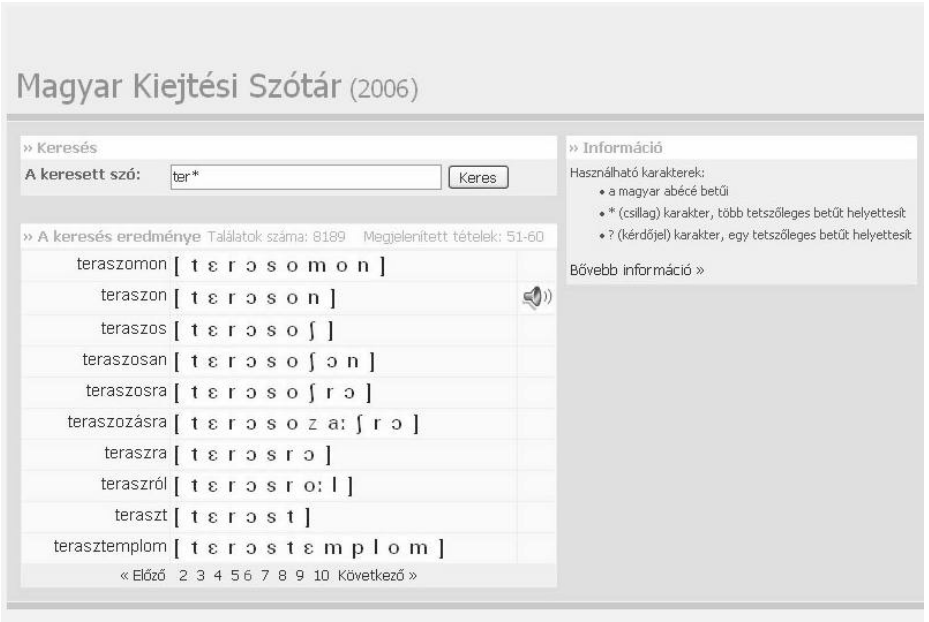


Fig. 4. Példa a kiejtési szótár találati listájára, a *ter** betűkapcsolatot tartalmazó szavakkal

A * karakter mellet használhatjuk a ? (kérdőjel) helyettesítő karaktert is, amely pontosan egy tetszőleges karaktert helyettesít. Ha a beviteli mezőbe a *v?r* karaktersorozatot visszük be, akkor a találati listában a *var, vár, ver, vér* szavakat kapjuk.

Amennyiben a keresőkérdésünk nem elég pontos – vagyis túl sok szó illeszkedik a kért betű- és helyettesítőkarakterek mintájára – egy lapozható találati listát kapunk, melyben az *Előző* vagy *Következő* linkeken kattintva kapjuk a szomszédos 10 találа-

tot. A Fig. 4. ábrán látható, hogy a *ter** keresőkérdésre 8189 db szóalak illeszkedik, amelyből éppen az 51-től 60-ig terjedő tételeket jelenítettük meg az ábrán. Az éppen listázott elemek közül a *teraszon* szólakot meg is hallgathatjuk.

Bibliográfia

1. A magyar nyelv könyve. Főszerkesztő: A. Jászó Anna.. Trezor Kiadó. 1991.
2. Fekete László: Magyar Kiejtési szótár. Gondolat Könyvkiadó. 1992.
3. Olaszy Gábor: Hangidőtartamok és időszerkezeti elemek a magyar beszédben. Nyelvtudományi Értekezések 155. Akadémiai Kiadó. 2006
4. Olaszy Gábor: Profivox- a legkorszerűbb hazai beszédszintetizátor. Beszédkutatás-2000. Szerk.: Gósy Mária. MTA Nyelvtudományi Intézet. Budapest. 2000. 167-179
5. Szende Tamás: Spontán beszédanyag gyakorisági mutatói. Nyelvtudományi Értekezések 81. Akadémiai Kiadó, 1973.
6. Tóthfalusi István: Kiejtési szótár. Tinta Kiadó. 2006. november 5.
7. Zainkó, Cs., Németh G.: Statistical Text Processing for Automatic Synthesis of Speech, Proc. of ECMCS2001 (EURASIP Conference on Digital Signal Processing for Multimedia Communications and Services), 2001. 644-647

Koartikulációs modellek a magyar nyelvű gépi beszédfelismerésben

Mihajlik Péter

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
mihajlik@tmit.bme.hu

Kivonat: A koartikulációs jelenségek két elvi csoportjának, a fonológiai és a fonetikai koartikuláció modellezésének kérdéseit, megoldásait vizsgáljuk magyar nyelvű statisztikai-alapú gépi beszédfelismerés esetén. A koartikulációs modellek beszédfelismerési hálózatba integrálásának érdekében bevezetjük a súlyozott véges állapotú átalakító (WFST) alapú felismerési hálózatépítést. Bemutatjuk az alkalmazott explicit és implicit koartikulációs modelleket, melyeket az általunk elérhető legnagyobb magyar nyelvű – részben publikus – telefonbeszéd-adatbázisok segítségével értékelünk ki. Az eredmények meggyőzően mutatják, mely típusú koartikuláció modellezésére érdemes nagyobb hangsúlyt fektetni a folyamatos beszédfelismerési pontosság jelentős növelésének érdekében.

1 Bevezetés

A koartikuláció – az egymást követő hangok egymásra hatása, „együtt ejtése” – a beszéd alapvető jellegzetessége. Nem különbözik e tekintetben a magyar nyelv más nyelvektől, amit az is jelez, hogy a korszerű magyar nyelvű gépi beszédelőállítási módszerek mindegyike elemi vagy magasabb szinten explicit hangkapcsolati modelleket használ (diádok, triádok, stb.) [5].

A magyar nyelvű gépi beszédfelismerés területén ugyanakkor – a nemzetközi trendekkel ellentétben – a különböző koartikulációs jelenségek explicit modellezése nem jellemző. Tipikus a környezetfüggetlen modellek használata, melyeknél a fonéma – beszédhang szintek szétválasztása fizikailag nem történik meg. A kutatócsoportunkhoz kötődő publikációkon felül nem ismerünk kísérleteket a magyar nyelvű koartikulációs jelenségeket explicit módon kezelő beszédhang-modellezéssel kapcsolatban.

A következőkben rövid áttekintést kívánunk nyújtani a magyar nyelv koartikulációs jelenségeinek modellezésével kapcsolatos nehézségekről, jellegzetességekről, és megoldásairól. Az egyes megközelítések hatását a beszédfelismerés pontosságára a legnagyobb elérhető magyar nyelvű beszédadatbázisokon különféle konfigurációkban és felismerési feladatokban mértük. Az eredmények megbízhatóságát szignifikancia-vizsgálattal ellenőriztük.

2 A koartikulációs jelenségek osztályozása

2.1 Fonológiai koartikulációs jelenségek

A modern nyelvtudomány a "kiejtési szabályok" néven összegyűjtött hasonulási, összeolvadási, stb. jelenségeket fonológiai koartikulációs jelenségeknek hívja. Ezek főbb ismérve, hogy egy vagy több beszédhang fonémaértéke megváltozik a kiejtés során (pl. *azt → a sz t*). A megváltozás lehet összetettebb jelenség, beleértve a kiesést vagy betoldást is (pl. *értsd → é r dzs d, tea → t e j a*). Külön említendők a szóhatárokon fellépő fonológiai változások (pl. *értds te → é r dzs d _ t e* vagy *é r cs t e*), melyek attól is függhetnek, hogy tart-e szünetet a beszélő a két szó között vagy sem, illetve, természetesen attól is, hogy milyen hanggal kezdődik a következő szó.

A fonológiai koartikulációs jelenségek egy lehetséges csoportosítása a következő:

- Zöngésségi (részleges és teljes) hasonulások: *adta → a tt a, lékbe → l é g b e*
- Képzés helye szerinti (részleges és teljes) hasonulások:
  azonban → a z o m b a n, önmaga → ö mm a g a
- Mássalhangzó-rövidülések: *állt → á l t*
- Összeolvadások: *látja → l á tty a, utca → u cc a, kétség → k é ccs é g*
- Egyéb kiesések, betoldások:
  parasztkolbász → p a r a sz k o l b á sz, tea → t e j a

2.2 Fonetikai koartikulációs jelenségek

A fonetikai koartikuláció a beszéd nagyon fontos jelensége. Lényege, hogy a beszélőszervek tehetetlenégének, folyamatos mozgásának következtében a hangátmenetek nagy része is folyamatos, így a beszédhangok jelentős része az önmagában való ejtéshez képest megváltozik. A fonetikai koartikuláció segít például a felpattanó zárhangok felismerésénél, ahol a környező magánhangzók formánsátmenetei engednek következtetni a zárhang identitására.

A fonetikai koartikuláció természetesen szóhatárokon is felléphet, illetve a beszédszünet is hatással lehet a környező hangokra.

3 A koartikulációs jelenségek modellezése

3.1 A WFST keretrendszer

Az előzőekben leírt koartikulációs jelenségek modellezése különösen a folyamatos beszédfelismerés esetén jelent nagy kihívást, hiszen attól függően lép fel egyik vagy másik jelenség, hogy az adott szó után melyik másik következik. Egyedi specializált megoldások helyett a súlyozott véges állapotú átalakítókkal (WFST – Weighted

Finite-State Transducers) történő tudásforrás reprezentációt és integrációt választottuk, mely általános, matematikailag is letisztult keretet biztosít a feladathoz.

A súlyozott véges állapotú átalakítók formálisan a *félgyűrűk* felett értelmezett matematikai objektumokként definiálhatók [4]. Praktikusan a véges állapotú gépek olyan általánosításának tekinthetők, melyek egy adott bejövő szimbólumsorozatnak nem csak az elfogadásáról vagy elvetéséről dönthetnek, hanem képesen súlyt és kimeneti szimbólumsorozatot is rendelni hozzájuk.

A WFST-keretrendszerben a beszédfelismerés során használt tudásforrásokat – úgymint nyelvi modell, kiejtési szótár, beszédhangmodellek, stb. – először súlyozott véges állapotú átalakító formára kell hozni, majd ezeket standard WFST műveletekkel lehet egybe komponálni és optimalizálni. Két tudásforrás ötvözésére a kompozíció (jelölés: o), az egyes tudásforrások optimalizálására pedig a determinizáció és minimalizáció használható (jelölés: det, min) [4].

A felismerési hálózat összeállításának szemléltetése környezetfüggetlen beszédhang-modellezés esetén:

$$\textit{Felismerési hálózat} = \min(\mathrm{H} \circ \mathrm{L} \circ \mathrm{G}), \qquad (\mathbf{1})$$

ahol G: a nyelvi modell
L: a kiejtési modell
H: a fonémák leképezése elemi akusztikus modellekre

A felismerési hálózat ilyenkor HMM (rejtett Markov-modell) állapot szimbólumsorozatot képez le szósorozatra a nyelvi, kiejtési és egyéb súlyoknak megfelelően, így közvetlenül használható a „hagyományos” Viterbi-féle HMM dekódolási algoritmus a felismerési eredmények valós idejű meghatározásához.

A keretrendszer óriási előnye a flexibilitás. Bármilyen kiejtési alternatívákat valószínűségekkel ellátó, vagy a legegyszerűbb fonológiai kiejtési modell integrálható, csakúgy mint a bigram helyett trigram vagy 4-gram nyelvi modell a rendszer bárminemű megbontása nélkül. Hátránya, hogy az inkrementális hálózatépítés (új szó hozzáadása) alapesetben a teljes felismerési hálózat újraépítését teszi szükségessé, mely jelentős számítási igénnyel járhat.

3.2 Fonológiai koartikulációs modellek

A hasonulási, egybeolvadási stb. fonológiai koartikulációs szabályok speciális esetei a környezetfüggő újraíró szabályoknak. Ezek WFST implementációjáról részletesen szól a [4], de az egyedi implementálás is járható út.

Az egyes zöngésségi, képzés helye szerinti hasonulások és opcionális összeolvadások súlyozott véges állapotú átalakítóinak kompozíciójával kaphatjuk meg az általános fonológiai koartikulációs modell WFST reprezentációját

A kísérletekben a [3]-ban bemutatott hierarchikus fonológiai koartikulációs szabályrendszer WFST megfelelőjét használtuk, melyet a következő oldalon részletezett módon állítottunk össze elemi szabálytípusoknak megfelelő véges átalakítókból.

P_1: Zöngésségi hasonulás /kötelező/

P_2: Összeolvadás + Rövidülés /kötelező/

P_3: Képzés helye, módja szerinti részleges hasonulások /opcionális/

P_4: Képzés helye, módja szerinti teljes hasonulások /opcionális/

Az fonológiai koartikulációs modell, P, [3] után az alábbi kompozíció-sorozattal adódik:

$$P = P2 \text{ o } P4 \text{ o } P3 \text{ o } P2 \text{ o } P1 \qquad (\mathbf{2})$$

Ez a modell a 2.1-ben említett fonológiai koartikulációs jelenségek közül mindegyiket explicit módon, *szóhatárokon átívelve* (is) kezeli. Kivételt csak az "egyéb kiesések, betoldások" képeznek, mert ezek esetlegesek, ritkák és automatizáltan nem állíthatók elő. Megjegyezzük, hogy a szóhatárokon átívelő koartikulációt csak akkor tesszük lehetővé, ha a két szó közé nem esik szünet a kiejtés során.

3.3 Fonetikai koartikulációs modellek

A fonetikai koartikuláció explicit modellezésére a környezetfüggő beszédhangmodelleket, ezen belül is a szóhatárokon átívelő („cross-word"), mindkét oldalon 1 hang távolságig környezetfüggő (trifón) modelleket választottuk. A trifón modelleket 3 állapotú, „left-to-right" struktúrájú rejtett Markov-modellek (HMM-ek) képviselik. Így a koartikulációt beszédhangonként a fonetikus környezettől függő kialakítású 3 kiejtési fázissal modellezzük.

Az általánosított trifón modellek állapotait fonémánként és állapotonkénti ML (Maximum Likelihood) fonetikus döntési fákkal csoportosítottuk [7]. Mivel az eljárás magyar nyelv esetén még nem bevett, ugyanakkor mind elméleti, mind gyakorlati szempontból fontos, a következőkben röviden vázoljuk.

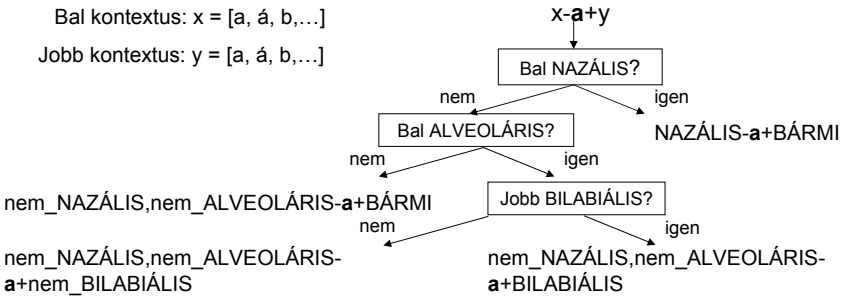
A módszer lényege, hogy az adott fonéma adott állapotához tartozó általánosított trifón állapotokat a kezdeti egy csoportból lépésről-lépésre úgy osztja további csoportokra, hogy a felhasználó által definiált fonetikus környezetre utaló kérdéseket sorban felteszi, és végül azt választja, mely ML értelemben a legjobb szeparációt jelenti. Az eljárás akkor ér véget, amikor egy csoportra már nem jut elég tanítóminta, vagy az új csoport kettéosztás már nem hoz érdemi hasonlósági mérték növekedést a tanítóadatbázison. Végeredményben egy döntési fa áll elő minden fonéma bal, középső és jobb állapotára (esetünkben összesen 3x39 döntési fa), melynek levelei reprezentálják a trifón állapot csoportokat.

Az eljárást a következő példával szemléltetjük.

Legyenek a bal és jobbkörnyezetekre utaló kérdések az alábbi módon definiálva:

NAZÁLIS: m, n, …
ALVEOLÁRIS: d, t, n, …
VELÁRIS: g, k, …
BILABIÁLIS: p, b, m…

Egy lehetséges döntési fa az „a” hang bal szélső állapotára:



1. ábra. A fonetikai döntési fa alapú trifón állapot csoportosítás szemléltetése.

Szemléltető példaként tekintsük a „pamacs” szó első és második „a” hangjának elemi akusztikus modellekre (HMM állapotokra) történő leképezését. Az első „a” általánosított trifónja a „p-a+m” a másodiké a „m-a+cs”: ezeket a fenti döntési fán kiértékelve kapjuk meg az adott – esetben a bal – állapot elemi akusztikus modelljét, mely a csoporthoz tartozó mintákkal lett tanítva. A „p-a+m” bal szélső állapota a fenti döntési fa alapján a „nem_NAZÁLIS,nem_ALVEOLÁRIS-**a**+BÁRMI” trifón állapot csoportba, míg az „m-a+cs” bal állapota pedig „NAZÁLIS-**a**+BÁRMI” csoportba kerül.

Az ML döntési fa-alapú trifón állapotcsoportosítás jó tulajdonsága, hogy a csoportok avagy az elemi akusztikus modellek száma két – a fa-építés leállításánál említett – küszöbérték segítségével széles határok közt állítható. Továbbá, hogy a tanító-adatbázis méretéhez jól alkalmazkodik, kisebb adatbázis esetén kevesebb, nagyobb adatbázisnál nagyobb számú elemi akusztikus modellre képez le azonos küszöbértékek esetén is. A megközelítés hátránya, hogy döntési fa építéshez igényel egy általánosított trifón szintű akusztikai modell tanítást is a tanító beszédadatbázison.

A fonetikai koartikulációs modell WFST-formátumra hozása két lépésben történhet. Az első lépés a fonémasorozat - általánosított trifónsorozat leképezés, melyet a CD véges átalakító végez. Súlyokra itt nincs szükség, mivel a leképezés egyértelmű. A CD átalakító képzésének kifejtésére itt nem vállalkozhatunk, az a [4]-ban megtalálható. A következő lépés az általánosított trifónok elemi akusztikus modellekre (HMM állapotokra) való leképzése, melyet a H_{tri} véges átalakító hajt végre. Ehhez az összes lehetséges általánosított trifón megfelelő döntési fán való kiértékelése szükséges, melynek eredménye egy trifón kiejtési táblázatba foglalható, amely már triviálisan alakítható véges átalakítóvá.

A fonéma-HMM állapot leképezés a környezetfüggő beszédhangmodelleknél tehát a H_{tri} o CD kompozícióval adódik, ahol a H_{tri} kialakítása tanító-adatbázis függő.

4 A koartikulációs modellek kiértékelése

A koartikulációs modellek kiértékelésénél természetes választás a beszédfelismerési tesztekkel történő minősítés. Ilyenkor fontos, hogy a felismerési feladat elég általános

legyen, valamint, hogy a teszt (és tanító) adatok elég változatosak, nagyszámúak és reprezentatívak legyenek. Továbbá a tanító és teszt adatok függetlenségének biztosítása is kívánatos, vagy legalábbis ennek kézbentartása.

A felismerési tesztek eredményei azonban önmagukban nemigen használhatók, ezért minden kísérletnél összehasonlításokat végeztünk. Általában az előző fejezetben tárgyalt explicit koartikulációs modelleket hasonlítottuk össze az implicit modellekkel. Fonológiai koartikulációnál az implicit modellt az jelentette, amikor a fonológiai szinten tanításnál sem vettük figyelembe a koartikulációs jelenségeket. A fonetikai koartikulációnál pedig az implicit modell a monofón, azaz a környezet független beszédhang-modell volt.

A következőkben röviden összefoglaljuk a kísérleti körülményeket, majd az elvégzett vizsgálatok lépéseit és eredményeit.

4.1 Kísérleti körülmények

Beszédadatbázisok:

A tanító- és tesztelő-adatbázisokat a legnagyobb magyar telefonos beszédadatbázisok, az MTBA, a Besztel, a SpeechDat és a Tesztel összességéből alakítottuk ki [6]. Ezek az adatbázisok elsősorban olvasott beszédet, valamint kisebb arányban spontán bemondásokat is tartalmaznak. Az első három adatbázis lényegében ugyanarra a szövegkorpuszra épül, és mindegyiknek az általunk elérhető része 500 beszélőtől tartalmaz hanganyagot. A Tesztel adatbázis 100 beszélős, és jellegzetessége, hogy szándékosan nagy és természetes háttérzajban felvett bemondásokat tartalmaz. Az adatbázisokban a vonalas és mobil telefonos felvételek összességében körülbelül ugyanolyan számban képviseltetik magukat.

Tanítóhalmazok:

Tanítás céljára az MTBA, Besztel, és a SpeechDat adatbázis 500-400-450 beszélőjének azon felvételeit jelöltük ki, melyek nem „o”, és „z” jelzésűek, azaz nem tartalmaznak tulajdonneveket és bizonyos típusú mondatokat. A SpeechDat esetén csak egy szűkebb halmazt, a fonetikailag változatos szavakat és mondatokat (kivéve a „z” jelzésűeket) használtuk.

A teljes tanítóhalmaz mellett annak bizonyos részhalmazait is képeztük, hogy a különféle koartikulációs modellezési eljárások tanító adatbázisméret-függését is vizsgálhassuk.

Sem a tanítóhalmazokban, sem a későbbi teszthalmazokban nem végeztünk szűrést az annotációnál zajosnak minősített felvételekre. Kizárólag azokat a felvételeket hagytuk ki, melyeknek az eleje vagy vége az annotáció szerint nem került rögzítésre.

A tanítóhalmazok jelölése és tartalma:

- **M**: Az MTBA fonetikailag változatos mondatai és szavai, 500 beszélő, 6000 felvétel
- **MM**: Az MTBA összes tanítófelvétele, 500 beszélő, 19000 felvétel.
- **MM_BS**: Az MTBA és a Besztel összes tanítófelvétele, 900 beszélő, 39000 felvétel.
- **MM_BS_SD**: Az MTBA, a Besztel és a SpeechDat tanítófelvételei, 1350 beszélő, 44000 felvétel.

A felismerési feladat:

Az általános tapasztalat szerint a beszélőfüggetlen folyamatos beszédfelismerés támasztja a legnagyobb igényeket a kiejtési – koartikulációs modellekkel szemben. Ezért olyan *általános* folyamatos beszédfelismerési feladatot próbáltunk definiálni, ami a rendelkezésre álló adatbázisokkal megvalósítható. Természetesen adódott, hogy az adatbázisok azon mondatait tartalmazó bemondásokat ismertessük fel, melyek nem szerepelnek a tanító halmazokban. A beszélőfüggetlenség követelménye miatt azon felvételeket is ki kellett zárnunk, melyeknek a beszélőjét felhasználtuk a tanítás során.

Teszthalmazok:

A teszthalmazokat tehát úgy állítottuk össze, ne legyen átfedés a tanítóhalmazban szereplő beszélőkkel. Így a tanításnál fel nem használt 170 beszélőtől (Besztel 100, SpeechDat 50, TeszTel 20) kerültek felvételek a teszthalmazokba. Összesen 2385 felvételt kaptunk, melyeket a tanító-adatbázishoz való illeszkedés mértéke szerint két halmazra bontottunk.

A folyamatos beszédfelismerésnél fonológiai és nyelvi illeszkedésről is beszélhetünk. Az egyik halmazba azokat a mondatokat válogattuk, amelyeknek *szöveges tartalma* egyezett az akusztikus modelltanításnál használt mondatokéval (fonológiai illeszkedés), valamint amelyeknek szöveges tartalma a nyelvi modell tanításakor is felhasználásra került (nyelvi illeszkedés), ez az „illeszkedő” halmaz. A másik, „nem illeszkedő” teszthalmazba azok a felvételek kerültek, melyek szövegtartalma sem az akusztikus, sem a nyelvi modell tanításkor nem lett felhasználva. Egyéb halmazt nem vizsgáltunk.

A teszthalmazok jelölése és tartalma:

- **I**: Nyelvi és fonológiai szempontból a tanításhoz illeszkedő mondatok, 170 beszélő, 1973 felvétel: „s” jelzésű mondatok a Besztel-ből és a SpeechDat-ból, „s1” és „s2” jelzésű mondatok a TeszTel-ből.
- **N**: Sem nyelvi és sem fonológiai szempontból a tanításhoz nem illeszkedő mondatok, 170 beszélő, 412 felvétel: „z” jelzésű mondatok a Besztel-ből és a SpeechDat-ból, „s3” jelzésű mondatok a TeszTel-ből.

Beszédfelismerési paraméterek, beállítások:

Lényegkiemelés: Lényegkiemelési paraméterekként a bemenő beszédjelből MFCC (Mel Frequency Cepstral Coefficients) 12 dimenziós vektorokat képeztünk, melyekhez *logE* (keretenkénti logaritmikus energia) paramétert is csatoltunk, majd dinamikus Delta és Delta-Delta értékeket számítottunk. A statikus energiát végül kicsatolva összesen 38 dimenziós jellemzővektorokat kaptunk. Mind a tanítás, mind a tesztelés során alkalmaztuk a vak csatornakiegyenlítés (Blind Equalization) módszerét [2].

Elemi akusztikus modellek: Az atomi modellek rejtett Markov-modell állapotok voltak rögzített hurok és továbblépési valószínűségekkel. Állapotonként maximum 10 Gauss függvényből álló folyamatos megfigyelési sűrűségfüggvényeket használtunk.

Fonetikai koartikulációs modellek (H_{mono} és H_{tri} o CD): Mind a monofón mind a trifón modelleknél a beszédhangokat 3 elemi akusztikus modellre képeztük le, az előbbi esetben a környezettől függetlenül az utóbbi esetben az ML döntési fa alapján a fonetikus környezettől függően. A döntési fákat - és így a H_{tri} leképezést - tanítóhalmazonként és fonológiai modellenként újraépítettük.

Fonológiai koartikulációs modell (P): A 3.2-ben ismertetett módon állítottuk össze a „kiejtési szabályok" néven közismert fonológiai koartikulációs jelenségek túlnyomó részét modellező véges állapotú átalakítót.

Lexikai modell (L): A kiejtési modellek nyers, fonológiai koartikulációkat nem tartalmazó fonemikus átiratait automatikusan állítottuk elő. Allofónikus változatokat nem jelöltünk, továbbá a hosszú és rövid mássalhangzókat sem különböztettük meg. Így – a szünetmodelleket nem számítva – összesen 39 fonológiai kategóriát használtunk. A szünetmodell háromállapotú környezetfüggetlen modell volt.

Az alkalmazott 5561 elemű szótár az összes előforduló szót tartalmazta (beleértve az illeszkedő és a nem illeszkedő teszthalmaz szavait), így szótáron kívüli elemek kezelésére nem volt szükség.

Nyelvi modell (G): A folyamatos felismerésnél szó-trigram nyelvi modelleket alkalmaztunk Katz-féle visszametszéssel és Good-Turing valószínűség-újraelosztással [1]. A tanítószöveg az illeszkedő tesztmondatok szövege alapján készült úgy, hogy minden különböző mondatot csak *egyszer* szerepeltettünk. Így az illeszkedő mondatokon PP=40-es perplexitást, a nem illeszkedő tesztmondatokon PP=6230-as (nagyon magas, azaz igen kedvezőtlen) perplexitás értéket kaptunk.

4.2 A fonológiai koartikulációs modellek kiértékelése manuálisan, fonémaszinten szegmentált tanító-adatbázis mellett

Első lépésként a fonológiai koartikuláció modellezés vizsgálatát tűztük ki célul adott, beszédhang-szinten *kézzel szegmentált és ellenőrzött* tanítóadatbázis-feldolgozás mellett. Erre egyedül az M-jelű tanítóhalmaz volt alkalmas (MTBA, fonetikailag változatos szavak, mondatok).

Explicit fonetikai koartikulációs modellezés – azaz trifón modellek – mellett végeztük az összehasonlítást, mert egyéb vizsgálataink szerint (lásd a 4.4 pontot) ez jelentette a nem vizsgált paraméterek optimális beállítását.

Az alábbi két felismerési hálózattal végeztünk kísérleteket:

- H_{tri} o CD o L o G – nincs fonológiai koartikuláció-modellezés
- H_{tri} o CD o P o L o G – explicit fonológiai koartikuláció-modellezés

Mivel a beszédhang-modelleket kézzel ellenőrzött – tehát a fonológiai koartikulációkat jelölő – fonetikus szegmentáció mellett tanítottuk, azok nem modellezték még implicite sem a fonológiai koartikulációs jelenségeket. Így a P modell alkalmazásától szignifikáns javulást vártunk. Az eredményeket az 1. és 2. táblázat mutatja.

1. a) és b) Táblázat. Az illeszkedő és nem illeszkedő teszthalmazok folyamatos beszédfelismerési eredményei manuálisan szegmentált tanító-adatbázis mellett.[75]

a) I (PP = 40)	FA	FP
H_{tri} o CD o L o G	93.05	*91.40*
H_{tri} o CD o P o L o G	93.99	*92.57*
ΔH	-13.6	*-13.6*

b) N (PP = 6230)	FA	FP
H_{tri} o CD o L o G	60.84	*49.45*
H_{tri} o CD o P o L o G	62.02	*51.09*
ΔH	-6.1	*-3.2*

Az illeszkedő teszthalmaz esetén kétszámjegyű relatív hibacsökkenés (ΔH) figyelhető meg, ugyanakkor a nem illeszkedő halmaz esetén a javulás szerényebb. Az eredmények szignifikanciáját *2 mintás Z-próba* segítségével ellenőriztük. 5% szignifikanciaszint mellett (95% konfidencia szint) az illeszkedő teszthalmaz esetén valóban szignifikáns javulást tapasztaltunk, míg a nem illeszkedőnél nem.

Az I-hez képest az N teszthalmazon – ugyanazon felismerési feladatban – mért sokkal gyengébb felismerési eredményeket a vonatkozó igen magas nyelvi modell perplexitás (PP) magyarázza.

4.3 A fonológiai koartikulációs modellek kiértékelése következetes tanító-adatbázis feldolgozás mellett

Az előző vizsgálatnál a referencia rendszerben egyáltalán nem modelleztük a fonológiai koartikulációs jelenségeket, mégis csak az egyik teszthalmaznál kaptunk szignifikáns javulást az explicit modell alkalmazásával. Ezért felmerült a kérdés, hogy következetes gépi szegmentációt alkalmazva és nagyobb tanító-adatbázisokat használva is tapasztalható-e érdemi felismerési hiba csökkenés a P véges átalakítónak köszönhetően.

A következő gépi fonetikus szegmentációs módszert dolgoztunk ki a következetes fonológiai modellezés érdekében. A legnagyobb tanítóhalmazra (MM_BS_SD) képeztük a lineáris Gtr „nyelvi modellt”, majd előállítottuk a *tanítóadatokra* vonatkozó felismerési hálózatokat:

- H_{tri} o CD o L o Gtr – *implicit* fonológiai koartikuláció-modellezés
- H_{tri} o CD o P o L o Gtr – explicit fonológiai koartikuláció-modellezés

Kezdeti beszédhangmodelleket tanítottunk be az M tanító halmaz manuális szegmentációja alapján. Ezekkel kényszerített felismerést („forced alignment”) végezve megkaptuk a fonológiai koartikulációt implicit valamint explicit módon tartalmazó gépi fonetikus szegmentációkat.

A különböző tanítóhalmazok és felismerési hálózatok esetén mindig a megfelelő tanítású beszédhangmodelleket alkalmaztuk. Összesen tehát 4x2 akusztikus modell halmazt vizsgáltunk 2 felismerési hálózattal.

---

[75] *Magyarázat a táblázatokhoz:*
FA: felismerési arány [%]. FP: felismerési pontosság [%]. ΔH: a hiba relatív megváltozása [%]. A definíciók részletezését lásd a [7]-ben.

A felismerési hálózatok a 4.2-ben vizsgáltakkal azonosak voltak:

- H_{tri} o CD o L o G – *implicit* fonológiai koartikuláció-modellezés
- H_{tri} o CD o P o L o G – explicit fonológiai koartikuláció-modellezés

Fontos megjegyezni, hogy a következetes tanítás és tesztelés miatt a P modell kihagyása már nem jelenti azt, hogy a fonológiai koartikulációt egyáltalán nem, hanem, hogy implicite, vagyis alacsonyabb, beszédhang szinten modellezzük.

2. a) és b) Táblázat. Az illeszkedő és nem illeszkedő teszthalmazok folyamatos beszédfelismerési eredményei következetes gépi adatbázis-feldolgozás mellett.

a) I (PP = 40)	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	FA	FP	FA	FP	FA	FP	FA	FP
H_{tri} o CD o L o G	94.13	*92.54*	93.82	*91.97*	94.22	*92.55*	94.47	*92.93*
H_{tri} o CD o P o L o G	94.24	*92.69*	93.41	*91.66*	94.14	*92.54*	94.78	*93.05*
ΔH	-1.9	*-2.0*	+6.6	*+3.9*	+1.4	*+0.1*	-5.6	*-1.7*

b) N (PP = 6230)	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	FA	FP	FA	FP	FA	FP	FA	FP
H_{tri} o CD o L o G	61.95	*48.16*	61.27	*47.66*	64.24	*52.34*	64.42	*52.02*
H_{tri} o CD o P o L o G	62.34	*50.14*	61.24	*48.20*	64.09	*51.91*	65.13	*53.20*
ΔH	-1.0	*-3.8*	+0.07	*-1.0*	+0.4	*+0.9*	-2.0	*-2.5*

Ahogy a 2 a) és b) táblázatok mutatják, az implicit és explicit fonológiai koartikulációs modellek beszédfelismerési eredményei között a különbség minimális. A sziginfikancia-vizsgálatok egyetlen esetben sem mutattak ki érdemi különbséget, sőt a hiba nem is csökkent minden esetben.

Észrevehető, hogy a kézi helyett gépi fonetikus szegmentáció az M halmaz esetében nem rontott, hanem még javított is az eredményeken. Gyakorlatilag tehát következetes gépi tanitóadatbázis-feldolgozás mellett ugyanolyan jó eredmények érhetők el fonológiai modell *nélkül* is, mint a manuálisan szegmentált adatbázissal *és* explicit, szóhatárokon átívelő hasonulási modellel.

Nem várt tapasztalat volt ugyanakkor, hogy a tanítóadatbázis méretének növelése alig javított a felismerési eredményeken annak ellenére, hogy minden tanítási konfigurációban újraépítettük a trifón állapotcsoportosítást végző ML döntési fákat. Itt a nagyobb tanítóhalmazoknál a tesztfelvételekhez képesti nagyobb fonológiai illesztetlenség, illetve az adatbázisok gyakorlatilag közös szövegkorpuszra épülése lehetnek a mögöttes okok.

4.4 A fonetikai koartikulációs modellek kiértékelése

A *fonetikai* koartikulációs modellek kiértékelésekor a fenti tapasztalatok alapján az implicit fonológiai koartikuláció kezelést választottuk. Ez a gyakorlatban azt jelentette, hogy a tanító-adatbázis gépi szegmentálásához csakúgy, mint a tesztekhez a P-modell alkalmazása nélkül készítettük a felismerési hálózatokat. Vagyis következetes gépi tanítóadatbázis-feldolgozást alkalmaztunk.

A vizsgált felismerési hálózatok tehát a következők voltak:

- H_{mono} o CD o L o G – implicit *fonetikai* koartikuláció-modellezés
- H_{tri} o CD o P o L o G – explicit *fonetikai* koartikuláció-modellezés

A monofón, vagy környezetfüggetlen beszédhangmodellek a folyamatos megfigyelési sűrűségfüggvényeik révén implicite modellezik a fonetikai koartikulációt, hiszen több fonetikus környezet mellett történik a tanításuk. A kérdés az, hogy ez az implicit modell hasonlóan viselkedik-e, mint a *fonológiai* koartikulációnál az implicit modell.

A különféle adatbázis-méretek és teszthalmazok melletti eredményeket mutatja a következő táblázat.

3. a) és b) Táblázat. Az illeszkedő és nem illeszkedő teszthalmazok felismerési eredményei implicit és explicit *fonetikai* koartikulációs modellek mellett.

a) I (PP = 40)	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	FA	FP	FA	FP	FA	FP	FA	FP
H_{mono} o L o G	85.34	*84.34*	80.19	*78.60*	79.87	*78.54*	80.74	*79.52*
H_{tri} o CD o L o G	94.13	*92.54*	93.82	*91.97*	94.22	*92.55*	94.47	*92.93*
ΔH	-60	*-52*	-69	*-62*	-71	*-65*	-71	*-65*

b) N (PP = 6230)	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	FA	FP	FA	FP	FA	FP	FA	FP
H_{mono} o L o G	29.22	*23.65*	25.40	*19.36*	24.72	*18.54*	25.58	*20.54*
H_{tri} o CD o L o G	62.34	*50.14*	61.24	*48.20*	64.09	*51.91*	65.13	*53.20*
ΔH	-50	*-38*	-52	*-40*	-53	*-42*	-54	*-40*

Mint láthatjuk az implicit és explicit fonetikai koartikulációs modellezés közti különbség drámai, és természetesen minden esetben szignifikáns a hiba csökkenése a „legszigorúbb", 0.01%-os szignifikancia küszöb mellett is.

Felhívjuk a figyelmet a nem illeszkedő teszthalmaz abszolút felismerési eredményeire. Itt nem a felismerési hibák, hanem a *felismerési arányok és pontosságok* között figyelhető meg 2 – 3-szoros különbség.

Az is észrevehető, hogy a környezetfüggetlen beszédhang-modelleknél az adatbázisméret növelése szinte csak rontott a felismerési eredményeken.

Terjedelmi korlátok miatt nem tudjuk részletesen közölni a monofón modellek esetén a P fonológiai koartikulációs modell alkalmazása mellett mért eredményeket. Röviden összefoglalva, abban az esetben relatíve nagyobb mértékben javultak a felismerési mutatók, mint a trifón esetben, azonban az abszolút felismerési arányok továbbra is messze leszakadva követik csak az explicit fonetikai koartikulációs modellek eredményeit.

A kísérletekben a beszédfelismerés számításigénye a trifón modellek esetén kb. 1.5x-ös volt a monofón modellekhez képest, azonban hatékony felismerési hálózat-optimalizációval ezt később 0.8-as, tehát a monofón modellekhez képes *gyorsabb* szintre tudtuk vinni azonos felismerési hiba mellett. Az abszolút számítási igény P4 3GHz-es számítógépen tipikusan valós idő alatt mozgott.

5 Összefoglalás

A koartikuláció két elvi csoportjának, a *fonológiai* és a *fonetikai* koartikuláció modellezésének kérdéseiről, megoldásairól értekeztünk magyar nyelvű statisztikai alapú gépi beszédfelismerés esetén. A koartikulációs modellek integrálhatóságának érdekében súlyozott véges állapotú átalakító (WFST)-alapú beszédfelismerési hálózatépítést használtunk. Az általunk elérhető legnagyobb magyar nyelvű – részben publikus – telefonbeszéd-adatbázisok segítségével értékeltük ki a koartikulációs modelleket, ahol referenciaként általában az implicit koartikulációs modelleket alkalmaztunk.

Megállapítottuk, hogy a *fonológiai* koartikuláció explicit modellezése következetes tanítóadatbázis-feldolgozás mellett nem jár előnnyel az implicit modellezéssel szemben. Ugyanakkor a *fonetikai* koartikuláció explicit modellezésénél mért felismerési eredmények már igen jelentősen túlhaladják az implicit (monofón) modellekkel kapottakat. Különösen a tanításhoz nem illeszkedő teszthalmaz esetén láthatunk drasztikus változásokat az explicit (trifón) modell hatására: itt a *felismerési arányok* nőttek több mint kétszeresükre. Összességében tehát arra jutottunk, hogy a magyar nyelvű statisztikai gépi beszédfelismerésben a fonológiainak a fonetikai koartikulációtól való megkülönböztetésére nincs szükség, míg „a koartikuláció" explicit modellezése úgymond „létszükséglet".

Fontos megjegyezni, hogy az ML döntési fa-alapú trifón állapotcsoportosításnak köszönhetően nem szükséges több száz óra tanítóadatbázis a környezetfüggő beszédhangmodellek betanításához. Amint az eredmények mutatják, a legkisebb, alig pár órás tanítóadatbázis mellett is nagyon jó felismerési eredmények érhetők el. Ezt az adatbázis gondos tervezésének és kifinomult kialakításának tulajdonítjuk, melyet ezúton is szeretnénk megköszönni az alkotóinak.

Bibliográfia

1. Church, K. W. – Gale, W. A. A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of English bigrams. Computer Speech and Language, 5 (1991) 19–54.
2. Mihajlik, P – Tobler, Z. – Tüske, Z. – Gordos, G. "Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech" in In Proc. of InterSpeech'05, Vol 1, pp. 2677-2680, Lisbon, September (2005)
3. Mihajlik, P – Tatai, P. – Gordos, G. "Automatic Phonetic Transcription and Its Application in Speech Recogniser Training – A case study for Hungarian", IOS Press, Amsterdam, NATO ASI series, under the title "Dynamics of speech production and perception;" co-edited by Georg Meyer and Pierre Divenyi, (2006)

Eredmények a magyar nyelvű beszédfelismerési konfidencia-becslésben

Tarján Balázs, Györki Milán, Mihajlik Péter és Gordos Géza

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék, Távközlési és Beszéd-jelfeldolgozási Laboratórium
{btarjan, gyorki, mihajlik}@tmit.bme.hu

Kivonat: A beszédfelismerési konfidencia-becslés célja, minden felismeréshez egy megbízhatósági mérőszámot rendelni. Valódi konfidenciáról viszont csak akkor beszélhetünk, ha a mérőszám jól közelíti a felismerési valószínűséget. Megbízható becslését számos gyakorlati feladat igényli, hiszen nem működhet egy felismerő optimálisan, ha nem képes különbséget tenni biztos és bizonytalan felismerés között. Kísérleteink során sikerült olyan számítási módszert kidolgoznunk, mellyel a konfidencia jól közelíthetővé vált. Cikkünkben összefoglaljuk módszerünk elméleti alapjait, a gyakorlati rendszer felépítését és a rajta elvégzett méréseink eredményeit.

1 Bevezetés

A beszéd gépi felismerésénél sohasem lehetünk biztosak abban, hogy adott felismerési eredmény nem csak optimális, de helyes is egyben. Ezért a gyakorlati alkalmazásoknál jelentős segítséget nyújthat egy olyan eljárás, amely minden felismeréshez egy megfelelő megbízhatósági mérőszámot, azaz **konfidenciát** rendel. Konfidencia birtokában lehetőségünk nyílik egyetlen valószínűségi mérőszámban megragadni mindazt a bonyolult folyamatot, mely az emberi bemondáshoz, egy lehetséges felismert szót rendel.

Megbízható mérőszám előállításának feltétele az optimális felismerő hálózat és a hozzá illeszkedő pontos becslés. Tehát két területen van lehetőség az előrelépésre. Egyfelől különböző felismerési konfigurációk alkalmazásával, másfelől a konfidencia-számítás módszerének változtatásával javíthatunk a becslésen. A cikkünk alapjául szolgáló kísérletsorozatban ezek számos változatát vizsgáltuk pontosságuk alapján, és minden esetben a legjobban teljesítőket emeltük ki. Az így adódott összesen négy kísérleti elrendezés szolgáltatja mérési eredményeinket.

Cikkünk elején összefoglaljuk a konfidencia közelítésének általános alapelveit, és áttekintjük az általunk használt kísérleti elrendezést és annak fontos elemeit. Kitérünk a kísérleti adatbázisra, majd cikkünk második felében a konkrét implementációkon végzett vizsgálatok eredményeit és az azokból levonható következtetéseket ismertetjük.

2 Az alkalmazott konfidencia-becslési alapelvek

2.1 Mintaillesztés a normál nyelvtani hálózathoz

A beszédfelismerés első szakaszában történik a lényegkiemelés. Itt az emberi beszéd időfüggvényének minden 10msec hosszú szakaszához rendelünk egy leíró vektort. Ezt nevezzük jellemzővektornak.

Második lépésben történik a mintaillesztés **[6]**, ennek során a jellemzővektorok sorozatához (jelölése: *O*) hozzárendeljük a szótárban található felismerhető eredmények (jelölése: *W*) közül a legjobban illeszkedőt. Ez a szótár két előzetesen betáplált modellje alapján működik. Az egyik az úgynevezett akusztikai modell, mely az adott nyelvre jellemző fonémák jellemzővektor eloszlását rögzíti. Segítségével közelítőleg megkapjuk, hogy egy szótári elemhez milyen valószínűséggel rendelődik egy jellemzővektor sorozat ($P(O \mid W)$).

A másik fontos elem a nyelvi modell. Ez a felismerhető szavak egymás közti viszonyát leíró egyszerű vagy igen bonyolult hálózat, amiben az élek szavakat kötnek össze, és minden átmenethez rendelődik egy átmeneti valószínűségi érték. A szavak helyébe helyébe fonémák és az azokhoz tartozó akusztikai modellből származó eloszlások helyettesíthetőek be. Ilyen módon az akusztikai modell szervesen beágyazódik a nyelvi modellbe. Ez a modell folyamatos felismerésnél általában szöveges tanítás útján jön létre, ilyenkor valószínűségével szerepelnek benne ($P(W)$). Parancsszóvezérlésnél pedig egy lista megadásával rendelhetünk 1-es értéket az elfogadott szavakhoz.

Összefoglalóan ezt a hálózatot nevezzük a normál **nyelvtan**hoz (grammar) tartozó felismerési hálózatnak. Amikor a jellemzővektorhoz legjobban illeszkedő nyelvtanban rögzített elemet keressük ($\hat{W}$), akkor tulajdonképpen $P(W \mid O)$ valószínűséget kívánjuk W argumentuma mentén maximalizálni (**1**). **[2]**

$$\hat{W} = \arg\max_{W} P(W \mid O) \tag{2}$$

Ezzel ekvivalens problémához jutunk, ha alkalmazzuk a Bayes – formulát (**2**).

$$\hat{W} = \arg\max_{W} \frac{P(W) * P(O \mid W)}{P(O)} \tag{2}$$

Az utóbbi, már átalakított képlet tovább egyszerűsíthető figyelembe véve, hogy P(O) értéke független W-től (**3**).

$$\hat{W} = \arg\max_{W} P(W \mid O) = \arg\max_{W} P(W) * P(O \mid W) \tag{3}$$

Tehát a jellemzővektor sorozatot végigfutatva a nyelvtani hálózaton lépésről lépésre vehetők az illeszkedési, illetve átmeneti valószínűségek szorzatai. A legvalószínűbb felismerést az az útvonalvonal jelöli ki, mely mentén ez a szorzat maximális. Ezt a szorzatot nevezzük a felismeréshez rendelt **hasonlósági mérték**nek.

2.2 A konfidencia-becslése

A bemondás normál nyelvtani hálózathoz illesztésével megkereshetjük a legvalószínűbb illeszkedő szótári eleme(ke)t, ám az így kapott hasonlósági mértékből nem lehet következtetni az illeszkedés valószínűségére. Ennek triviális oka, hogy a hasonlósági mérték függ a végigjárt állapotok számától. Emellett, mivel minden bemondáshoz csak egyetlen legjobban illeszkedő szó rendelődik, arról sem rendelkezünk információval, hogy vajon a többi lehetséges felismerés közül lényegesen kiemelkedik-e a legjobbnak választott.

Tehát valódi konfidencia számításához vissza kell nyúlni eredeti valószínűségi képletünkhöz **[1]** illetve annak Bayes - formulával átalakítottjához (**4**).

$$P(W \mid O) = \frac{P(W) * P(O \mid W)}{P(O)} \tag{4}$$

Az egyenlet bal oldala pontosan a keresett felismerési valószínűséget fejezi ki. A jobb oldalon lévő tagok közül *P(W)* illetve *P(O|W)* számíthatóságát már korábban láttuk, így *P(O)* megfelelő közelítése jelentheti a megoldást. Az ilyen *P(O)* jellemzővektor sorozat előfordulási valószínűségeket azonban megkaphatjuk *P(O|W)* együttes eloszlásának peremeloszlásaként, mint azt az alábbi képlet is kifejezi (**5**).

$$P(O) = \sum_{W} P(W) * P(O \mid W) \tag{5}$$

A beszédfelismerésben alkalmazott nyelvtani hálózatokkal meg tudjuk határozni ennek az összegnek egy tagját, ezt használtuk ki a mintáink illesztésénél, ám a véges szótár nem alkalmas ezen végtelen összeg közelítésére. Szóba jöhetne az úgynevezett N-best módszer **[3]** alkalmazása, amikor a nyelvtani hálózat N legjobb (N>1) illeszkedéséhez tartozó hasonlósági mértékek összegével közelítjük a sort. Izolált szavas felismerésnél azonban ez nem alkalmazható, mert szótáron kívüli bemondás esetén előálló alacsony hasonlósági mértékek közül az N legnagyobb sem dominál, így azok összege és a valódi sorösszeg jelentősen eltér.

Ezért felmerül az igény egy felismerési hálózat összeállítására, mely minden jellemzővektor sorozathoz jól illeszkedik. Egy ilyen **mindenhez illeszkedő hálózat**ban a legjobb illeszkedéséhez tartozó hasonlósági mérték bemondástól függetlenül mindig kiemelkedő. Ezáltal domináns tagja is a P(O)-t meghatározó sornak, azaz vele a sorösszeg is jól közelíthető (**6**).

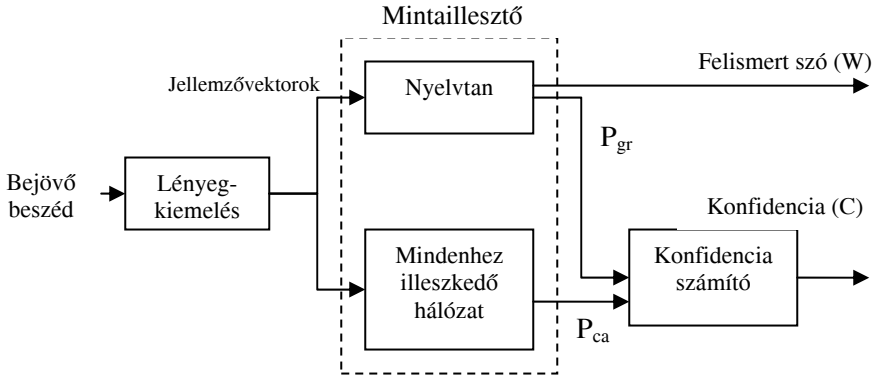
$$P(O) \cong P(W') * P(O \mid W') \tag{6}$$

A mindenhez illeszkedő hálózat, tehát becsülhetővé teszi a P(O) értékét, ami a konfidencia képletének (4) utolsó, idáig ismeretlen tagja volt.

3 A kísérleti elrendezés

Az illeszkedés mértékét a hasonlósági mérték fejezi ki, melyet a normál nyelvtannal végzett mintaillesztés szolgáltat, ám nagysága természetszerűen függ a bemondás jellemzővektor-számától, így magában nem használható konfidencia számítására.

Esetünkben a normál nyelvtani hálózat (nyelvtan) mellett egy különleges, mindenhez illeszkedő felismerési hálózatot is használunk. Konfidencia számításához e két felismerési hálózatot kötjük párhuzamosan, így egy bemondás kiértékelése mindkettőn megtörténik. (**1. ábra**)



1. ábra A párhuzamosan alkalmazott két felismerési hálózat blokkvázlata

A módszer lényege, hogy a nyelvtanhoz való illeszkedést kifejező hasonlósági mértéket (P_{gr}) hasonlítjuk össze a mindenhez illeszkedő hálózaton mért hasonlósági mértékkel (P_{ca}). E két mennyiség aránya alapján a konfidencia már becsülhető (**7**). P_{gr} és P_{ca} hányadosának értéke azonban rendszerint nem esik a kívánatos [0 - 1] valószínűségi tartományba, így „a", „b" értékekkel transzformáljuk, majd az esetlegesen mégis kívül eső elemeket 0 alatt mindig 0-nak, 1 fölött mindig 1-nek tekintjük.

$$C = \begin{cases} C_0 = f\left(\frac{P_{gr}}{P_{ca}}\right) = b * \left(\lg P_{gr} - \lg P_{ca} + a\right), & 0 \le C_0 \le 1 \\ 0, & C_0 < 0 \\ 1, & C_0 > 1 \end{cases} \quad (7)$$

Vizsgálataink során kétféle típusú mindenhez illeszkedő hálózattal dolgoztunk, egy úgynevezett **fonéma-bigram** modellre épülővel valamint egy **multi-Gauss** típusúval. Emellett mindkét esetben kétféle konfidencia variánst használtunk, a pontosabb becslés érdekében. Az első, alap variáns az imént ismertetett módón számítódik, a második, normált variáns is csak annyiban tér el ettől, hogy a logaritmikus valószínűségek különbségét osztjuk a bemondás jellemzővektor számával, és ehhez az értékhez illesztjük „a"-t illetve „b"-t.

4 A mindenhez illeszkedő felismerési hálózatok

Mint láttuk, P(O) közelítéséhez szükségünk van egy olyan felismerő hálózatra, mely minden magyar nyelvű bemondáshoz jól illeszkedik. Így, ha a normál nyelvtanhoz

illeszkedő a bemondás, akkor közel ugyanolyan hasonlósági mértéket rendel hozzá, mint a nyelvtani hálózat. Ha azonban a normál felismerési hálózathoz nem illeszkedik a bemondás, akkor ott ugyan alacsonyabb lesz a hasonlósági mérték, de minden elfogadó hálózatnál a jó illeszkedés miatt marad a magas hasonlósági mérték. A kérdés tehát, hogyan készíthetünk ilyen hálózatokat.

4.1 Multi-Gauss hálózat

Mivel a beszédhangmodellek tanítása során a hozzájuk tartozó rejtett Markov-modell állapotokat eleve GMM (Gauss Mixure Model)-lel jellemezzük, kézenfekvő, hogy az összes beszédhangmodell Gauss-komponenseit összefésüljük, és így egy olyan „meta beszédhangot" használjunk, mely mindenbeszédhangra illeszkedik. Ugyanezt tettük a szünetmodellekkel, és a két „meta modellt" párhuzamosan kapcsolva és visszahurkolva elkészült az általunk multi-Gauss-nak nevezett mindent felismerő hálózat.

A tapaszatalataink szerint az összes Gauss függvény használata messzemenően redundáns modell, ezért a [7]-szerinti „greedy"-algoritmussal a Gauss függvények számát a kezdeti közel 3000-ről 60-ra csökkentettük.

4.1 Fonéma-bigram hálózat

Az alternatív – jóval nagyobb számításigényű – mident felismerő hálózatunk a fontikailag változatos szöveganyagon tanított fonéma-bigram model volt, amiben szavakon átívelő „cross-word" trifón beszédhangmodelleket használtunk.

5 A vizsgálati adatbázis és az alkalmazott beállítások

5.1 A beszédadatbázisok

Tanítás és tesztelés céljára a legnagyobb magyar telefonos adatbázisokat használtunk (MTBA, a Besztel, a SpeechDat és a Tesztel) **[5]**. Ezek az adatbázisok elsősorban olvasott beszédet, valamint kisebb arányban spontán bemondásokat is tartalmaznak. Az első három adatbázis lényegében ugyanarra a szövegkorpuszra épül, és mindegyiknek az általunk elérhető része 500 beszélőtől tartalmaz hanganyagot. A Tesztel adatbázis 100 beszélős, és jellegzetessége, hogy szándékosan nagy és természetes háttérzajban felvett bemondásokat tartalmaz. Az adatbázisokban a vonalas és mobil telefonos felvételek összességében körülbelül ugyanolyan számban képviseltetik magukat.

5.2 Tanító- és teszthalmazok

Tanítás céljára az MTBA adatbázis 500 beszélőjének azon fonetikailag változatos mondatait és szavait jelöltük ki, melyek nem „o", és „z" jelzésűek, azaz nem tartal-

maznak tulajdonneveket és bizonyos típusú mondatokat. Ez összességében 6000 tanító felvételt eredményezett.

A 2475 felvételes izolált szavas teszthalmazunkat úgy állítottuk össze, hogy ne legyen átfedés a tanítóhalmazban szereplő beszélőkkel. Így a tanításnál fel nem használt felvételek összesen 170 beszélőtől (Besztel 100, Speechdat 50, Tesztel 20) kerültek a teszthalmazba.

5.3 Beszédfelismerési paraméterek, beállítások

Lényegkiemelési paraméterekként a bemenő beszédjelből MFCC (Mel Frequency Cepstral Coefficients) 38 dimenziós vektorokat képeztünk statikus energiát nem használva, de egyébként dinamikus Delta és Delta-Delta értékeket is számítva. Mind tanítás, mind tesztelés során alkalmaztuk a vak csatornakiegyenlítés módszerét.

Akusztikus modellként balról-jobbra struktúrájú, háromállapotú rejtett Markov-modelleket használtunk mind a monofón, mind a trifón beszédhangmodellek esetén. Állapotonként maximum 10 Gauss függvényből álló folyamatos megfigyelési sűrűségfüggvényeket használtunk. Szóhatárokon átívelő, azaz „cross-word" trifón modelleket alkalmaztunk.

6 Vizsgálat ROC görbék segítségével

6.1 A használt mérőszámok

Kísérleteink kiindulási alapja az egyszerű konfidencia-becslő ($\log P_{gr}$ és $\log P_{ca}$ különbsége). Ez az érték bár nem tartalmazza a helyes felismerés valószínűségét már alkalmas arra, hogy segítségével döntsünk a felismert szavak elfogadásáról, azaz beengedési küszöbként használható. Ezen küszöb változtatásával munkapontok jelölhetőek ki. A munkapont kijelölése után a küszöb alá került (elutasított), de egyébként helyesen felismert szavak számának (*KaH*), és összes helyesen felismert szó számának (*H*), valamint a küszöb fölötti félreismert (*KfF*) és összes félreismert szó számának (*F*) ismeretében a munkapontot jellemző két mennyiség definiálható (**8**).

$$FRR = \frac{KaH}{H} * 100[\%] \quad FAR = \frac{KfF}{F} * 100[\%] \tag{8}$$

A tévesen elutasított szavak arányát fejezi ki a False Rejection Rate (FRR), illetve a tévesen beengedettekét a False Acceptance Rate (FAR). A felismerés minősége általánosan jellemezhető az úgynevezett **Equal Error Rate** (**EER**) **[4]** mérőszámmal, ami FAR és FRR értéke annál a küszöbnél, ahol a két mennyiség megegyezik.

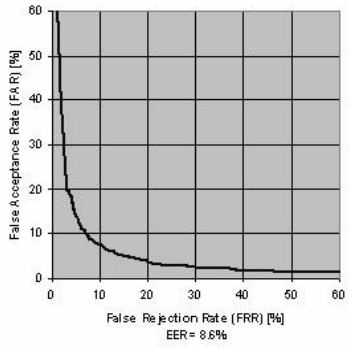
Az EER csökkentése egy fontos, de nem mindenek előtt való feladat a felismerési kísérleteknél. Ennek oka, hogy eltérőek lehetnek a felismerővel szemben támasztott elvárások. Például egy telefonközpontban használt dialógus rendszert érdemes alacsony FRR -tel rendelkező munkaponton üzemeltetni (pl.: 5%), mert nem kívánunk a helyes bemondásokat felesleges elutasításával kellemetlenséget okozni. Ilyen esetek-

ben fontosabb paraméter a kötött FRR -hez tartozó munkaponti FAR, ami nem feltétlenül a legkisebb EER -tel rendelkező felismerőben a legkedvezőbb.

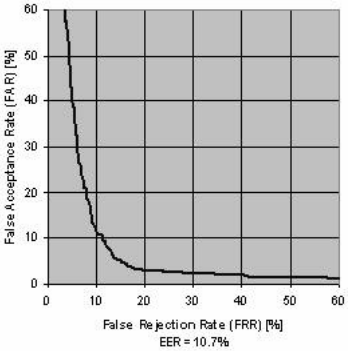
6.2 Az ROC görbék kiértékelése

Vizsgálatainkban két kísérleti elrendezést alkalmaztunk, ezeket az alkalmazott mindenhez illeszkedő hálózat különböztette meg (fonéma-bigram, multi-Gauss). Mindkét elrendezésre a korábban ismertetett két konfidencia-számító módszert alkalmaztuk, azonban számításuknál különbözőképpen definiáljuk az egyszerű konfidencia-becslőt. Alap esetben ez a hasonlósági mértékek logaritmusának különbsége (EKB_{alap}), normált esetben ugyanez a jellemzővektorok számával normálva ($EKB_{normált}$).

Az egyes munkapontok mentén összetartozó FRR és FAR értékeket egy grafikonon ábrázolva jutunk a **Receiver Operating Characteristic** (**ROC**) **[2]** görbéhez. Az ROC görbe jól szemlélteti a felismerő azon képességét, hogy milyen mértékben képes szótárban nem rögzített szavakat a szótári szavaktól szeparálni Vizsgáljuk először a két alapesetet! (**2. ábra**)



2.1. ábra Fonéma-bigram mindenhez illeszkedő hálózat, alap becslés



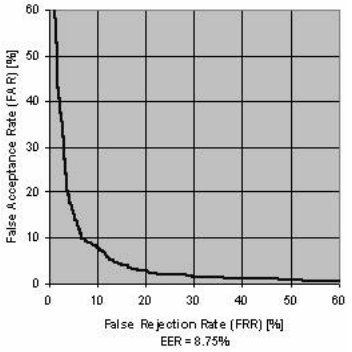
2.2. ábra Multi-Gauss mindenhez illeszkedő hálózat, alap becslés

A mindenhez illeszkedő hálózatok cseréje látványos változást idéz elő az ROC görbék alalkulásában. Multi-Gauss hálózat használtánál (**2.2. ábra, 10.7%**) lényegesen nagyobb EER mérhető, mint fonéma-bigram (**2.1. ábra, 8.6%**) modell esetén, de ennél is szembeötlőbb a görbe viselkedésének különbsége FAR tengely közelében. Ez azért különösen lényeges része az ábrának, mert a már korábban említett és eltertjedten használt FRR érzékeny alkalmazásokat az alacsony FRR értékkel rendelkező, tehát FAR tengely közelében lévő munkapontokon érdemes működteni.

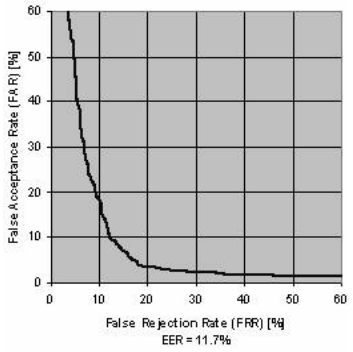
Összességben tehát az EER még alul is becsülte a különbséget ebben az esetben a két elrendezés között. Például, ha a kívánt munkaponti FRR=5%, akkor fonéma-bigramm esetén FAR=14.7%, multi-Gauss esetén 42.7%.

Az ROC görbéken végzett vizsgálataink alapján a fonéma-bigram mindentfelismerő hálózat alkalmasnak tűnik a gyakorlati alkalmazásra, bár számítási igénye kétségtelenül magasabb, mint a multi-Gauss-os változaté, de teljesítménye jóval felülmúlja azt, főként ha FRR érzékeny alkalmazásokat tekintünk.

Második típusú konfidencia-becslésünket alkalmaztuk mindkét hálózati konfigurációra, és ez újabb két görbét eredményezett. (**3. ábra**)



3.1. ábra Fonéma-bigram mindenhez illeszkedő hálózat, normált becslés



3.2. ábra Multi-Gauss mindenhez illeszkedő hálózat, normált becslés

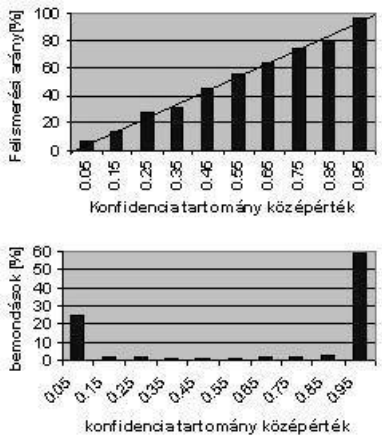
Ez az eset annyiban különbözik tehát az előzőtől, hogy a $EKB_{normált}$ érték alapján került az ROC görbe felrajzolásra. A szóhosszal való normálástól azt vártuk, hogy megbízhatóbb EKB-t szolgáltat majd, mely alapján a felismerő hatékonyabban szét tudja majd választani a szótáron belüli, illetve kívüli bemondásokat. Ezzel ellentétben mindkét esetben rosszabb EER volt mérhető. FRR=5%-os megkötésnél is romlottak a hozzá tartozó FAR értékek (15.8%, 46.7%). Kijelenthetjük tehát, hogy gyakorlati alkalmazásokban az ROC görbén végzett vizsgálat alapján nem érdemes a normált becslési módszert alkalmazni.

7. Konfidencia pontossági kiértékelés

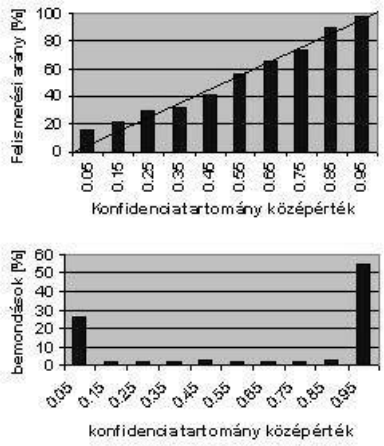
7.1. A felismerési arány grafikon

Ha a felismerés biztonságáról is szeretnénk információhoz jutni, szükség van arra is, hogy minden bemondáshoz egy hozzárendelési szabállyal (az általunk használt szabályt a cikk korábbi fejezeteiben ismertettük) egy konfidencia értéket társítsunk. Ez a fejezet a hozzárendelési szabály hatékonyságát próbálja értékelni. Ehhez szükséges bevezetni a részhalmazon mért felismerési arány fogalmát, ami azt mutatja meg, hogy a részhalmaz bemondásainak hány százalékát ismerte fel helyesen a felismerőnk. Fontos fogalom még a konfidencia tartomány, ami alatt két konfidencia érték közötti intervallumot értünk, amibe minden olyan bemondás beletartozik, amihez e két érték közötti valószínűséget rendeltünk. Minden ilyen tartományra értelmezhető egy felismerési arány is. Ha kiszámítjuk és ábrázoljuk ezt az arányt minden konfidencia tartományra, akkor jutunk a felismerési arány grafikonhoz. **[4]**

Önmagában ez az ábra még nem ad megfelelő értékelést a konfidencia-becslésről, mivel nem állapítható meg, hogy az egyes konfidencia tartományokba a szavak hány százaléka esett. Ha egy tartományban a felismerési arány jelentősen eltér a becsült konfidenciától, nem jelenti azt, hogy az egész becslés rossz volt, ha az adott tartományba a bemondásoknak elhanyagolható százaléka esett. Ezért kell kiegészíteni a felismerési arány ábrát egy grafikonnal, ami megmutatja, hogy az egyes konfidencia tartományokban a bemondások hány százaléka található (**4. ábra**).



4.1. ábra: Konfidencia-becsléshez tartozó felismerési arány grafikon, valamint a bemondások eloszlása a konfidencia függvényében. Fonéma-bigramm mindenhez illeszkedő hálózat, alapeset.



4.2. ábra: Konfidencia-becsléshez tartozó felismerési arány grafikon, valamint a bemondások eloszlása a konfidencia függvényében. Multi-Gauss mindenhez illeszkedő hálózat, alapeset.

7.2. Numerikus kiértékelés

Egy mindenhez illeszkedő hálózat felismerési arány grafikonja szemléletes, de sokszor célszerű az ábra által közölt információt egyetlen mérőszámba összevonni. Ilyen a **confidence error rate** (**CER**) százalékos mérőszám, ami megmutatja, hogy a konfidencia-becslés felismerési arány grafikonja hány százalékban tér el az ideálistól (**9**).

$$CER = \sum_{i=1}^{10} \frac{B_i}{B_f} * \left| F_i - C_i^k \right| * 100[\%] \tag{9}$$

Ahol B_f az összes bemondásnak a száma B_i pedig az i-edik konfidencia tartományba eső bemondások száma. F_i az i-edik tartomány felismerési aránya, C_i^k konfidencia tartomány középérték, pedig a tartomány alsó és felső határának számtani közepe.

7.3. Konfidencia pontosság a teljes vizsgálati halmazon

A legjobb konfidencia-becslést a korábban definiált „a” és „b” szabad paraméterek változtatásával CER minimalizálásával kerestük meg. Így kaptunk hálózatonként két konfidencia-becslést alap és normált esetre.

A két mindenhez illeszkedő hálózat CER értékei egy táblázatban kerültek összefoglalásra (**1. Táblázat**). Ezek alapján kijelenthető, hogy a fonéma-bigram mindenhez illeszkedő hálózat konfidencia-becslése hatékonyabb, ami a 4. ábrán is megfigyelhető, ha figyelembe vesszük, a 0-0.1-es konfidencia tartományt, ahol a multi-Gauss hálózat igen rosszul teljesít. Jól látható továbbá, hogy a normálás is ront konfidencia-becslés hatékonyságán.

1. Táblázat: A CER értékei az eredeti bemondáshalmaz konfidencia-becsléseire

	Fonéma-bigram	Multi-Gauss
Alap eset	2.2%	5.09%
Normált eset	3.1%	5.48%

7.4. Konfidencia-becslés pontosságának vizsgálata bemondáshalmaz szétbontással

Önmagában a teljes bemondáshalmaz vizsgálata nem elegendő, mivel az optimalizálás maga is ezen a halmazon történt. A további értékeléshez felbontottuk a bemondások adatbázisát több kritérium szerint diszjunkt bemondás-halmazokra. Először három, elemeit véletlenszerűen összeválogatott, körülbelül megegyező számú bemondást tartalmazó halmazra. A második kritérium a szótagszám szerint bontotta szét kis szavakra (1-2szótag), átlagos hosszúságú szavakra (2-4) és végül 5-nél több szótagot tartalmazó nagy szavakra. Az utolsó kritérium két csoportot hozott létre az alapján, hogy a bemondás szótáron belüli vagy kívüli.

A véletlenszerű felbontással képet kaphatunk arról, hogy az adott hálózat konfidencia-becslése (a teljes adatbázisra beállított „a" és „b" paraméterek itt már konstansok) mennyire adatbázis-független. A másik két módszerrel megfigyelhető, hogy a bemondott szó bizonyos tulajdonságai (hossza, szótáron belülisége) mennyire vannak hatással a konfidencia-becslésre.

Véletlenszerű szétbontás

A véletlenszerűen háromfelé bontott multi-Gauss és fonéma-bigram hálózatos becslések látványosan nem romlanak. A változás mértéke jó közelítéssel azonos a jellemzővektor számmal normált és nem normált becslés esetén mindkét mindenhez illeszkedő hálózat esetében. Az eredmények arra engednek következtetni, hogy a konfidencia-becslési paraméterek nem csak az optimalizációs adatbázishoz, hanem annak valamennyi részhalmazához is jól illeszkednek. (**2. Táblázat**)

2. Táblázat: A CER értéke a véletlenszerűen szétbontott bemondáshalmaz konfidencia-becslésére

	Fonéma-bigram (Alap / Normált eset)	Multi-Gauss (Alap / Normált eset)
1. Bemondáshalmaz	3.6% / 2.9%	5.7% / 4.4%
2. Bemondáshalmaz	3.3% / 3.1%	5.7% / 6.6%
3. Bemondáshalmaz	4.6% / 3.7%	6.7% / 6.3%

Szóhossz alapján történő szétbontás

A multi-Gauss hálózatos felismerőnél a szótagszám alapján történő szétbontása már változó képet mutat: a kis méretű szavak esetében a legnagyobb a változás. A legtöbb elemet tartalmazó közepes méretű szavak halmazán a romlás átlagos mértékű, a nagyméretű szavaknál javulás figyelhető meg. A normálásnak itt már volt látható hatása, kisméretű szavaknál javított a becslés hatékonyságán, 2-4 szótagos szavaknál

bár jobb nem lett, de kevésbé romlott az eredetihez képest. A fonéma-bigram típusú hálózat becslései csak a hosszú bemondásokra igazán érzékenyek. Az utóbbi csoporton egy keveset segített a normálás, sajnos annak árán, hogy az előbbi két csoport konfidencia-becslése rosszabb lett. (**3. Táblázat**)

3. Táblázat: A CER értéke a szóméret alapján szétbontott bemondáshalmaz konfidencia-becslésére

	Fonéma-bigram (Alap / Normált eset)	Multi-Gauss (Alap / Normált eset)
1-2 szótag	4% / 2.5%	6.8% / 6.9%
2-4 szótag	4.1% / 3.5%	5.8% / 5.6%
5- szótag	8.4% / 9.7%	4.5% / 3.9%

Szótári tartalom alapján történő szétbontás

A romlás minden esetben itt a leglátványosabb. A rendszer általában felismeri a szótáron belüli elemeket, néha mégis alacsony valószínűséget rendel hozzájuk. Ugyanez fordítva is igaz, több szótáron kívüli elem, túl nagy valószínűséget kap. Az ilyen jellegű hiba a nyelvi modell tulajdonságaiból ered, konfidencia-becsléssel nem javítható.

4. Táblázat: A CER értéke a szótári tartalom alapján szétbontott bemondáshalmaz konfidencia-becslésére

	Fonéma-bigram (Alap / Normált eset)	Multi-Gauss (Alap / Normált eset)
Szótáron belüli	10.2% / 9.2%	14.5% / 14.5%
Szótáron kívüli	15.5% / 15.9%	20.1% / 20.1%

8 Összefoglalás

Cikkünk folyamán egy az elméleti alapelvekre épülő megoldási lehetőséget mutattunk be a konfidencia-becslésére. Mint az látható volt, egy jellemzővektor sorozat előfordulási valószínűségének közelítése mindenhez illeszkedő hálózatok használatával célszerű és jó pontosságú becslést eredményez.

Eredményeinket áttekintve szembeötlő a két mindenhez illeszkedő hálózat alkalmazásakor jelentkező különbség. Multi-Gauss hálózat csak az ROC görbe FAR érzékeny részén volt képes a fonéma-bigram teljesítményét megközelíteni, ilyen alkalmazás azonban kevésbé gyakori. A konfidencia is jobban becsülhetőnek bizonyult a fonéma-bigram esetben, bár ez nem meglepő, mivel a konfidencia értéke EKB alapján számítódik. Ennek szótáron belüli, illetve kívüli szavak közti szeparációs képességét viszont az ROC segítségével vizsgáltuk. Az, hogy ezen vizsgálatok során multi-Gauss rosszabbul teljesített előrevetítette a rá adható pontatlanabb konfidencia-becslést. Ugyanezen szeparációs képesség hiányával magyarázható, hogy minden esetben magas CER érték adódott a bemondáshalmaz szótári tartalom alapján történő szétbontásánál. Ez mindenképpen javításra szorul a jövőben.

Fontos megjegyeznünk cikkünk egy kutatássorozat első eredményeit foglalja összse. Természetesen a konfidencia-számítási módszerünk egyszerűsége is okolható az a tapasztalható eltérésekért, ám magában rejti a továbbfejleszthetőséget, hiszen EKB-kön alkalmazott mostani lineáris transzformáció helyettesíthető magasabb szintű közelítésekkel. Sokkal meglepőbb inkább az a tény, hogy a módszer jelenlegi szintjén is már sok tekintetben pontos becslést ad.

A korábban elmondottak alapján több terület is kínál továbblépési lehetőséget. A mindenhez illeszkedő hálózatok tökéletesítése, a normált becslés átalakítása valamint összetettebb közelítések mind pontosíthatják az eljárást. Az ezek által kijelölt irányvonalak mentén kívánjunk mi is továbbfejleszteni felismerőnket. Célunk egy minden tekintetben pontosabb konfidenciát szolgáltató rendszer tervezése, ami nagy hatásfokkal képes detektálni a szótáron kívüli szavakat anélkül, hogy a szótáron belülieket elutasítaná. Egy ilyen eszköz várhatóan szükséges a valóban nyílt szótáras felismeréshez is, ami külön motiválja a kutatásainkat.

Bibliográfia

1. B. Dong, Q. Zhao, Y Yan: A fast confidence measure algorithm for continuous speech recognition (2005)
2. J. Pinto, R. N .V. Sitaram: Confidence measures in speech recognition based on probability distribution of likelihoods (2005)
3. J. Razik, O. Mella, D. Fohr, J.-P. Haton: Local word confidence measure using word graph and N-best list (2005)
4. G. Skantze: The use of speech recognition confidence scores in dialogue systems (2003)
5. Vicsi Klára et al: http://alpha.ttt.bme.hu/speech/databases.php (2005)
6. D.A.G. Williams: Knowing what you don't know: Roles for confidence measures in automatic speech recognition (1999)
7. Young, S. – Kershaw, D. – Odell, J. – Ollason, D. – Valtchev, V. – Woodland, P. The HTK Book (Version 3.0), (2000)

Látható beszéd: beszédhang alapú fejmodell animáció siketeknek

Feldhoffer Gergely[1], Bárdi Tamás[1]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar, 1083 Budapest, Práter utca 50/a, Magyarország
{flugi, bardi}@itk.ppke.hu

Kivonat: Elkészült egy fej-animációs rendszer, ami siketek számára beszédjelből olyan szájmozgást állít elő, hogy a siket felhasználó azt megérthesse. Egy olyan audiovizuális adatbázis készült el hozzá, amihez professzionális jeltolmácsok közreműködését kértük. Az animációs részhez szabványos MPEG-4 fejmodellt használtunk. Sikerült olyan reprezentációt találni főkomponens analízis segítségével, aminél egy neuronháló képes volt az akusztikus jellegvektorokból kiszámítani az animációs paramétereket. A rendszer fontosabb elemei mobiltelefonra is elkészültek. Siket felhasználókkal végzett teszt 50% körüli felismerési pontosságot mutatott ki a képi adatokból és a hangból számolt animációra is.

1 Bevezetés

Ennek a cikknek a témája egy olyan rendszer, ami siketeknek próbál segítséget nyújtani, hogy a kommunikációs lehetőségek egy újabb irányát használni tudják. A legtöbb siket ember gyermekkorában a halló emberekkel való kommunikálásnak legalapvetőbb módjaként a szájról olvasást tanulja meg. Emellett jeltolmácsok, illetve az írásbeli lehetőségek azok, amiket használni tudnak, hogy a halló emberek üzeneteit megértsék. A másik irány, a siket ember üzenete a halló emberekhez, szintén tanulható. A siket fiatalok megtanulnak bizonyos szintig beszélni, ami ugyan árulkodik a képességbeli elmaradásról, de kis tanulással illetve megszokással érthető. Ez azt jelenti, hogy a telefonos kommunikációnak csak az egyik iránya hiányzik, a halló ember üzenetét nem tudja fogadni a siket, de a siket tud beszélni. Ha tehát a beszédjelből olyan szájmozgás animációt tudunk előállítani, amit a siket meg tud érteni, akkor nincs akadálya a kétirányú kommunikációnak.

A szakirodalom szerint a szájmozgás hangból történő teljes pontosságú visszaállítása lehetetlen feladat, mert többféle hanghoz tartozhat ugyanaz a mozgás, illetve többféle mozgáshoz tartozhat ugyanazon hang. Amit ebben a cikkben bemutatunk, az nem mond ellent ennek a tézisnek. Ennek az az oka, hogy nem az a cél, hogy az eredeti szájmozgást számoljuk ki a beszédjelből, elegendő, ha az egyik olyan animációt sikerül előállítani, ami az adott beszédhang-sorozatra olyan mértékben jellemző, hogy a siket ember képes megérteni.

A rendszerünk egyik legfontosabb tulajdonsága, hogy elkerültük a diszkrét osztályozást, többek között a fonémákra vagy vizémákra való bontást, ezzel elvi akadálya nincs a nyelvfüggetlenségnek, az kizárólag az adatbázis összeállításán múlik. Létezik vizéma alapú rendszer [5][2], ami egy beszédfelismerő és egy arc szintetizátor összekötése. Ez a rendszer siketek számára nem elegendően élethű. A mi megoldásunk egyik fontos előnye, hogy megtartja a természetes ritmust, és nem vét hibás fonéma meghatározásból eredő hibát.

Szintén fontos tulajdonsága a rendszernek, hogy programozható mobiltelefonon megvalósítható módszerekre szorítkoztunk. Ennek az oka az, hogy a siket felhasználó nem szívesen használ olyan eszközt, ami felhívja a figyelmet a fogyatékosságára, a mobiltelefon viszont társadalmilag nem megbélyegző.

2 Adatbázis

2.1 Előzetes felmérések

A munka előtt felmértük, hogy milyen típusú kommunikációs formát tudnának a siketek megérteni és elfogadni. Ennek a felmérésnek már az elején világossá vált, hogy nem szívesen használnának olyan vizualizációt, amit tanulniuk kell, így a beszélő száj animációjában találtuk meg azt a formát, ami megoldható és használható is. Természetesen kényelmesebb lenne a siket embernek a szöveggé alakítás, mert szájról olvasni kimerítő tevékenység, de folyamatos beszéd felismerése telefonon keresztül máig egy igen nehéz, nagy adatbázisokat és számítási kapacitást igénylő feladat, ami programozható telefonon a mai teljesítmény mellett nem reális cél.

A feladat rögzítése után felmértük, hogy milyen paraméterek befolyásolják a beszélő emberről készült felvételek érthetőségét. A felmérések azt igazolták, hogy egy mobiltelefon méretű kijelzőn, aminek a felbontása 320x208 pixel, érthető egy olyan felvétel, amin csak a száj környéke látható. Ugyanakkora felismerési arányt tapasztaltunk nagyobb felbontású, vagy nagyobb kijelzőkön, és időbeli felbontásban elegendőnek bizonyult 12 kép/másodperc is. Ez azt jelenti, hogy a mobiltelefonon megvalósított rendszer érthetőségét a készülék technikai paraméterei nem korlátozzák.

Ennek a felmérésnek a legfontosabb tanulsága az volt, hogy az érthetőség minden technikai paraméternél jobban függ attól, hogy a beszélő mennyire képes igazodni a siketek igényeihez. Azt a felvételt, amin a beszélő már sokat foglalkozott siket emberekkel, a legrosszabb technikai körülmények között is majdnem pontosan értették, míg a gyakorlatlan beszélőket a felszereléstől függetlenül kevéssé.

Egy másik felmérésben arra voltunk kíváncsiak, hogy a felismerésben számít-e az, hogy a felvételen látható arc árnyalataiból az eredeti háromdimenziós fej rekonstruálható. Az erre a kérdésre adott válasz dönti azt el, hogy szükség van-e a száj mozgásának háromdimenziós rögzítésére, vagy elegendő kétdimenziós, pontkövető vagy kontúrkereső algoritmusok használata. A felvételeket torzítottuk, hogy eltűnjenek az árnyalatok, bizonyos felvételeken csak a szájkontúr volt kivehető. A felismerendő szavak véletlen kétjegyű számok voltak. A siketek a torzított videókat ugyanolyan pontosan megértették, mint az eredetieket. Ezért használtunk a későbbiekben olyan

rögzítési eljárást, ami egyetlen kamerának a képéből dolgozva nyeri ki a száj állapotát.

A teszteléshez fontos a siketek kontextus követésének elemzése. Írásbeli és írásos-mozgóképes vegyes teszteket végeztünk, ahol meggyőződtünk arról, hogy annak ellenére, hogy írásban sok nyelvtani hibával kommunikálnak, a kontextust jól megértik, történetek szereplőinek cselekményeit akkor is követni tudják, ha az nyelvtanilag csak a toldalékokból derül ki, holott a toldalékokat írásban nem használják. Kiderült az is, hogy a siketek szókincse szűk, például, amikor egy állatról szólt a szövegkörnyezet, akkor a „gebe” szót azért nem ismerték fel, mert nem tartozik a szókincsükhöz. A feladat szempontjából ez nem okoz problémát.

2.2 MPEG-4

A világban sokféle arc-animációs rendszer terjedt el, többségükben az MPEG-4 szabvány szerint paraméterezhetőek. Ez a szabvány rögzíti az arc főbb pontjait (feature point, továbbiakban tartópont), azokat képes normálni, és így arc állapotokat lehet különböző alakú fejre illeszteni. Léteznek MPEG-4 szabvány szerint működő szövegfelolvasó rendszerek, amik egy beszédszintetizátor és egy fejmodell párhuzamos meghajtásával működnek. Munkánkhoz a Cosi és társai [3] által létrehozott Lucia fejmodellt használtuk.

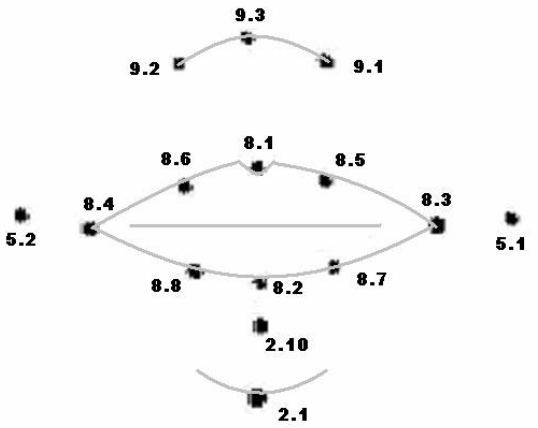


Fig. 1. Az MPEG-4 általunk is használt tartópontjai.

A tartópontokon kívűl létezik az MPEG-4-ben magas szintű animációs leírás is, ami vizémákra alapul, ami a fonémákhoz tartozó száj-állapotokat jelenti. Ezeket nem használtuk, mert a rendszerünk nem dolgozik fonémaszinten.

2.3 Az adatbázis rögzítése

Az előzetes felmérések tanulsága szerint az adatbázist hivatásos jeltolmácsok szereplésével készítettük. A száj körüli tartópontokat kellett tehát adatbázisban rögzíteni. Ezeket olyan videó felvételekkel készítettük el, ahol a modell arcán megjelöltük az FP pontokat.

A videó felvételt automatikus módszerekkel dolgoztuk fel, ami a következő lépéseket jelenti: színkiemelés, binarizálás, dilatáció, erózió. A kapott eredményeket manuálisan javítottuk, ahol az automatikus módszer hibázott. Ezek a helyzetek jellemzően a felpattanók gyors mozgása, vagy a csücsörítés megváltozott fényviszonyai miatt léptek fel.

A felvétel közben jó minőségű hangfelvétel készült (48kHz, 16 bit), amit szinkronizáltunk a videó felvétellel. (Fig. 2)

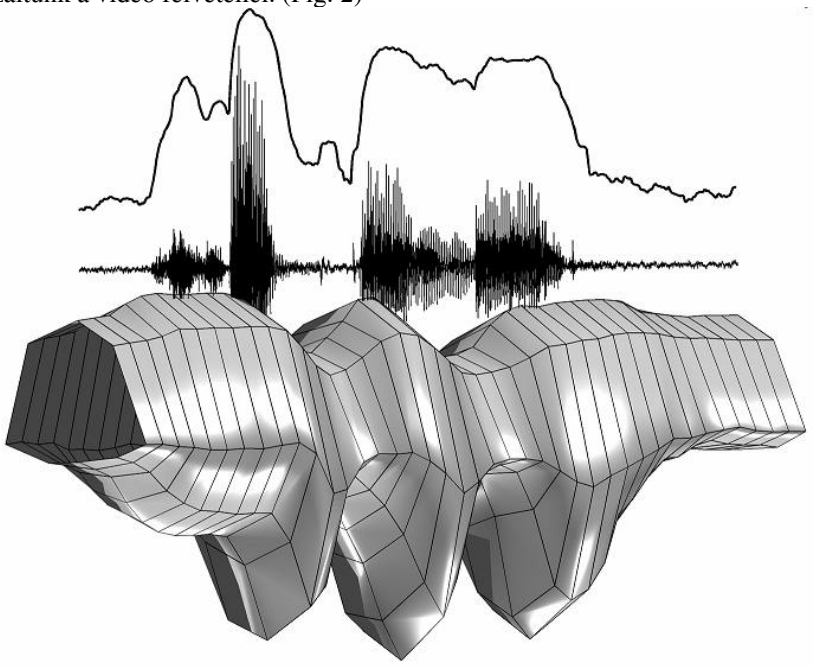


Fig. 2. A „szeptember” szó az adatbázisban. Az ábrán az energia, a beszédjel, és a száj külső kontúrja látható az időben.

A beszédjelen a videó sebességéhez igazított ablakozást használtunk, a PAL rendszer 25 kép/másodperc sebességénél ez 40ms. Így minden képhez tartozott egy hangszakasz. Ez a hangszakasz 48kHz mintavételezés mellett 1920 mintát jelent. Ezen szakaszhosszból kiválasztottuk a maximális kettőhatványt, ez 1024 jel, amin Radix-2 FFT algoritmust futtattunk Hamming ablakkal. Az eredményt mel skála szerint 16 sávban összegeztük, majd cosinus transzformációval kiszámítottuk a cepstrumot (MFCC).

A telefonon futó alkalmazásnál 8kHz mintavételezés mellett az ablakméretek 320-nak illetve 256-nak adódnak. A magasabb mintavételezéssel tanított hálózatot használni lehet az alacsonyabb mintavételezésű készülékeken a mel skála igazításával.

A videó felvételből nyert pontokon lényegkiemelést végeztünk főkomponens analízissel (1). Az első 6 főkomponens használatával az eredeti adatok 1-2%-os hibával történő rekonstruálása lehetséges, ami átlagosan 1 pixel elmozdulást jelent a PAL szabványú képen. Ez a hiba elhanyagolható, hiszen 720x576 pixeles felbontás mellett van tartópont, ami 120 pixelnyi tartományban mozog.

$$w_{1\ldots6} = P^{-1}B\Big|_{p_1^{-1}\times..\times p_6^{-1}} \tag{3}$$

A 16 dimenziós hangi és a 6 dimenziós képi adatokat neuronháló tanításával kapcsoltuk össze. A neuronháló bemenete 5 egymást követő MFCC ablak, a kimenete pedig a középső ablakhoz tartozó képkocka főkomponens-térben adódó koordinátái. A neuronháló 40 rejtett neuront tartalmaz.

Felmerül a kérdés, hogy az itt használt 40ms hosszú ablak alkalmas-e a beszédből szájmozgássá alakításhoz. A beszédfeldolgozó rendszerek mindig rövidebb ablakot használnak. A beszédre jellemző szájmozgásnál a külső szemlélő azt a hatást látja, amit a szájmozgató izmok okoznak. Ezek az izmok jóval lassabban képesek csak mozogni, mint a nyelv, vagy más, a hangot befolyásoló izmok. Még ennél is fontosabb, hogy megkülönböztethetünk fonémákat dominancia szempontból. [4] Azt nevezzük domináns fonémának, ami meghatározza a száj állását, ilyenek a magánhangzók és az ajakhangok. A domináns fonémák a szomszédos nem domináns fonémákra eső szájállást is befolyásolják. A rendszer 5 egymást követő szakasszal dolgozik, melyek összesen 200ms időtartamot fognak át. Ez az átlagos fonémahosszak alapján valószínűsíti, hogy az 5 ablak közül legalább egy domináns fonéma tiszta fázisába essen. (Fig. 3) Ez a neuronhálózat tanulása szempontjából már elegendő. A 200ms késés nem jelent problémát, mert az információ csak ebben a modalitásban érkezik, és a 200ms belül van az emberi dialógusok toleranciahatárán.

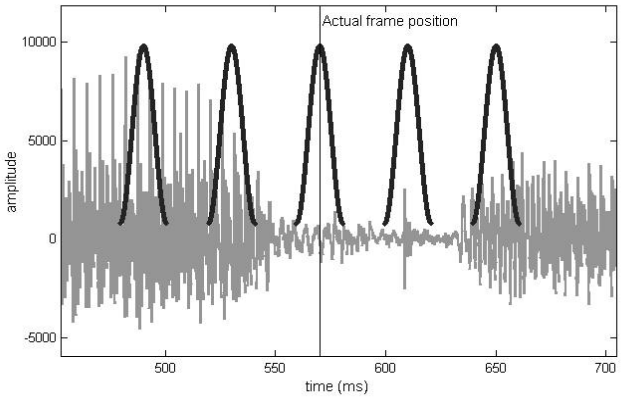


Fig. 3. Az öt egymást követő ablak. Két ablak kezdőpontja között 40ms a távolság.

2.4 Az adatbázis tartalma

Az adatbázisban szerepelt egy rövid általános tanító rész, ami a magánhangzókat egyenlő számban tartalmazta. Az adatbázis nagyobbik részét olyan anyag teszi ki, ami lehetőséget ad közvetlen tesztelésre siket felhasználókkal. A rendszer minőségének a mérésére ugyanis úgy kerülhet sor, ha a felhasználók látnak olyan szintetizált képet is, amit az adatbázis képi anyagából nyerünk, és ennek a lehető legjobb minőségű módja az, ha az adatbázisban rögzítünk olyan frázisokat, amik a tesztelés során egyben felhasználhatóak. A teszt során olyan szituációt modelleztünk, amiben figyelembe vettük a siketek jó kontextus követését, ezért megszorítottuk az anyagot olyan tartalomra, mint egy- és kétjegyű számok, hónapok nevei, hét napjai. Ezek folyó szövegben mindig úgy helyezkednek el, hogy a kontextusból tudható a halmaz, de az, hogy melyik elem a több közül, az bizonytalan.

3 A rendszer

A kész rendszer főbb alkotóelemei a hangkezelő rész, ami a telefon mikrofonját olvasva elvégzi az adott hosszúságú ablakokra vágást, a jellegvektorok kiszámítása, neuronháló, főkomponens analízis, és fejmodell szintetizálás. (Fig. 4)

Az MFCC adatok előállításához Radix-2 FFT algoritmust használunk, ami egy programozható mobiltelefonon is alkalmas valósidejű elemzésre.

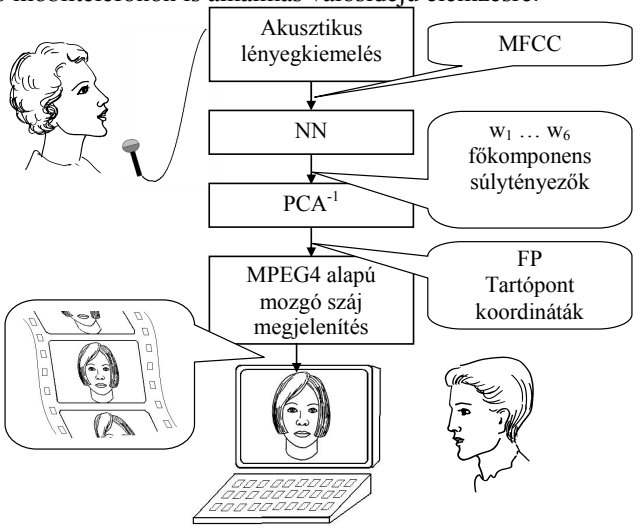


Fig. 4. A rendszer áttekintése.

3.1 Neuronháló

A tesztelt rendszer neuronhálózata 80 bemenettel rendelkezik, ami 5 keret, egyesével 16 MFCC együttható. A kimeneten 6 érték jelenik meg, ami az adatbázis adataiból nyert főkomponensek első 6 összetevő által kifeszített térben 6 koordináta. A neuronhálózat egy hagyományos hiba visszaterjesztéses neuronháló (backpropagation), aminek a számítási kapacitását egy mátrixszorzásra alapuló technikával gyorsított fel Davide Anguita. [1] Ez a neuronháló 1 000 000 epoch tanítás után került tesztelésre siket felhasználókkal.

3.2 PCA

A főkomponens analízisnek több előnye van. Elsősorban a rendszer működőképességének szempontjából fontos, hogy komolyabb adatveszteség nélkül dimenziót lehet csökkenteni. Ez esetünkben a 30 dimenziós pixeltér és a 6 dimenziós főkomponensek által kifeszített tér közötti különbség. A PCA használata előtt futtattunk neuronháló tanításokat pixeltérben is, és több mint 300 millió epoch után sem volt használható a hálózat kimenete. Az ilyenformán tanított neuronhálózat is egyetlen szabadsági fok mentén mozdult el: az állkapocs nyílása. A probléma az volt, hogy neuronhálózatnak kellett megtanulnia azt is, hogy a száj körüli pixelek általában hogyan helyezkednek el, és az együtt mozgásuk általában milyen. Ezt a terhet veszi le a neuronhálóról a PCA térben felírt száj állapot leírás. A PCA koordinátarendszerében ugyanis az általános mozdulat-összetevők már megvannak, egyszerűen ezek együtthatóival reprezentáljuk az állapotot.

A másik előnye a PCA-nak, hogy igen egyszerű használni. A pixeltérbe való visszaszámolás egy konstans mátrixszorzással megoldható. (2)

$$\overline{B}_k = \left(\underline{w}_k + \underline{c}\right) \cdot P \tag{4}$$

Matlab kódot készítettünk, ami a gyűjtött adatbázisból automatikusan kiszámítja a főkomponenseket, átírja az adatbázis pixeltérben rögzített értékeit PCA térbe, és elkészíti azt a C kódrészletet, amivel a visszakódolás elvégezhető. Azért használunk generált kódot kimentett értékek fájlkezeléssel való feldolgozása helyett, mert a mobil platformon így jóval egyszerűbb dolgozni.

A PCA további lehetőségeket is rejt. Ezek közül a legérdekesebb az, hogy a PCA alkalmas arra, hogy objektív véleményt formálhassunk az adatbázishoz készített felvétel minőségéről az olvashatóság tekintetében. Azt figyeltük meg, hogy a siketekkel dolgozó, a szájmozgására tudatosan figyelő és rutinos személyek főkomponensei jól megkülönböztethetőek a képzetlen szereplőkétől. A különbség a főkomponensek sorrendjében van. A főkomponens analízis ugyanis fontossági sorrendbe állítja a főkomponenseket, az első főkomponens az az iránya a sokdimenziós pontfelhőnek, amerre a szórás a legnagyobb. Esetünkben az állkapocs nyitásához kapcsolódó főkomponens az, aminek a súlya lényegesen megelőzi a többit, a szórás körülbelül 70%-áért felelős. Ebben minden szereplő közös: az első főkomponens mindig ez. A másodiktól a negyedikig azonban eltér a sorrend. A képzett, professzionális beszélő főkomponensei mind vizéma megkülönböztető szerepet töltenek be, olyan mozdulatokat, mint a száj széthúzása („csííz”), vagy a csücsörítés. (Fig. 5) A képzetlen beszélőknél igen komoly helyezést érnek el, a vizémákat nem megkülönböztető, beszéd-

szokásokhoz, emóciókhoz kapcsolható mozdulatok, mint például a száj tekerése, oldalra elhúzásával nyitása.(Fig. 6)

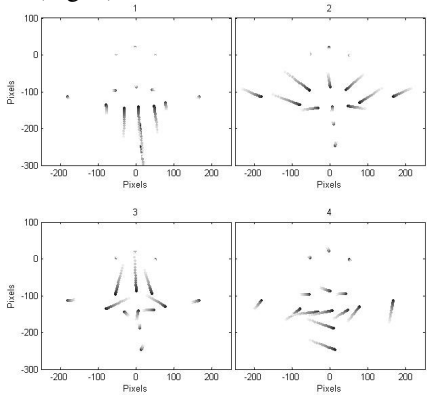


Fig. 5. Profi beszélő főkomponensei. Látható az első három mozdulatkomponens vizéma megkülönböztető szerepe.

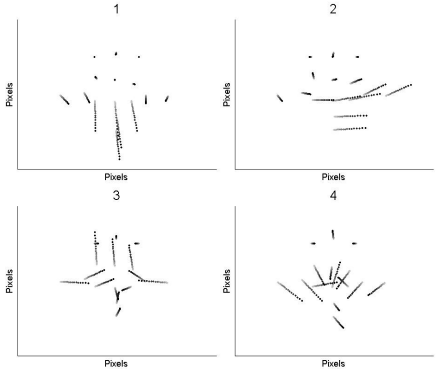


Fig. 6. Gyakorlatlan beszélő főkomponensei. Már a második főkomponens

3.3 Fejmodell

A folyamat végén a fejmodell áll, ami megkapja azon pixelek koordinátáját, ahol a tartópontoknak lenniük kell a 2 dimenziós felvételen. Ebből ki lehet számolni a 3 dimenziós MPEG-4 tartópontokat.[6] Ez úgy lehetséges, hogy a legtöbb mélységi komponenst elhanyagoljuk, csak az állkapocs hátramozdulását számítjuk bele a koordinátákba, illetve, hogy a száj normál állapotától vett távolságokból fejezzük ki az egyes tartópontok helyét. Az így kapott tartópontok alapján az arc szintézise azon alapul, hogy értelmezünk minden tartópont köré egy hatókört, ami azt a bőrfelületet reprezentálja, amit a tartópont magával húz, amikor mozog. Ez a technika a száj vonalánál igényel körültekintést, a hatóköröket úgy kell megadni, hogy az alsó és a felső ajak között ne legyen átfedés, különben a száj belső kontúrja nem nyílna ki. Az állkapocsnál is van egy kis probléma: az állkapcson elhelyezett tartópontnak az egész állkapcsot mozgatnia kell, de ez nyers formában olyan rajzfilmszerű mozgást eredményez, mintha az állkapocscsont függőlegesen eltolódva nyitná a szájat, nem pedig

elfordulva. Ezt a problémát úgy oldottuk meg, hogy tettünk egy el nem mozduló tartópontot az állkapocs tengelyének két oldalára. Ezzel elértük, hogy a súlyozott elmozdulások eredője jól imitálja az állkapocs lefelé elfordulását.

4 Eredmények, fejlesztési lehetőségek

A rendszer által előállított animációkat bemutattuk siketeknek. Kontrollként valódi videó felvételt is mutattunk, illetve azt a fej-animációt is, aminek a paramétereit nem a hangból, hanem egyenesen az adatbázisban rögzített mozgásból állítottunk elő. Ezt az tette lehetővé, hogy az adatbázisban tárolt paraméterek és a fejmodell paramétereit igyekeztünk egyformára csinálni. A arcra felfestett pontok pontatlanságát utólag korrigáltuk, hogy a fejmodell mozgása a lehető legpontosabban kövesse az eredeti felvételt.

Összesen 70 rövid mozgóképet mutattunk, egy- és kétjegyű számokat, hónapneveket és a hét napjait. A mozgóképek között volt valódi felvétel, az adatbázishoz készült felvétel egy-egy részlete. Volt a képi adatok felhasználásával meghajtott szintetizált fej, és volt hangból számított fejmozgás. A valódi videó felismerési aránya 97% volt. Annak az animációnak, ami a képi adatok felhasználásával készült, 55%-os lett. A hangból számított animáció felismerése 48%-os volt.

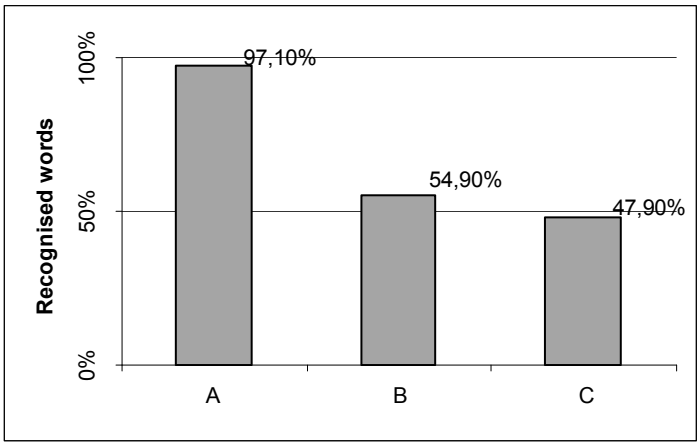


Fig. 7. Felismerési pontosságok a különböző mintákon: eredeti felvétel professzionális jeltolmáccsal (A), szintetikus arc, aminek a paraméterei az adatbázis képi adataiból származik (B), illetve hangból számolt szintetikus arc (C)

Ebből az a következtetés vonható le, hogy a fejmodellünk részletessége a fő probléma. Nyilvánvaló hiányosság, hogy nincs információ a száj belső kontúrjáról, a fogak és a nyelv láthatóságáról. Ezeket a paramétereket nem lehet pontok felfestésével megoldani, ezért a jelenlegi kutatás ebben az irányban is zajlik, új képfeldolgozó eljárásokat vizsgálunk meg. Ugyancsak kutatás folyik abban az irányban is, hogy hogyan lehet összekapcsolni több személy adatait, a jelenlegi rendszer ugyanis csak egy ember hangjának és szájmozgásának összekapcsolását tudja megtanulni.

Köszönetnyilvánítás

Ezúton köszönjük az együttműködést a Siketek és Nagyothallók Országos Szövetségének, és a siketeknek, akik segítettek tesztelni, a T-Mobile-nak, Dr Takács Györgynek, Tihanyi Attilának, Srancsik Bálintnak, és Harczos Tamásnak.

Bibliográfia

1. Anguita D.:Matrix Back Propagation – An efficient implementation of the BP algorithm. Technical report, DIBE – University of Genova (1993)
2. Beskow J.: Talking Heads, Model and Applications for Multimodal Speech Synthesis: Doctoral Dissertation, Stockholm (2003)
3. Cosi P., Fusaro A., Tisato G.: Lucia a New Italian Talking Head Based on a Modified Cohan-Massaro's Labial Coarticulation Model. Proceedings of Eurospeech 2003, Geneva, Switzerland (2003) 2269–2272
4. Czap L., Mátyás J.: Virtual Speaker, Híradástechnika, selected papers (2005) 2-5
5. Granström B., Karlsson I., Spens K-E.: SYNFACE – a project presentation, Proc of Fonetik, TMH-QPSR (2002) 93-96
6. Takács Gy., Tihanyi A., Bárdi T., Feldhoffer G., Srancsik B.: MPEG-4 modell alkalmazása szájmozgság megjelenítésére, Híradástechnika 2006 8.

VII. Pszichológiai szempontú szövegfeldolgozás

A személy- és csoportközi értékelés pszicholingvisztikája

Bigazzi Sára[1], Csertő István[2], Alessio Nencini[3]

[1] Pécsi Tudományegyetem, Pszichológiai Intézet
Ifjúság útja 6 Pécs 7624
bigazzisara@hotmail.com
[2] Pécsi Tudományegyetem, Pszichológiai Intézet
Ifjúság útja 6 Pécs 7624
csertopi@gmail.com
[3] Universitá di Padova, Dipartimento di Psicologia Sociale
alessio.nencini@unipd.it

Kivonat: Jelen kutatási projekt az NKFP 6/074/2005 számú pályázat támogatásával készült, melynek célja a szövegek automatikus elemzésének kidolgozása különböző pszichológiai dimenziók mentén. Munkánk célja egy olyan szövegelemző program kidolgozása, amely képes a történetekben kifejezésre kerülő értékelések automatikus azonosítására. A szöveg egy társadalmi valóságot tár elénk, melyben, ahogy a történet halad előre, személyközi és csoportközi kapcsolatok születnek, állnak fenn és változnak meg a szereplők között. Az értékelések ezen kapcsolatok dinamikus építőelemei.

1 Elméleti háttér

A Jerome Bruner által megalapozott narratív pszichológia elméleti keretei szerint az emberi tapasztalat narratív formákba szerveződik [1], [2]. Ahogyan azt Bruner kifejti, a narratív pszichológiának és a szociális reprezentáció konstrukcionista elméletének léteznek közös pontjai. Bruner szerint a szociális reprezentációk narratív szerveződésükön keresztül jönnek létre és terjednek el. Ebből a perspektívából a narratívum vagy bármilyen történet már a nyelv előtt létezett, természetes eszközt nyújtva az embereknek, hogy világnézetüket megkonstruálják és társadalmi valóságukat megértsék. Ennek következménye, hogy egy szociális reprezentációnak a megosztása és megvitatása kötelezően tartalmának narratív szerveződésén keresztül történik. Így a kommunikáció folyamata gondolatmenetünk kiindulópontja: az emberek a kommunikáción keresztül konstruálják és osztják meg tapasztalataikat és reprezentációikat. Így a nyelven keresztül, társadalmilag elismert szabályainak és konvenciónak felhasználásával, az emberek képesek világnézetüket egyértelműsíteni és érthetőbbé alakítani. Ennek következménye, hogy a nyelvre és konvencióira építünk, amelyeket az emberek arra használnak fel, hogy közösen elfogadott jelentéseket hozzanak létre.

Ez az elméleti keret lehetőséget ad a személyes és szociális identitás többarcú fogalmának vizsgálatára is. A személyes identitás tekinthető úgy, mint egy biográfia,

amely múltbéli tapasztalatokat gyűjt össze, és a jelentől függően folytonosan újraírja őket [3]. A szociális identitásnak, hasonló módon, számolnia kell a csoport múltjával. A múlt és a jelen elemeinek szervezésében és strukturálásában az időbeli komponens alapvető: kontinuitást és koherenciát nyújt a reprezentációnak [4], [5]. Ez a reprezentáció, mely a csoport legrelevánsabb elemeiből, gyakran történelmi elemeiből épül fel, egy szimbolikus rendszert jelent, amelyet a csoporttagok más emberekkel való mindennapi kapcsolataikban, identitásuk – összehasonlításon és értékelésen keresztüli – tárgyalásában használnak fel [6].

Ebben a kontextusban kell értelmeznünk az identitás és az értékelés közti kapcsolatot is. Úgy gondoljuk, hogy sem a személyes, sem a szociális identitást nem értelmezhetjük úgy, mint statikus és változhatatlan entitásokat. Ellenkezőleg, az identitás a más emberekkel folytatott kommunikáción keresztül egyfolytában újradefiniálódik és rekonstruálódik, és ebben a folyamatban a pszichológiai értékelés alapvető szerepet játszik.

Mindannyian rendelkezünk egy olyan értékrendszerrel, melynek kategóriái vagy dimenziói mentén megítélünk egy másik embert. Az értékelésnek ebben a folyamatában a másikra ruházott tulajdonságok a róla alkotott reprezentációnk szerves részévé válnak, és e reprezentáció meghatározó szerepet kap a másik viselkedésének észlelésében, a rá irányuló illetve őt érintő kommunikációs aktusaink és egyéb cselekedeteink megtervezésében, kivitelezésében, valamint hatással van az önmagunkról alkotott reprezentáció szerveződésére és tartalmára, beleértve értékeink rendszerét. Az értékelés így kizárólag egy relációs perspektívában létezik és létrehozott elemei csak e kapcsolaton belül értelmezhetőek.

1.1 Az értékelés pszicholingvisztikai modellje

Semin és Fiedler [7], [8] nyelvi kategória modelljéből (LCM) kiindulva igyekeztünk megragadni a narratívában megjelenő pszichológiai értékelést. Az értékelés pszicholingvisztikai modelljét referenciarendszerként kezeltük a további kategorizálásokhoz.

A modell két alapdimenzióra épül:

- Az első „Konkrét-absztrakt” dimenzió Semin & Fiedler nyelvi kategória modelljéből származik [7], [8], és a különböző absztrakciós szinteknek felel meg, amelyeket az emberek cselekvések és történések leírásánál használnak fel. A mi modellünkben ez a dimenzió az értékelésben megjelenő különböző absztrakciós fokozatoknak felel meg, mivel az értékelt személyre való utalás kifejezhető fizikai tulajdonságokkal vagy konkrét cselekvések leírásával, figuratív képekkel vagy absztrakt fogalmakkal. A konkrétság erősen kapcsolódik az értékelés alatt álló személy észlelhető tulajdonságaihoz és helyzeti beágyazottságához.
- A második „Explicit-Implicit” dimenzió az értékelést kifejező kommunikáció megértéséhez szükséges közös tudásra utal. Minél explicitebb egy értékelés, annál jobban dekontextualizált jelentéssel bíró és általános tudásra épülő szavakkal és kifejezésekkel lesz megformálva. Ez alacsonyabb szintű értelmezést és kevesebb inferenciát igényel a befogadó részéről. Egy implicit értékelés magasabb szintű háttértudást igényel. Ebben az esetben az értelmezés folyamata erősen kontextus függő lesz.

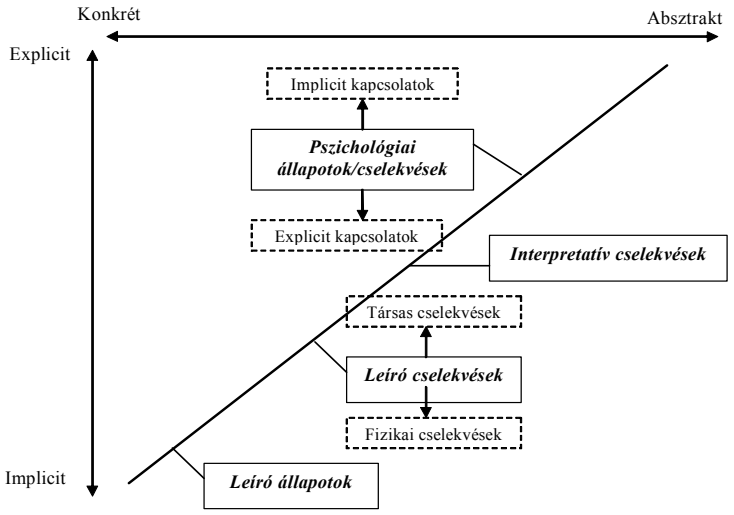


Fig. 1. Az értékelés pszicholingvisztikai modellje

A kapott diagonális a kontextualizáltság versus általánosság fokát jelenti. Az implicit és konkrét pólusból kiindulva a *leíró álapotokat* találhatjuk meg. Ez a kategória azokat a nyelvi állításokat tartalmazza, amelyekben az értékelés az értékelt fizikai leírásain és szerepein keresztül jön létre. Ezek a nyelvi formák, amennyiben tartalmaznak értékelést, erősen kontextualizáltak és megértésükhöz több következtetésre van szükségünk. Az értékelő nem vállal egy aktív és határozott pozíciót, inkább egy távoli és konkrét perspektívát választ. Nem direkt és explicit minőségeket fogalmaz meg, hanem többé-kevésbé megosztott társadalmi kategóriákat, szerepeket, metaforákat, konkrét képi leírásokat, amelyekkel ugyanakkor az értékelő hovatartozásából adódóan attribútumokat, normákat és értékeket is kommunikálhat. Az értékelésnek ezt a nyelvi kategóriáját talán a legnehezebb azonosítani kijelentéseinek kontextusfüggősége és implicit jelentésekkel való telítettsége miatt.

- Anna egy szőke nő.
- A cigány fiú megfürdött a patakban.
- Az alacsony szicíliai a szembefújó szélben összegörnyedve ment tovább.

A *leíró cselekvéseket* az értékelés pszicholingvisztikai modelljének második szintjén találjuk. Az értékelés itt egy konkrét és fizikálisan megjelenő cselekvés, eset leírásán keresztül történhet. Az ilyen igék használatát az a szándék vezérli, hogy a beszélő egy adott, megfigyelhető eseményt írjon le, megtartva annak perceptuális jellegét. Ebben az esetben az esetleges értékelés szintén implicit és értelmezéséhez szükséges jelentésének kontextusba helyezése.

- János átöleli Máriát
- Péter elrohant.

A cselekvés interperszonális mivoltától függően két alkategóriába osztható. Ahogy, az ábrán is látszik, fizikai cselekvéseknek hívjuk azokat a kifejezéseket, ame-

lyekben nincs egy explicit kapcsolat. Társas cselekvéseknek nevezzük azokat a cselekvéseket, amelyek egy explicit kapcsolatot is kifejeznek (pl.: János átöleli Máriát).

Az *interpretatív cselekvések* kategóriája olyan cselekvéseket tartalmaz, amelyek elvesztették a cselekvés konkrét és perceptuális elemeit. Ezeknél az értékelő állításoknál a használt igék figuratívak: a cselekvéseket egy absztrakt síkon írják le. Az absztraktság következménye, hogy egy morális tartalmat és értékelést közvetíthetnek használatukban. Az interpretatív igék nem megfigyelhető cselekvéseket írnak le, a befogadónak kell értelmezni a kijelentésben rejlő értékelés tartalmát.

- Józsi bántotta Gabit.
- Péter udvarol Annának.

A *pszichológiai állapotokhoz* olyan nyelvi elemek tartoznak, amelyek egy egyén belső állapotát írják le. Az értékelések itt az értékelt személy érzelmeiről való nyilatkozások. Ezek a pszichológiai állapotok absztraktabb és dekontextualizáltabb helyzetekről számolnak be, mint az előző nyelvi elemek. Ebben az esetben egy empatikus perspektívát kap a befogadó. Az értékelés az értékelt érzelmi állapotának átélésében keletkezik. Ez egy szubjektív értékeléshez vezet, amely nem a személy diszpozíciós jellemzői alapján történik, ezért még mindig változtatható és helyzeti.

- Sándor utálja azt az embert.
- Lóri boldog.

Ebben az esetben az értékelő sajátos szándékától függően szintén kétfajta alkategóriát találunk: a pszichológiai állapotok önállóan (pl.: Lóri boldog), vagy egy kapcsolati helyzetben vonatkozhatnak az értékelés tárgyának érzelmi, tudati állapotára (pl.: Sándor utálja azt az embert). Ezek az értékelések még nagyobb mértékben függetlenednek a szituációtól, és lehetőséget adnak a befogadó számára, hogy identifikálódjon az értékelés tárgyával.

Az *értékelő állapotok* kategóriája a legabsztraktabb és legexplicitebb kifejezéseket tartalmazza, amelyekben megjelenhet egy személyközi értékelés [9]. Az értékelő állapotok a leggyakrabban használtak attitűdök kifejezésénél. Általánosítják az értékeléseket különböző eseményeken keresztül, és csak az értékeltet írják le. A pszichológiai állapotoktól eltérően az értékelő állapotok bizonyos érzelmi távolságból kezelik az értékeltet: olyan társadalmilag elfogadott normák és értékek szerint minősítik, amelyek fontosak az értékelő számára.

- Géza agresszív
- Helyes az a lány

Az értékelő állapotok az értékelés tárgyával kapcsolatos általánosításokat fejeznek ki. Ezek az értékelések nagyon hatékonyak, mivel olyan elterjedt és jelentőségteljes tulajdonságokon keresztül írják le az értékeltet, amelyek esetében magas szintű társadalmi konszenzus feltételezett.

1.2 Értékstruktúra és értékelések

Az értékelést kifejező nyelvi forma konnotációja nagyon gyakran társadalmilag meghatározott. Amikor valakit értékelünk, egy pozíciót jelölünk ki számára egy vagy több tulajdonság mentén. Az értékelés tekinthető úgy, mint az értékelő értékeinek és

vélekedéseinek tudatosan az értékelt tárggyal kapcsolatos megfogalmazása. Ahogyan a világot és benne másokat látunk, lehetővé teszi számunkra, hogy jelentéssel töltsük fel társadalmi valóságunkat, melynek része mások értékelése. Ebből adódóan a fentiekben leírt modellt az értékelő értékrendszerével kell kapcsolatba hoznunk. Ehhez a Schwartz [10] által felállított egyetemes értékstruktúrára támaszkodtunk, melyben a szerző az emberi cselekvés mögött meghúzódó célokat és motivációkat csoportosította [10], [11]. Modellünk kategóriáiban az értékelő kifejezések tartalmi csoportosításához Schwartz értékstruktúrájának négy értékdimenzióját használjuk fel.

- **Nyitottság a változásra**, amely az önirányítás, az ingerlés és a hedonizmus értékeire utal
- **Konzervativizmus**, amely a tradíció, a konformitás és a biztonság értékeire utal.
- **Szelf-transzcendencia**, amelyhez a jóindulat és az univerzalizmus értékei tartoznak.
- **Ön-érvényesítés**, amely a hatalom és a teljesítmény értékeit érinti.

Az értékelő állapotok az adott értékre vonatkozóan pozitívak illetve negatívak lehetnek.

1. Táblázat: Schwartz értékdimenziói és az értékelő állapotok

Schwartz-dimenziók	**Értékelő állapot**	
	+	-
Nyitottság a változásra	A független. A kíváncsi. A aranyos.	A függő. A passzív. A visszataszító.
Konzervativizmus	A fegyelmezett. A udvarias. A tiszteletteljes.	A fegyelmezetlen. A udvariatlan. A tiszteletlen.
Szelf-transzcendencia	A jóindulatú. A őszinte. A becsületes.	A rosszindulatú. A hűtlen. A becstelen.
Önérvényesítés	A tekintélyes. A ambiciózus. A sikeres.	A engedékeny. A motiválatlan. A sikertelen.

2 Az értékelő állapotok nyelvtani vizsgálata

Ebben a fejezetben az automatizált szövegelemzés kidolgozásának eddig megtett lépései kerülnek bemutatásra, amely értékelési modellünk legabsztraktabb és a szövegen belül legkönnyebben beazonosítható szintjét, az értékelő mellékneveket tartalmazza. Egy szövegen belül értékelést kifejezhető érzelmi állapotok és mentális cselekvések tulajdonításának azonosításához Fülöp Éva által elkészített érzelmi szótárt

és Vincze Orsolya által létrehozott mentális igék kategóriáját fogjuk felhasználni. Az interpretatív igéket elkülönítettük a deskriptív igéktől, további feldolgozásuk még folyamatban van. Végső célunk, olyan mondatfeletti lokális nyelvtanok megírása melyek képesek a szövegben megjelenő különböző szintű, tartalmú és konnotációjú személy- és csoportközi értékelések azonosítására.

2.1 Az értékelő állapotok szótárai

Az MTA Nyelvtudományi Intézetének melléknév-szótárából, amely a magyar nyelvben leggyakrabban használt tizenötezer melléknevet tartalmazza, két független bíráló döntése alapján kiválogattuk a személyre (is) vonatkoztatható mellékneveket illetve melléknévi igenévképzős alakokat. Ezeket két kategóriába, a leíró (pl. szőke) és az értékelő melléknevek (pl. szép) kategóriájába csoportosítottuk. További kategóriát képeztek azok a melléknevek, amelyek szövegkörnyezettől függően lehetnek leírók és értékelők is (pl. vízfejű). Összesen 2133 leíró, 2010 értékelő és 1177 kontextusfüggő szót találtunk.

2. Táblázat: Részlet az általánosan pozitív illetve negatív konnotációval bíró melléknevek szótárából

Önmagukban pozitív konnotációval bíró melléknevek	**Önmagukban negatív konnotációval bíró melléknevek**
türelmes	Aberrált
udvarias	Abnormális
utolérhetetlen	Ádáz
üdítő	Aggályos
üdvös	Aggasztó
üdvözítő	Agresszív
ügyes	Agyalágyult
ügyi	Agyatlan
vájtfülű	Agymosott
választékos	Alábbvaló
változatos	Alacsony
varázslatos	Alacsonyrendű
vitális	Alamuszi
vonzó	Alantas
zseniális	Alattomos
mélyérzésű	Alávaló
tündöklő	Aljas

Az értékelő mellékneveket további két kategóriába csoportosítottuk konnotációjuktól függően. 475 olyan szót találtunk, amelyek a szövegtől dekontextualizáltan, alapvetően pozitív konnotációval bírnak, 689 negatívat és 846 olyan melléknevet (2. táblázat), amelyek értékét a szűkebb vagy tágabb szövegkörnyezet, határozza meg (ilyenek pl. az –ú/-ű képzős melléknevek – gyenge/jó képességű, illetve azok a melléknevek, melyek konnotációja a beszélőnek a szövegben kifejeződő attitűdjétől függ – pl. agilis, ironikus).

Az általános pozitív-negatív konnotáción túl az értékelő mellékneveket a fentebb bemutatott Schwartz-féle értékmodell négy értékdimenziója mentén is megítéltük. A

melléknevek minden egyes dimenzión pozitív, negatív vagy semleges értéket vehettek fel. Ily módon minden melléknév négy értéket kapott a dimenzionális kódolás során.

A szelf-transzcendencia dimenzión való kódolást például a következő, ellentétes jelentésű tulajdonságpárok alapján végeztük:

3. Táblázat: Részlet a szelf-transzcendencia értékdimenziót reprezentáló melléknévi szótárból

pozitív szelf-tr.:	**Negatív szelf-tr.**
Jóindulatú	Rosszindulatú
Univerzalista	Soviniszta
Önzetlen	Önző
Védelmező	Támadó
megértő, együttérző	Elutasító
Barátságos	közönyös, barátságtalan
Őszinte	Hazug
Becsületes	Becstelen
Szerető	Szeretettelen

Az értékelő melléknevek grammatikája

Az MTA Nyelvtudományi Intézetének segítségével a NOOJ nevű szövegelemző programban a fent leírt szótárakat lefuttattuk. Az értékelő melléknevek morfológiai és szintaktikai sajátosságainak figyelembevételével válik lehetővé olyan, a számítógépes tartalomelemzést vezérlő grammatikai szabályok („lokális nyelvtanok") megalkotása, amelyek alapján a NOOJ nevű program kimutatja az értékelő melléknevek konkordanciáit a vizsgált szövegben.

Elsőként egy olyan vezérlő szabályt kívánunk létrehozni, amely az egyes *tagmondatokon belül* elkülöníti az értékelő jelzős szerkezeteket, valamint az alanyként illetve névszói állítmányként szereplő értékelő mellékneveket.

Példák az azonosítandó elemekre:

- *Jelzős szerkezet:* A bölcs ember kevés szóból is ért.
- *Alany*: A bölcs kevés szóból is ért.
- *Névszói állítmány:* A dalai láma bölcs.

Ahhoz, hogy a Nooj el tudja végezni ezt a műveletet, első lépésben az azonosítandó elemek köréből ki kell zárnunk a következőket:

- az értékelő melléknevek ragozott alakjait (pl. szépen, jól, biztatólag);
- azokat a ragozott alakokat, amelyekben a melléknév mint kettős szófajú szó főnévként funkcionál (pl. szépre, jót, biztatóval);
- azokat a múlt idejű igealakokat, amelyek formájukban megegyeznek az igékből képzett befejezett melléknévi igenevekkel (pl. tanult, szeretett);
- azokat az értékelő mellékneveket, amelyek nem személyekre vonatkoznak (pl. szép táj, ez igaz);

Ily módon a program csak azokat az értékelő mellékneveket azonosítja, amelyek alanyesetben és egyes vagy többes számban vannak. Az alanyesetben lévő alakok egyes számban szerepelhetnek jelzős szerkezetekben, valamint egyes vagy többes számban alanyként illetve névszói állítmányként.

Második lépésben be kell vennünk az azonosítandó elemek körébe a következőket:

- az értékelő melléknevek közép, felső, és túlzófokú – alanyesetben és egyes vagy többes számban lévő – alakjait (pl. jobb, legjobb, legeslegjobb);
- azokat az értékelő mellékneveket, melléknévi igeneveket, amelyek szószerkezetekben nyerik el értéküket (kontextusfüggők); ilyenek többek között a következők:
  - az –ú/-ű képzős melléknevek (pl. jó képességű);
  - egyes melléknévi igenevek (pl. rosszul alkalmazkodó, jól felkészült)

Az értékelő melléknevek lokális nyelvtanai

A szótárak alkalmazásakor azzal a nehézséggel kerültünk szembe, hogy a program a keresés során a szótári elemeknek nem az általunk megadott alakját, hanem a tövét vette figyelembe, és az egyes tövek minden toldalékolt alakját azonosította (pl. a „kedves" esetében listázta a „kedvel", „kedvvel", „kedvenc", „kedvét alakokat is).

Ezt a problémát úgy küszöböltük ki, hogy a következő formában annotáltuk a keresett mellékneveket: <[abszolút vagy relatív] tő + melléknévképző + alanyeset>. Ily módon a program csak a keresett elemeket azonosítja, azoknak egyes ill. többes számú, alanyesetű alakjait (pl. a gondoskodó a <gondoskodik+imppart+nom> annotációt kapta). Az annotált mellékneveket képzőik szerint csoportosítottuk. Ez azért célszerű, mert egyes, a későbbiekben megírandó nyelvtanok csak egy-egy csoportra fognak vonatkozni (pl. a befejezett melléknévi igenevek olyan lokális nyelvtant igényelnek, amely elkülöníti azokat a múlt idejű igealakoktól). A képzőtípusok alapján a következő kategóriákat hoztuk létre: az -ú/-ű képzős (uattrib); -s képzős (sattrib); -i képzős (iattrib) melléknevek; a folyamatos (imppart), befejezett (perfpart), beálló történésű (futpart) és modális (modalpart) melléknévi igenevek kategóriáit; valamint az „egyéb" kategóriát, amelyben a nem képzett melléknevek (pl.: szép), illetve az olyan képzett alakok szerepelnek, amelyeket nem annotáltunk, mivel képzőjüket a Nooj nem azonosítja (pl. fosztóképző). Az annotált formában bevitt szótárak lehetővé tették az alanyesetű pozitív értékelő melléknevek kizárólagos azonosítását a szövegben. Jelzős szerkezetben, névszói állítmányként és alanyként minden esetben alanyesetű melléknévi alakok szerepelnek. A következő nyelvtan a pozitív értékelő melléknevek jelzős szerkezetben való azonosítására utasítja a Nooj-t (Fig. 2.).

Az „<A" és a „+ synsem=poz jelzős szerkezet>" parancsokat követve a program az azonosított nyelvi egységeket a „poz jelzős szerkezet" címkével látja el. A „pozjelszerk"-kel megjelölt, zárójelben található doboz beágyazott gráfokként tartalmazza a különböző képzővel rendelkező pozitív értékelő melléknevek egyes kategóriáit. A dobozba visszatérő rekurzív görbe és a dobozon belüli vessző együttesen a szövegben előforduló olyan felsorolásokat azonosítja, amelyekben a mellékneveket kizárólag vesszők választják el egymástól. A „<$pozjelszerk=A+sg-pssg-ps>" parancs mint megszorítás arra utasítja a programot, hogy a „pozjelszerk" egységben szereplő szótári elemek közül csak azokat az alakokat azonosítsa a szövegben, amelyek egyes számban és birtokos személyjel illetve birtokjel nélkül állnak (kizárva így pl. az „igaza van" kifejezést, azonosítva ugyanakkor az „igaz ember" szerkezetet). A következő egység, a dobozban szereplő „<N>" parancs szerint a program csak azokat a szerkezeteket azonosítja, amelyek egy vagy több (az előzőekben meghatározott kritériumoknak

eleget tevő) pozitív értékelő melléknévből, valamint egy közvetlenül utána álló (toldalékolt vagy toldalék nélküli) főnévből állnak. A személyre vonatkozó értékelő jelzős szerkezeteket egyelőre nem tudjuk elkülöníteni más entitásokra (pl. állatokra, tárgyakra) vonatkozóktól, mert a rendelkezésre álló főnévszótár tartalmi kategorizációja jelenleg még folyamatban van. A humán főnevek szótárának bevitelével az elkülönítés minden további nélkül lehetségessé válik.

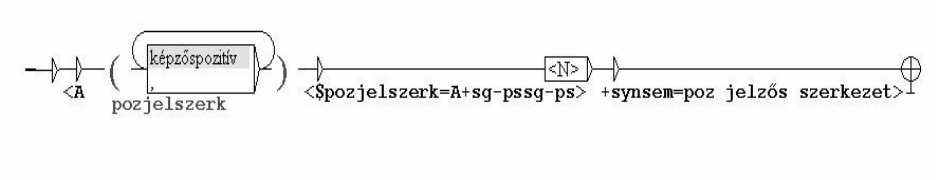


Fig. 2. A pozitív értékelő jelzős szerkezetek lokális nyelvtana

A lokális nyelvtan más melléknév-szótárak alkalmazásával is lefuttatható, ily módon azonosítani lehet a negatív értékelő mellékneveket illetve az egyes schwartzi dimenziókhoz tartozó mellékneveket is.

2.4 További feladatok

A fentebb bemutatott lokális nyelvtan egyelőre nem tudja kezelni az olyan jelzős szerkezeteket, amelyben a jelző vagy jelzők illetve főnévi referenciájuk nem közvetlenül egymás mellett helyezkednek el, hanem más elemek (pl. kötőszók) ékelődnek közéjük (pl. az okos és nagyon jó magaviseletű Péter). Ezen szerkezetek azonosításának problémája még megoldásra vár.

Azok az alanyesetű melléknevek, amelyek nem jelzős szerkezetek részei, vagy alanyként, vagy állítmányként szerepelhetnek a mondatban. A következő feladat, hogy ezeket a mellékneveket elkülönítsük egymástól, illetve azonosítsuk őket a szövegben. Ez azért szükséges, mert a különböző nyelvi formákban kifejezett értékelések az értékelés eltérő szubjektív hangsúlyosságát közvetíthetik.

A későbbiekben a következő alakok azonosítására fogunk lokális nyelvtanokat írni: a melléknévhez járuló ragokkal ellátott alakokra (pl. –n/-on/-an/-en, -l/-ul/-ül, -lag/-leg), amelyek nem közvetlenül az értékeltre vonatkoznak, hanem valamely hozzá tartozó entitásra (tárgyra, viselkedésre vagy tulajdonságra), ezzel indirekt módon kommunikálva az értékelést. Továbbá a főnévi ragokkal ellátott alakokra (pl. tárgyeset, részeseset), amelyek helyettesítik az értékeltet, és ily módon közvetlen értékelést fejeznek ki.

Bibliográfia

[1] Bruner, J.: Actual Mind, Possible worlds. Cambridge, MA: Harvard University Press. (1988)
[2] Bruner, J.: Acts of meaning. Cambridge, MA: Harvard University Press. (1990)

[3] Ricouer, P.: Construing and constructing. Review of the book The aims of interpretation. In M. J. Valdes (Ed.), A Ricoeur reader: Reflections and imaginations (pp. 195-199). New York: Harvester/Wheatsheaf. (1991)
[4] Liu, J. & Hilton, D. How the past weighs ont he present: Social representations of history and their role in identity politics. British Journal of Social Psychology (2005)
[5] Liu, J. & all.: Social representations of events and people in world history across twelve cultures. Journal of Cross-cultural Psychology. (2004)
[6] László, J. A történetek tudománya. ÚMK, Budapest (2005)
[7] Semin, G. R., & Fielder, K.: The cognitive functions of linguistic categories in describing persons: Social cognition and language. Journal of Personality and Social Psychology, 54, 558–568. (1988).
[8] Semin, G. R., & Fielder, K.: The Linguistic Category Model, Its Biases, Applications and Range. European Review of Social Psychology, 2, 1-30
[9] Smith, E. R. & Semin, G. R.: Socially situated cognition: Cognition in its social context. Advances in Experimental Social Psychology , 36 , 53-117 (2004)
[10] Schwartz, S. H.. Universals in the content and structure of values: Theory and empirical tests in 20 countries. In M. Zanna (Ed.), Advances in experimental social psychology (Vol. 25) (pp. 1-65). New York: Academic Press. (1992)
[11] Myyry, L. & Helkama, K.: University students' value priorities and emotional empathy. Educational Psychology, 21, 25-40. (2001)

NooJ fejlesztések a szubjektív időélmény tartalomelemzéses vizsgálatára[76]

Ehmann Bea[1]
Garami Vera[2]
Szabó Júlia[3]

[1] MTA Pszichológiai Kutatóintézet, 1132 Budapest, Victor Hugó u. 18-22,
ehmannb@mtapi.hu
[2] PTE BTK Pszichológiai Doktori Iskola, 7624 Pécs, Ifjúság útja 6.
garamivera@yahoo.com
[3] ELTE PPK Pszichológiai Intézet, 1064 Budapest, Izabella u. 46.
szabo.julia@gmail.com

1 Előzmények és háttér

Korábbi közleményeinkben már beszámoltunk a szubjektív időélmény élettörténeti szövegekben történő tartalomelemzéses vizsgálatairól (Ehmann [2]; Ehmann, et al. [3,4,5]; László [6];). Jelen előadásunkban a NooJ tartalomelemző szoftvernek az MTA Nyelvtudományi Intézete Korpusznyelvészeti Osztálya által számunkra elérhetővé tett magyar változata segítségével végzett fejlesztéseket mutatjuk be (Váradi [8]).

2 Funkcionális és tartalmi időkategóriák

Korábbi elgondolásunkhoz képest újdonság, hogy a szubjektív időélmény pszichológiai vizsgálatában megkülönböztetünk tartalmi (az elbeszélt szöveg témájától függő), illetve funkcionális időkategóriákat.

A tartalmi időkategóriák közé soroljuk az idő léptékére vonatkozó szavakat a pillanatnyiságtól az örökkévalóságig (pillanat/perc/óra, nap/hét/hónap/év/évtized, évszázad, stb.), valamint a személyes életperiódusok (pl. gyermekkor/iskolaévek/katonakor, stb.), a személyes, családi, vallási és nemzeti ünnepek és események (pl. születésnap, karácsony, a háború, az 56-os forradalom, stb.) nyelvi markereit. Ezek szövegbeli gyakorisága önmagában nem utal pszichológiai korrelátumokra, hiszen előfordulásuk az elbeszélt témakör függvénye.

A pszichológiai korrelátumokat a funkcionális időkategóriák közvetítik. A továbbiakban bemutatott fejlesztésekkel ilyen funkcionális időkategóriák vizsgálatát

[76] A kutatást az NKFP 2001-2004/5/26, és az OTKA 2004-2007/T-046522 számú pályázatai támogatták.

kíséreltük meg. Elsőként az idő szubjektív tempójára (gyors – lassú) és tartamára (pillanatnyi – huzamos) utaló nyelvi markereket próbáltuk megragadni.

3 NooJ gráfok a funkcionális időkategóriák tartalomelemzéses vizsgálatára

NooJ fejlesztéseink lényege, hogy a funkcionális időkategóriák szólistáit egymásba ágyazott gráfokká építjük össze, és e gráfokat használjuk a tényleges szövegbeli találatok konkordancia listájának elkészítésére. Az egyes szókategóriákat a Korpusznyelvészeti Osztály munkatársaitól kapott, a tíz-tízezer leggyakoribb magyar igét, határozószót és melléknevet tartalmazó szólistákból állítottuk össze.

Jelenleg elkészült illetve tervezett funkcionális időszótáraink:

(a) Gyorsaság – Lassúság;
(b) Kezdet – Befejezés;
(c) Pillanatnyiság – Huzamosság;
(d) Ismétlődés

Ezeket a szótárakat építjük össze a tartalmi időszótárakkal:

(a) A hét napjai;
(b) Az év hónapjai
(c) Ünnepek;
(d) Életperiódusok, stb.

A gráfok alapelve, hogy a funkcionális kategóriák felülírják a tartalmiakat. Például:

'egész' + éjjel/nyáron/augusztusban' = huzamosság
'minden' + éjjel/nyáron/augusztusban' = ismétlődés

3.1 Az idő szubjektív tempója

Az időmodul egyik kategóriája az idő szubjektív tempójával foglalkozik. Úgy gondoljuk, hogy a szubjektív időélmény egyik alapvető, pszichológiailag is jelentős tényezője a gyors – lassú dimenzió. Fontos lehet az időélménynek ez a dimenziója például a klinikai pszichológiában, ahol különböző mentális betegségeknél beszélnek az idő szubjektív tempójának patológiás módosulásairól. Bschor és munkatársainak [1] kutatási eredményei megerősítették a szakirodalomban konzisztensen a depressziós személyeknek tulajdonított 'meglassult', illetve a mániás személyekre jellemző 'felgyorsult' időélményt. A kontroll személyek élménye az idő áramlásának szubjektív tempójáról ebben a vizsgálatban kiegyensúlyozott volt. Bár itt a szerzők vizuális analóg skálát alkalmaztak, úgy gondoljuk, hogy az időélmény szubjektív

tempójának különbségei megjelennek a spontán nyelvhasználatban is, így ez a jelenség tartalomelemzéssel is megragadható.

A 'Gyorsaság – Lassúság' funkcionális szemantikai kategória alapja az igeszótár és a határozószó lista. A 'Gyors' illetve a 'Lassú' igék szótárait a tízezer leggyakoribb igéből válogattuk ki. A következő szempontok szerint kerültek be szavak ezekbe a szótárakba:

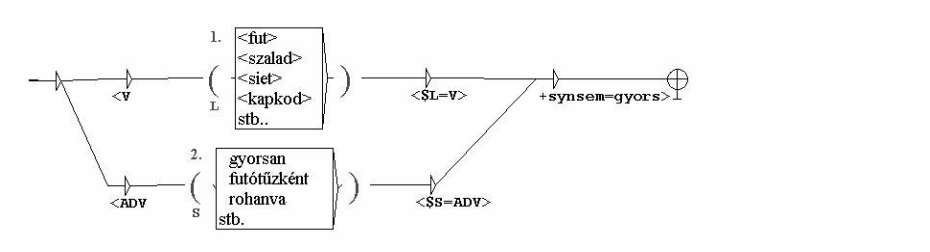
1.'Gyors' igék (606 db):
(a) Az ige gyors mozgást fejez ki. Például: száguld, fut, rohan
(b) Az ige sietséget fejez ki. Pl. elhamarkodik, elkapkod, siet, összecsap, lezavar

2. 'Lassú' igék (174 db):
(a) Az ige lassú mozgást fejez ki. Pl. ballag, sétál, cammog
(b) Az ige valamilyen módon akadályozott mozgást fejez ki. Pl. vergődik, biceg, sántikál

A 'Gyors' és a 'Lassú' határozószókat is elkülönítettük a gyakoriság alapján készült lista szavai közül. Ezek jelentése a tempóra vonatkozó melléknevekből képzett szavak a 'Lassú' kategóriában: csigalassan, lassan, lassanként stb. Ez hasonló a 'Gyors' határozószók esetében is - pl. élénken, elevenen, sebesen, gyorsan - , de vannak olyan szavak is, amelyek a 'Gyors'/(b) igékhez hasonlóan sietséget vagy gyors mozgást fejeznek ki, és ezekből származó határozói igenevek: Pl. kapkodóan, rohanva. A 'Gyors' határozók egy harmadik csoportja – hamar, hirtelen, tüstént a magyar nyelvben ismert idő-határozószók közül kerülnek ki. Szóba jöhetnek még főnevek határozóraggal álló alakjai is, pl. futótűzként. Érdekes, hogy mennyivel kevesebb a lassú ige és határozószó az eredeti gyakoriság alapján összeállított szótárakban.

3.1.1 Nooj működés – 'Gyors' és 'Lassú' időkategória

A Nooj program lehetővé teszi, hogy olyan gráfokat alakítsunk ki, amelyek megtalálják és 'felcímkézik' számunkra a megfelelő időkategóriát, és a találatok konkordanciáit is megadják.



1. ábra. A 'Gyors' időkategória Nooj gráfjának vázlata

A fenti gráf 1. számmal jelölt felső része azokat az igéket ismeri fel, amelyeket a 'Gyors' igeszótárban megadtunk. Az <$=SV> kikötés biztosítja, hogy ha a szövegen

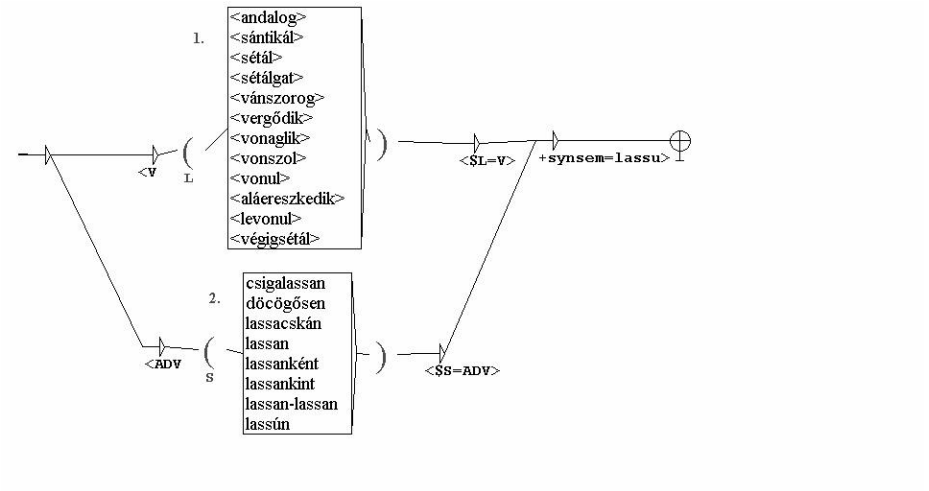
előzőleg lefuttatjuk a lexikai elemzést, akkor a gráf csak akkor címkézi majd kimenetként gyorsnak az adott szót, ha a szó annotációjában az „ige" (V) szerepel. Ezek az igék kimenetként „V+gyors" címkét kapnak. A gráf 2. számmal jelölt alsó része ugyanezt teszi a felsorolt határozószók esetében, és azokat „ADV+gyors" címkével látja el. A Nooj így többet tud, mintha egyszerűen csak megjelölné az általunk megadott szótárban feltüntetett kifejezéseket, hiszen a megadott szófaji kikötésekkel sok hibát kiküszöbölhetünk. Az itt közölt gráfok az eredetieknek csak vázlatai, mivel a könnyebb áttekinthetőség kedvéért nem tartalmazzák a teljes szótárakat.

A gráf segítségével próbaképpen egy regény (Vámos Miklós [7]: Anya csak egy van) szövegében próbáltuk megtalálni a 'Gyors' időélményre utaló mondatokat. A Nooj program által kiadott konkordancia output egy részét mutatja be a 2. ábra. Látható, hogy vannak olyan esetek, amikor a gráf nem azt találja meg, amit kerestünk, pl. a „tűz" esetében, amely bár az itt talált mondatokban főnév, mivel ige is lehet (eltűzni valahonnan), a lexikai elemzés során ige és főnév annotációt is kapott. Ezeket az eseteket utólagos ellenőrzéssel szeretnénk kiküszöbölni, és a gyakran előforduló téves riasztásokra a gráfok módosításával megoldást találni.

```
edjél, ha tudsz, de gyorsan/<ADV<gyorsan=ADV>+synsem=gyors>!különben
t sokat, viszonylag hamar/<ADV<hamar=ADV>+synsem=gyors> beadta a
összegubancolódott, rángatja/<V<rángatja=V>+synsem=gyors>,ahelyett
buci bácsi egyszer  befutott/<V<befutott=V>+synsem=gyors> délután,
n. E nyúlós hangtól végigszaladt/<V<végigszaladt=V>+synsem=gyors>
gyre erőteljesebben lüktetett/<V<lüktetett=V>+synsem=gyors> az
ll a haja, biztosan futott/<V<futott=V>+synsem=gyors>, jól van,
ó szája kínvigyorba szaladt/<V<szaladt=V>+synsem=gyors>: pont most?!
ájt, hogy a lábából kiszaladt/<V<kiszaladt=V>+synsem=gyors> az erő,
rt nem találtunk rá hamarabb/<ADV<hamarabb=ADV>+synsem=gyors>? miért
égitámadások idején kergetett/<V<kergetett=V>+synsem=gyors>körbe az
-papot hátra hagyva rohanna/<V<rohanna=V>+synsem=gyors> a kórházba, s
ndig Ladó kért neki tüzet/<V<tüzet=V>+synsem=gyors>a pincértől, s
orné, jól teszi, ha siet/<V<siet=V>+synsem=gyors>, az én férjem nem
ómot kapok, hogy ne ugráljak/<V<ugráljak=V>+synsem=gyors>, vegyem
dig vérvörös füllel elrohant/<V<elrohant=V>+synsem=gyors>, még a
ották a hevenyészve összecsapott/<V<összecsapott=V>+synsem=gyors>
k az hiányzik, hogy lezuhanjak/<V<akinek a r=V>+synsem=gyors>. Ladó
az anyját, azonnal  rohant/<V<feketé=V>+synsem=gyors>hozzá,
onnal orvoshoz kéne rohanni/<V<rohanni=V>+synsem=gyors>, vagy egyenest
határozott akciókat sürgetett/<V<sürgetett=V>+synsem=gyors>. Ladó
isten, hogy anya ne fusson/<V<fusson=V>+synsem=gyors>be este vagy
ak érezte magát, és sürgősen/<ADV<sürgősen=ADV>+synsem=gyors> a tettek
könyvet nem talált. Kapkodva/<V<Kapkodva=V>+synsem=gyors>
z anyja ölelését, s kirohant/<V<kirohant=V>+synsem=gyors>
```

2. ábra. A 'Gyors' gráf konkordanciái az „Anya csak egy van" című regény szövegében (kivonat)

A 'Lassú időkategória gráfja ugyanezzel a módszerrel működik. Az alábbiakban látható a gráf vázlata és a regény szövegében talált konkordancia részlete.



3. ábra. A 'Lassú' időkategória Nooj gráfjának vázlata

```
s falnak támasztva, lassan/<ADV<lassan=ADV>+synsem=lassu> lecsúszom a
Gellért-hegyen      sétáltak/<V<sétáltak=V>+synsem=lassu> Verával, s
m... - a kád közben lassacskán/<ADV<lassacskán=ADV>+synsem=lassu>
és ő vert seregként vonszolta/<V<vonszolta=V>+synsem=lassu> ma gát a
nkább a belvárosban sétálgatok/<V<sétálgatok=V>+synsem=lassu>,
vagy moziba, vagy   sétálni/<V<sétálni=V>+synsem=lassu>. Különben is
```

4. ábra. A 'Lassú' gráf konkordanciái az „Anya csak egy van" című regény szövegében (Kivonat)

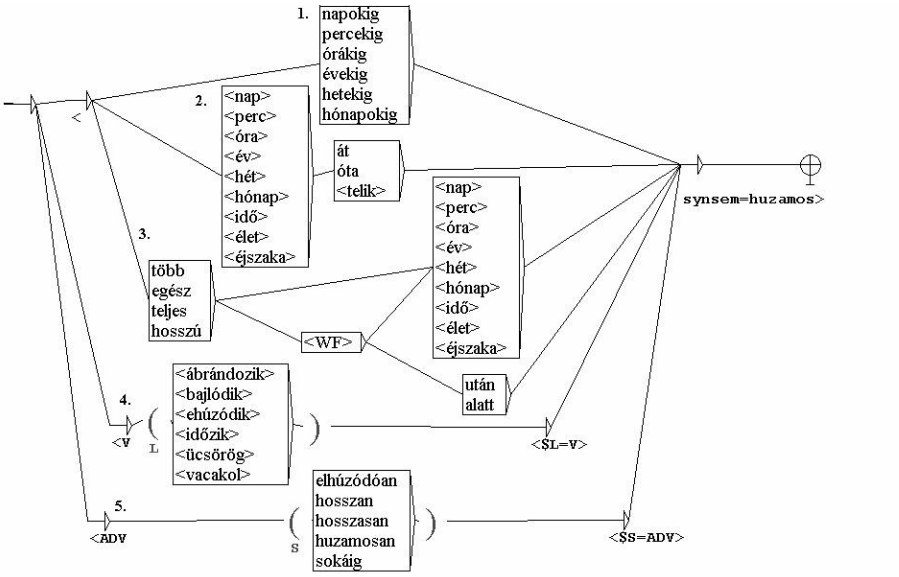
3.2 Időtartam

A szubjektív tempó vizsgálatára kidolgozott szótárak készítése közben felmerült a kérdés, hogy azok az igék, amelyek olyan cselekvést vagy történést fejeznek ki, amelyek nagyon rövid idő alatt történnek (begurul, belelapoz), illetve egy adott pillanatban (csattan, köhint, pillant) a 'Gyors' kategóriába kerüljenek-e. Hasonlóan ehhez a lassú időélménnyel is először kapcsolatba hoztuk a hosszú ideig tartó, huzamosságra, elhúzódásra utaló szavakat, és azokat, amelyek valamilyen „időigényes" cselekvésre, történésre utalnak (pl. ábrándozik, ácsorog, telik-múlik). Végül külön szótárt hoztunk létre a 'Pillanatnyi' illetve a 'Huzamos' időkategóriáknak, mivel úgy gondoljuk, hogy ezek a szavak kifejeznek időbeliséget, de ez nem az idő szubjektív tempójával, hanem egy adott cselekvés, történés várható tartamának hosszával van kapcsolatban.

3.2.1 Nooj működés – 'Huzamos' időkategória

A 'Huzamos' időkategória Nooj gráfja összetettebb, mint a fentebb bemutatott 'Gyors' és 'Lassú' gráfok, még inkább kihasználja a program adta lehetőségeket. Az 5. ábrán látható a gráfnak egy vázlata, amelyben a tartalmi időszótárakat (nap, óra, perc stb.) összeépítettük olyan kifejezésekkel, amelyekkel együtt huzamosságot fejeznek ki.

Elsőként létrehoztunk egy időtartamokat kifejező főnevekből álló listát (perc, óra, nap, éjszaka, hónap, év, évszak, idő stb.), és ezek konkordanciáit vizsgálva kerestük a huzamosságra utaló több szóból álló kifejezéseket. Ezzel párhuzamosan kialakítottunk egy 'Huzamos' igeszótárt és határozószó-listát.



5. ábra. A 'Huzamos' időkategória Nooj gráfjának vázlata

A gráf 1. számmal jelölt ága a legegyszerűbb, az időtartamokat jelölő főnevek olyan ragozott alakjait tartalmazza (napokig, percekig, stb.), amelyek önmagukban is elegendőek a huzamosság kifejezéséhez.

A gráf 2. ága egy tartalmi időszótár, amelynek elemeit akkor címkézi 'huzamosnak' a gráf, ha utána az „át" illetve az „óta" névutó, vagy a „telik" ige valamelyik alakja követi. Pl. „éveken át", „évek óta", vagy „évekbe telt".

A gráf 3. számmal jelölt középső része azokat a mellékneveket illetve számneveket veszi figyelembe, amelyek a megadott feltételekkel szintén huzamosságot fejeznek ki. A melléknév vagy számnév akkor esik bele az általunk keresett kategóriába, ha utána valamilyen időtartamot kifejező főnév áll, pl. „hosszú órák". A „<WF>" kifejezés azt az elemet képviseli a gráfban, amely beékelődhet a melléknév, illetve számnév és az időtartamot kifejező főnév közé. Pl. „egész álló nap". A 3. ágnak van még egy útvonala, amelyben az első helyen álló névszót követi egy másik szóalak, és ezt pedig az „után" vagy az „alatt" névutó. Erre példa a „hosszú csönd után", de itt találja meg a gráf az „évek hosszú sora alatt" kifejezést is.

A gráf 4. és 5. ága a fentebb bemutatott 'Gyors' és 'Lassú' gráfhoz hasonlóan megtalálja a 'Huzamos' igéket, illetve határozószókat. Mind az öt ág találatait egy közös konkordancia listában kapjuk meg, de az igéket „V+huzamos", a határozószókat pedig „ADV+huzamos" címkével ellátva.

kulcsa van, sárika **sokáig**/<ADV<sokáig=ADV>synsem=huzamos> vacakolt a
képességét, mindig **hosszan**/<ADV<hosszan=ADV>synsem=huzamos> szarakodik
a
valamit? - persze, **egész életében**/<synsem=huzamos> ilyen tehetetlen
volt
töltsék együtt az **egész napot**/<synsem=huzamos>. Szerelmeskedéssel. Ha
hát ez őrület, **órákig**/<synsem=huzamos> fogok itt ülni, de ő nem i
cipót, olyat,ami **hetekig**/<synsem=huzamos> se szárad meg, szegény
azt jelenti...?! - **napokig**/<synsem=huzamos> idézgette magában e
n, huszonkét percig **ücsörgött**/<V<ücsörgött=V>synsem=huzamos> azon a
nagy baj van, majd **hosszú csönd után**/<synsem=huzamos> hozzáteszi,
bejárati ajtóban - **hosszú hatásszünet után**/<synsem=huzamos> -
beszélek, miközben **napok telnek**/<synsem=huzamos> **el** úgy, hogy nincs is
nagyjából *harminc* **éven át**/<synsem=huzamos>, kimondani is szörnyű,
én vagyok a hibás. **Napok óta**/<synsem=huzamos> ígérgettem neki, hogy
félreértette, aztán **évekbe telt**/<synsem=huzamos>, mire az agya bevette
a
t volt, hanem *három* **teljes napig**/<synsem=huzamos> ki se bújtunk az
plakátot, állítólag **több éve**/<synsem=huzamos> kinn lóg az irodán és az
adó befejezte, Alfy **hosszú ideig**/<synsem=huzamos> töprengett, majd azt
merte. Ladó az *évek* **hosszú sora alatt**/<synsem=huzamos> összeállította
zoknom a protézist, **hosszú hetekig**/<synsem=huzamos> vissza-visszajártam
viselt. Ladó ezután **hosszú évekig**/<synsem=huzamos> nem hallott felőle.
es foglal kozását - **egész elszivárgó életét**/<synsem=huzamos>. A
sel töltötte, *három* **éjszakán át**/<synsem=huzamos> összesen nem aludt
en áldott nap *nyolc* **hosszú órát**/<synsem=huzamos> dolgozik a laborban,
én
végtelennek tetsző **idő óta**/<synsem=huzamos> nem tárcsázták. Itt a

6. ábra. A 'Huzamos' gráf konkordanciái az „Anya csak egy van" című regény szövegében (Kivonat)

3.3 Nehézségek és megoldási lehetőségek

A Nooj gráfokkal való munka közben felmerült néhány nehézség, például a fentebb már említett többértelműség problémája. A 'Gyors' kategóriánál például a „tűz" ige, amelyet felvettünk a 'Gyors' igeszótárba, mivel a „tűzzünk el innen" típusú kifejezéseket szerettünk volna megtalálni. Azonban a „tűz" egyben főnév is, ugyanakkor igeként is többértelmű, („tűzz a hajadba egy szalagot"), így a probléma megoldhatatlannak tűnik a gráfok szintjén. Ilyenkor a konkordancia utólagos javítása lehet a megoldás.

Egy másik probléma, hogy egy adott mondatban két egymást követő szó a gráf két különböző útvonalán is bekerülhet a konkordanciába. Például: „a Sokáig vacakolt a zárral" kifejezés kétszer is bekerült, mivel a „vacakol" „V+huzamos", a „sokáig" pedig „ADV+huzamos" címkét kapott.

Ezzel kapcsolatban dilemmánk merült fel. Egyfelől kézenfekvőnek tűnik, hogy ezeket a kettős előfordulásokat a statisztikai elemzésben csak egyszeri találatként szabad figyelembe venni. Másfelől az ilyen kettős hangsúlyozás pszichológiai szempontból különleges jelentőséggel bírhat, és ezért külön kategóriaként is kezelhető.

Tanulmányunk műhelybeszámoló; az elkészült lokális nyelvtanok egyrészt még fejlesztés alatt állnak (további időgráfok is készülnek), másrészt ezek szolgálnak kiindulópontként az elbeszélt önéletrajzi emlékezet idői szerkezetének feltárásához.

Köszönetet mondunk az MTA Korpusznyelvészeti Osztálya munkatársainak – Váradi Tamásnak, Nagy Viktornak és Vajda Péternek – a NooJ pszichológiai alkalmazásához nyújtott szakértői segítségért.

Bibliográfia

1. Bschor, T., Ising, M., Bauer, M., Lewitzka, U., Skerstupeit, M., Müller-Oerlinghausen, B., and Baethge, C.: Time Experience and Time Judgement in Major Depression, Mania and Healthy Subjects. A Controlled Study of 93 Subjects. Acta Psychiatrica Scandinavica. 109: 222-229. (2004)
2. 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)
3. Ehmann Bea, Kiss Balázs, Naszódi Mátyás, László János: A szubjektív időélmény tartalomelemzéses vizsgálata. A LAS Vertikum időmodulja. Pszichológia, 2005/2. 133-142. (2005)
4. Ehmann Bea: NooJ in Psychological Content Analysis: Time Structure of Short Traumatic Recalls. In: Vitas, D. and Silberztein, M. /Eds./: 9th Intex/NooJ Conference, Belgrade, Serbia, June 1-3, Abstracts. pp.73-74. (2006)
5. Bea Ehmann, Vera Garami, Matyas Naszodi, Balazs Kis, Janos Laszlo: Subjective Time Experience: Identifying Psychological Correlates by Narrative Psychological Content Analysis. Empirical Text and Cultural Research, In Press.
6. László János: A történetek tudománya. Bevezetés a narratív pszichológiába. Budapest, Új Mandátum Könyvkiadó. (2005)
7. Vámos Miklós: Anya csak egy van. AB OVO. (2006)
8. Váradi Tamás: Translation of Multiword Expressions in NooJ. In: Vitas, D. and Silberztein, M. /Eds./: 9th Intex/NooJ Conference, Belgrade, Serbia, June 1-3, Abstracts. pp.13-14. (2006)

Az intencionalitás modul kidolgozása NOOJ tartalomelemző programmal

Ferenczhalmy Réka[1], László János[2]

[1] Pécsi Tudományegyetem Pszichológia Doktori Iskola,
7624 Pécs, Ifjúság útja 6.
ferreka@freemail.hu
[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: A pszichológiai narratív elemzés és kutatás célja, hogy feltárja, milyen módon jelenik meg a szövegben a pszichológiai sík, mely túlmutat az események puszta leírásán, azaz milyen nyelvi kódok által fejezzük ki és tudjuk kikövetkeztetni ezeket az implikált tartalmakat. A narratívumok, illetve a világ általuk való leképezése egy konstruktív folyamat, melynek jellegzetességei pszichológiailag releváns információval szolgálnak a szöveg alkotójáról. Ennek egyik fontos dimenziója az intencionalitás, ami arra vonatkozik, hogy milyen szándékot tulajdonítunk saját és mások viselkedésének. A kutatásban a NOOJ tartalomelemző programmal dolgozunk, amely a lexikai egységek megragadásán kívül gráfok, azaz lokális nyelvtanok segítségével morfoszintaktikai elemzésre is képes. Távlati cél a különböző modulok kidolgozása után, ezek interakciója alapján egy, a személyiség komplex működésének megragadására is képes szövegelemző program kidolgozása.

1 Bevezetés

Az embert folyamatosan körülvevő, ámbár állandóan változó hatalmas mennyiségű információ felfogásában és feldolgozásában, mely alapján saját, egyedi világunkat megkonstruáljuk, a nyelvnek meghatározó szerepe van. A nyelv is, csakúgy, mint az általa reprezentált világ, állandóan változik, ilyen értelemben élő, dinamikus entitásként beszélhetünk róla. Egyfelől univerzális, másfelől kulturálisan determinált formaként és eszköztárként is jelen van a gondolkodásban, az egyén szempontjából pedig fontos, hogy ezt milyen mértékben és milyen módon tudja igénybe venni, mennyire tudja általa megragadni és megkonstruálni saját külső és belső élményvilágát. Pszichológiai szempontból rendkívül jelentős, hogy az élmények milyen szinten

ragadhatók meg verbálisan, mely tehát több - kulturális, társadalmi illetve egyéni - szinten is meghatározott.

Már kisgyermekkortól, a beszéd kezdetével a nyelv elsődleges szerepet tölt be a reprezentáció kialakulásában („mire van szavunk" – nyelvi relativitás elmélet), illetve az érzések, élmények megélésében és feldolgozásában. A nyelv egy olyan közeg, olyan csatorna, amely által az élmények megfoghatóvá és megoszthatóvá válnak, azaz lehetővé teszi az absztrakt gondolkodást illetve kommunikációt, mely az egyének közötti megértés alapja. Azt is mondhatjuk, hogy a személyek közötti, nonverbálisan bizonyos szempontból áthidalhatatlan szakadékon ível át. Ennek két jelentős aspektusát emelem ki: egyrészt az egyén élményfeldolgozása és identitásának alakulása szempontjából fontos, hogy mindez társas közegben zajlik[77], másrészt lehetővé teszi, hogy nagyobb létszámú csoportok és magasabb szervezettségi szinten éljenek együtt, mivel a szociális térben való eligazodást szolgálja. [1]

Nagyon fontos tehát, hogy miként reflektál az egyén az őt körülvevő szociális közegre, illetve magára, mint eme szociális térben működő individuumra, azaz hogyan érzékeli a saját maga és környezete között fennálló kölcsönhatást. Ennek meghatározó aspektusa az intencionalitás.

Az intenció szó jelentése az *Idegen szavak és kifejezések szótára* alapján a következő:

„intenció *lat* 1. szándék, törekvés, célzat 2. *fil* a tudatnak vmely tárgyra irányulása (élményben) 3. *orv* sebgyógyulás"[2]

Itt is megjelenik az a kettőség a fogalom jelentésében, mely a pszichológiában való használatára is jellemző: a szándékosság, illetve a valamire való vonatkozás/ vonatkoztatás, utalás.

Franz Brentano (1838-1917), a folyamatpszichológia atyja, a lelki jelenségek megközelítésében az intencionalitást, azaz a valamire való irányulást hangsúlyozza, melyet szembeállít a fizikai jelenségek önmagukért való létével. A lelki jelenségeknek ezek alapján két aspektusa különíthető el: amire irányul, a tárgya, és ahogyan irányul, azaz a művelet, aktus. Ez utóbbinak három típusát írja le: a képzetalkotást és ideációt, a megítélést illetve a szeretet-gyűlöletet. Elképzelésében fontos szerepet kap a reprezentáció – csak olyan dologhoz tudunk viszonyulni, ami számunkra már létezik, pl. nem tudunk *csak úgy* látni, mindig *valamit* látunk. [3]

Brentano két aspektusa – a lelki jelenségek tárgya és aktusa – mellett fontos, hogy mindig *valaki* lát valamit. Ez a továbbgondolás William Stern nevéhez fűződik, aki tehát ugyanennek a mozzanatnak a harmadik aspektusát hangsúlyozta. [3]

Az intencionalitás tágabb értelemben a szándék mellett a különböző belső állapotok tulajdonítását is magában foglalja. Kutatásunkban szűkebben, csak a szándéktulajdonítást vizsgáljuk. Bizonyos értelemben minden cselekvést szándékosnak tekinthetünk. Kutatásunkban azonban arra fókuszálunk, hogy a szándék az események puszta leírásán túlmenően, elsődlegesen milyen módon jelenik meg a szövegben, azaz hogy a személy hogyan jeleníti ezt meg, minek tulajdonítja és hogyan reflektál rá.

---

[77] Az identitás felfogható élettörténetünk szerveződéseként is – hogyan integrálja az egyén a különböző életeseményeit, tapasztalatait a róluk szóló narratívumok mentén koherens élettörténetté. Ilyen értelemben a pszichoterápia narratív újraírás, újrastrukturálás, mely az élmények megragadására és integrációjára irányul. [1]

2 A modul felépítése

2.1 A modul teljes szerkezete

Az intencionalitás kódolása a szövegben több szinten zajlik. Kutatásunk során az igékből indultunk ki, a modul alapját a Todorov által leírt igei transzformációk képezték, melyek arra vonatkoznak, hogy az igék milyen pszichológiai jelentéseket hordoznak, illetve ezek hogyan módosulhatnak. A hat transzformáció (modalitás, intenció, eredmény, mód, aspektus, státus) közül kettő releváns a modul szempontjából: az intenció és az eredmény. (Az eredmény megjelenése, pl. valami *sikerült* valakinek, implikálja a szándék előfeltevését, ezért tartozik ide.) [4]

Igei szinten intencionalitást a következő jegyek közvetítenek:

- Jövőidő (El fogok utazni. Meg fogom venni.)
- Feltételes és felszólító mód (Elmennék nyaralni. Ne álmodozz!)
- Műveltetés (Megjavíttatom az autómat.)
- Bizonyos mentális igék (tervez, remél, szándékozik, stb.)

Az igék esetében lényeges, hogy az alany személy legyen, melyet a közeljövőben már tudunk kezelni. (Ki fog sütni a Nap. A vízben lejönne a naptej a bőrömről. stb. nem intencionálisak.)

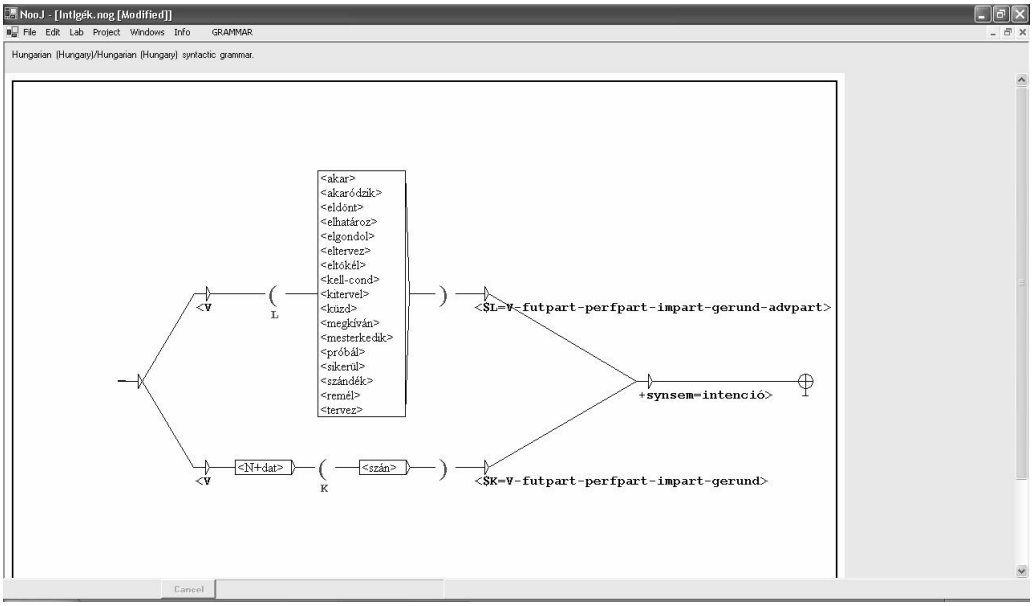
Egyéb nyelvtani elemek és szintek, melyek szintén a intenció nyelvi kódjainak tekinthetők.:

- Főnevek (szándék, cél, akarat, remény, stb.)
- Célhatározók
  - ragos (-ért) és névutós (végett, érdekében) főnevek (Elmegyek egy koktélért. Az elutazásunk végett kérdezem. A nyaralás érdekében tette.)
  - határozói névmás – avégett (Avégett adta ide az autóját, hogy járjak vele.)
  - főnévi igenév (Szórakozni jöttem. Ebédelni mentem.)
- Módhatározószók (szívesen, szándékosan, direkt, stb.)
- Célhatározói alárendelő mondatszerkezetek (Azért jöttem, hogy megmutassam a nyári fényképeket.)

A továbbiakban a határozószók, a feltételes mód és a mentális igék, illetve a főnevek kerülnek részletesebb bemutatásra.

2.2 Intencionális igék

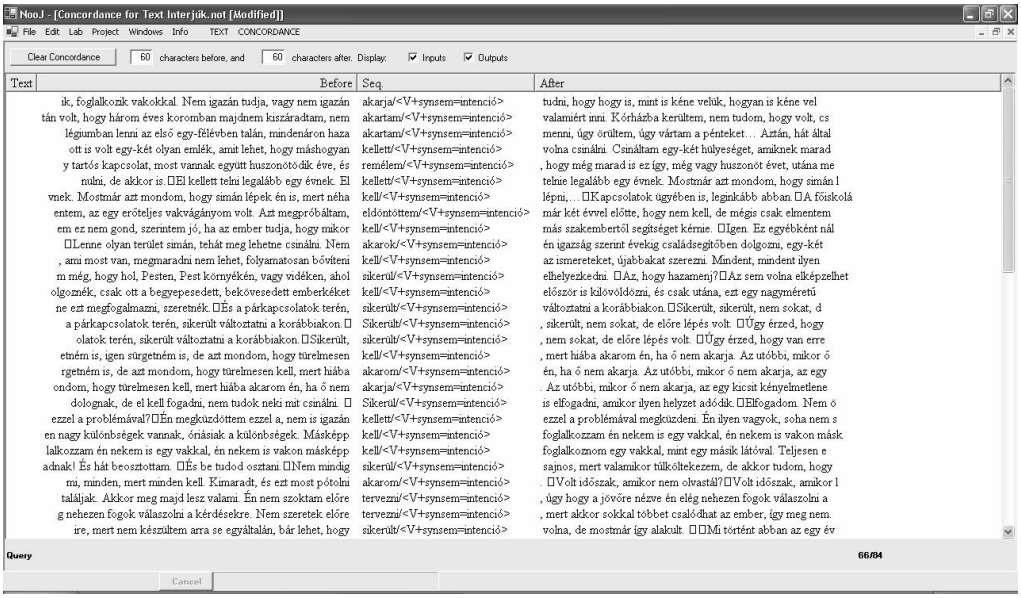
A korábbiakban már volt róla szó, hogy bizonyos értelemben minden cselekvő ige intencionális (ez egy lényeges kapcsolódási pont az aktív-passzív modulhoz), azonban ezen modul esetében arra irányul a keresés, amikor a szándékosságra (vagy a szándék hiányára) tett utalás megjelenik a szövegben, azaz amikor ez a szint elsődlegesen van jelen. Pl. *Ilma iszik egy pohár vizet.*, ezt önmagában még nem soroljuk ide, bár nyilvánvaló, hogy egy szándékos cselekvésről beszélünk. De ha azt mondjuk, hogy *Ilma egy pohár vizet akar inni.*, vagy *Ilmának sikerült innia egy pohár vizet.*, abban már benne van a szándék megjelölése is.



1. ábra Intencionális igék gráfja

A fenti gráfról leolvasható, hogy mely igéket soroltuk ide. Látható, hogy vannak kitételek: a felsorolt szavak (L-változó) minden alakját keresse a program (itt még bármilyen szófajú lehet), azonban a kimenetnél megszabtuk, hogy a találatok közül csak az igéket vegye figyelembe, azok közül is kizártuk a melléknévi vagy határozói igenévekként, illetve modális alakban szereplőket. Ez utóbbi a lehetőségre utal, melyről még a későbbiekben lesz szó. Az előző példamondatra visszatérve: attól még, hogy *Ilma dönthet úgy, hogy iszik egy pohár vizet.*, attól még nincs jelezve, hogy mi a feltételezett szándéka. (Pl. *Dönthet úgy, de szerintem nem akar.*)

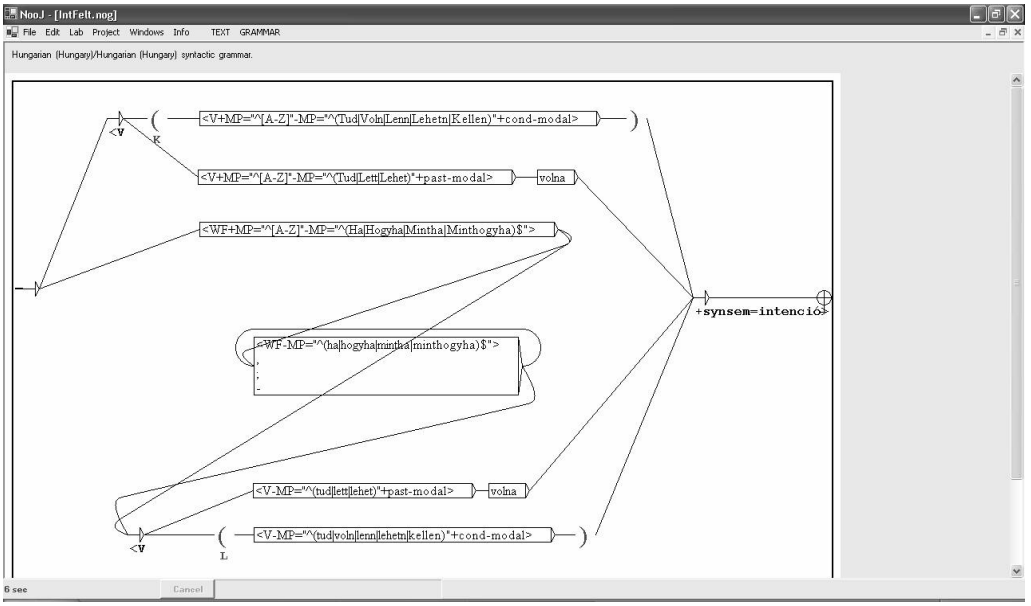
A gráfban szerepel a szán ige is, de csak abban az esetben, ha közvetlenül előtte áll, hogy minek szán valaki valamit. Pl. *Ilma viccnek szánta, hogy megissza a vizemet. (Én viszont megharagudtam érte.)* Itt szándékról van szó, akár a szándék és ersdmény eltérését is megjelenítve, mely pszichológiailag szintén nagyon fontos. Ha a bűnét szánja valaki, vagy a szán mint főnév szerepel, az már nem tartozik ide és ezt az esetet ezzel a gráffal ki is szűrtük. (Még szükséges ilyen kiegészítő gráfokat létrehozni, pl. a tervez ige sem intencionális, amennyiben konkrét cselekvésként, pl. egy ház tervrajzának elkészítését jelenti.)



2. ábra Intencionális igék konkordanciája

2.3 Feltételes mód

Azt mondhatjuk, hogy az ige feltételes módban, lévén, hogy nem konkrét cselekvést ír le, hanem utal arra, hogy a személynek milyen szándéka, vágya, stb. lenne, intencionális. Itt is beleütközünk azonban a lehetőség megjelenésébe, ami nem sorolható ide. Pl. Ilmánál maradva: *Ilma inna egy pohár vizet.*, vagy *Ilma szeretne (inni) egy pohár vizet.*, ezek mindenképpen intencionális kifejezések. De az *Ilma ihatna/ Ilma tudna inni egy pohár vizet.,* vagy a *Volna itt két pohár víz.* mondatok már nem feltétlenül. (Ez utóbbi alanya nem személy, tehát ezt a későbbiekben ki tudjuk zárni. De a *lenne egy kérdésem/kérésem* már megfontolandó eset, így az ilyen típusú közlésekre szintén gráfokat hozhatunk majd létre). Ugyanígy a *ha adnál, hogyha elmondanád, mintha nem itt lennék,minthogyha még húsz éves lennék, lehetne vagy lehetett volna rosszabb is,* stb. kifejezések sem utalnak a szándékra, hanem lehetőségeket ragadnak meg. Ezek kiszűrésére készült az alábbi gráf.



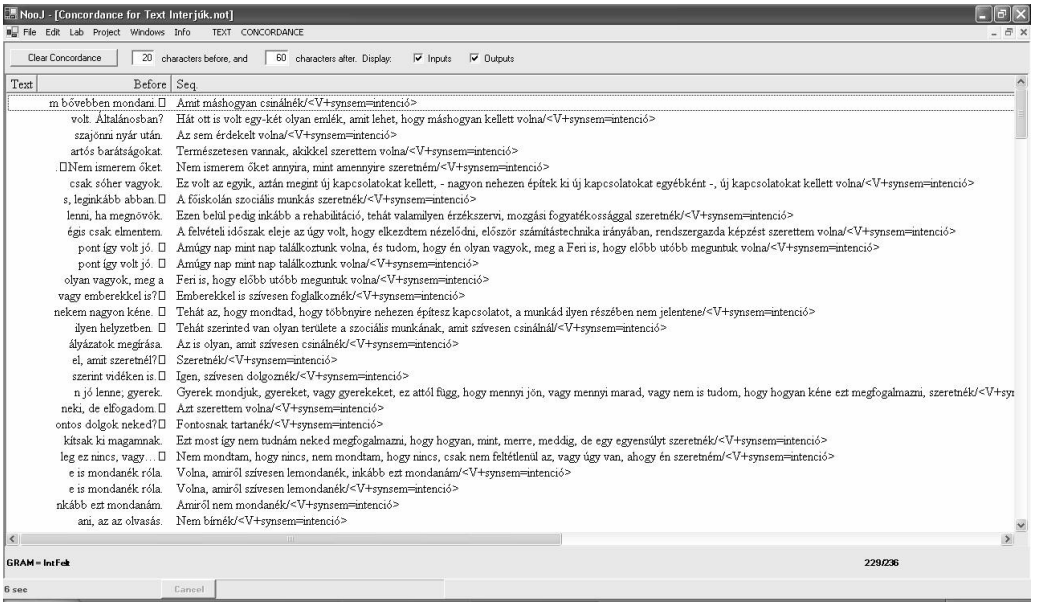
3. ábra Feltételes mód gráfja

A gráf célja tehát, hogy azokat a feltételes módú igéket azonosítsa, melyek előtt a mondatban nem szerepel a ha, hogyha, mintha, minthogyha kifejezés, illetve az ige nem a volna, lenne, lehetne vagy kellene valamely alakja. A gráf segítségével az is megoldható, hogy mindezek kizárása mellett figyelembe vegyük, hogy az ige állhat a mondat elején, illetve megelőzheti egy vagy több szó is. (Fontos szempont, hogy a szöveg mondatok mentén bontottuk egységekre, tehát a gráf a mondathatárokat nem tudja átlépni.)

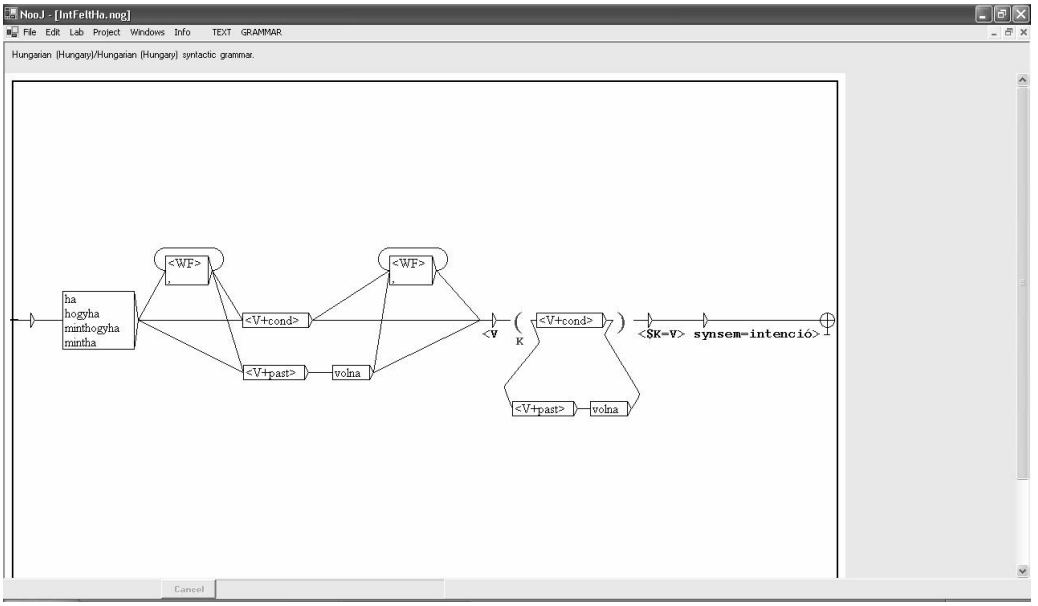
A konkordancia jól mutatja, hogy a gráf a megfelelő alakokat jeleníti meg (pl. a volna nem kap intencionális jelölést, illetve nem szerepelnek benne a nem kívánt alakok). A kézi lejelöléssel egybevetve azt mondhatjuk, hogy nem szerepelnek nem kívánt találatok, illetve a 236 helyes bejelölés mellett kilenc eset van, amit a gárf nem talált meg, tehát nem szerepel a konkordanciában. Ezek alapján még nem vonhatunk le konkrét statisztikai következtetéseket, de az eddigi eredmények bíztatóak.

Azonban azáltal, hogy a lehetőség eseteit kizártuk, vesztettünk intencionális feltételes alakokat is. Például: *Ha Ilma inkább bort inna, magam is vele tartanék*., *Ha velünk jöttél volna, te is jót mulattál volna.* Ezeknél a mondatoknál az első tagmondat a feltételt vagy lehetőséget jelöli, míg a második feltételes módú ige intencionális. Ezt az előző gráfunkal kiszűrtük, mivel a gráf elakadt a ha, hogyha, stb. kötőszavaknál. Így erre a lehetőségre egy újabb lokális nyelvtant alakíthatunk ki, mely, bár még nem végleges formában, de a következőképpen nézhet ki.

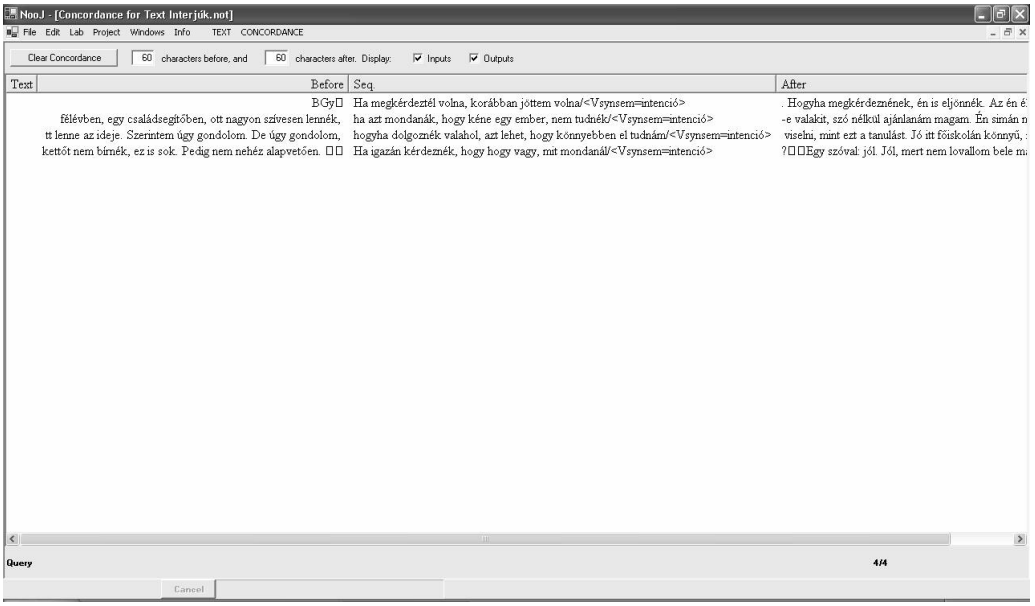
Az alábbiakban tehát az első feltételes-gráf konkordanciája, majd a második feltételes gráf és annak konkordanciája látható.



4. ábra A feltételes mód konkordanciája



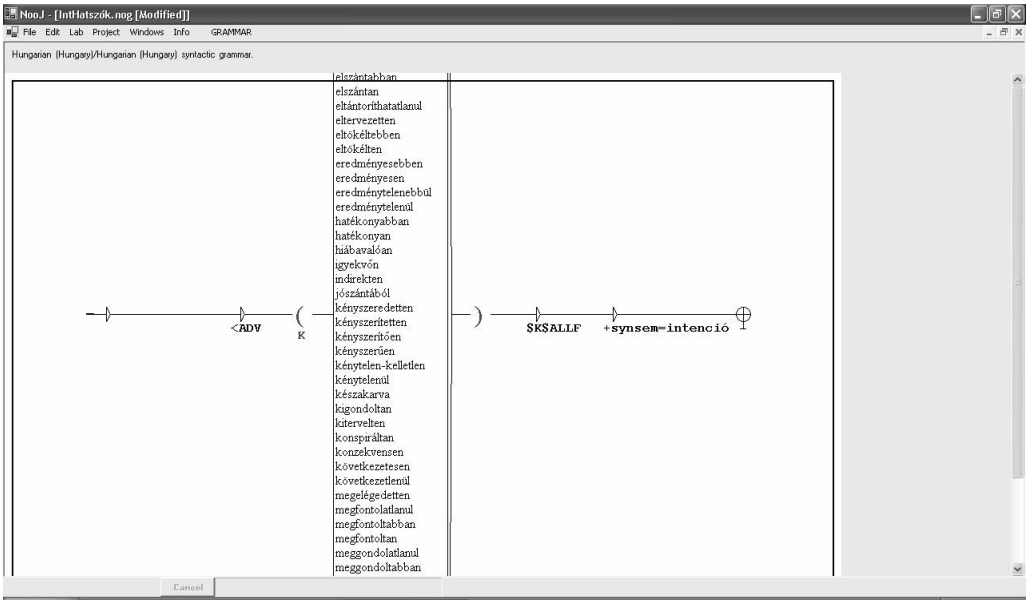
5. ábra Feltételes mód 2. gráfja



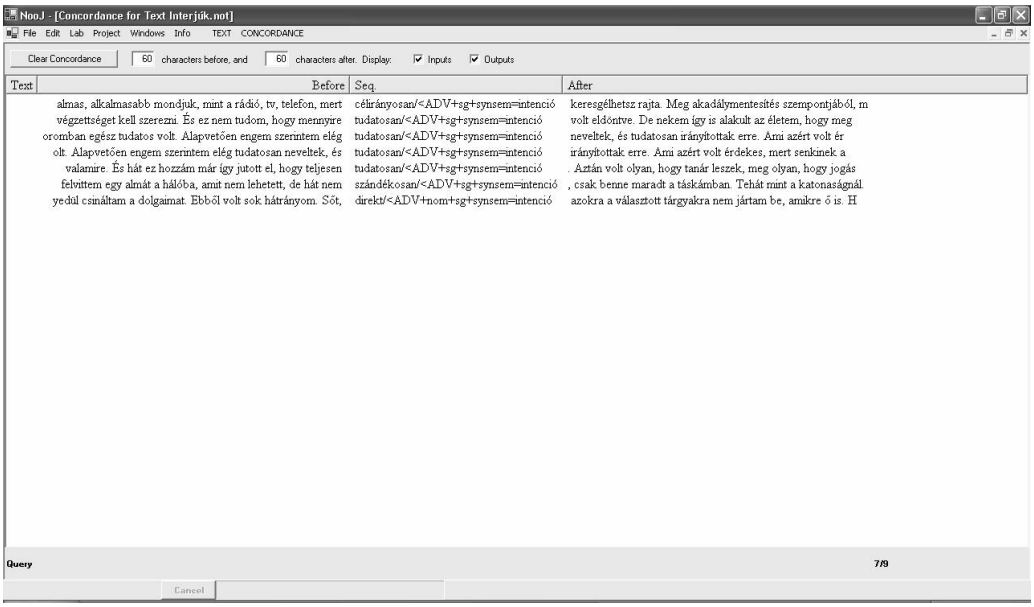
6. ábra 2. feltételes gráf konkordanciája

2.4 Határozószók

Az intencionalitást hordozó igei transzformációk egyik jelentős típusa a módhatározószavak használata. Ismét Ilmához visszakanyarodva: *Ilma direkt vizet iszik, vagy készakarva öntötte ki a vizét, mert nem szereti.* stb.. A száz elemből álló listát egy gráfba helyezve lehet lefuttatni a szövegen, méghozzá úgy, hogy csak az adott szóalakra keresünk, mivel nincs szükségünk például az azonos szótővel rendelkező melléknevekre.



7. ábra Határozószók gráfja



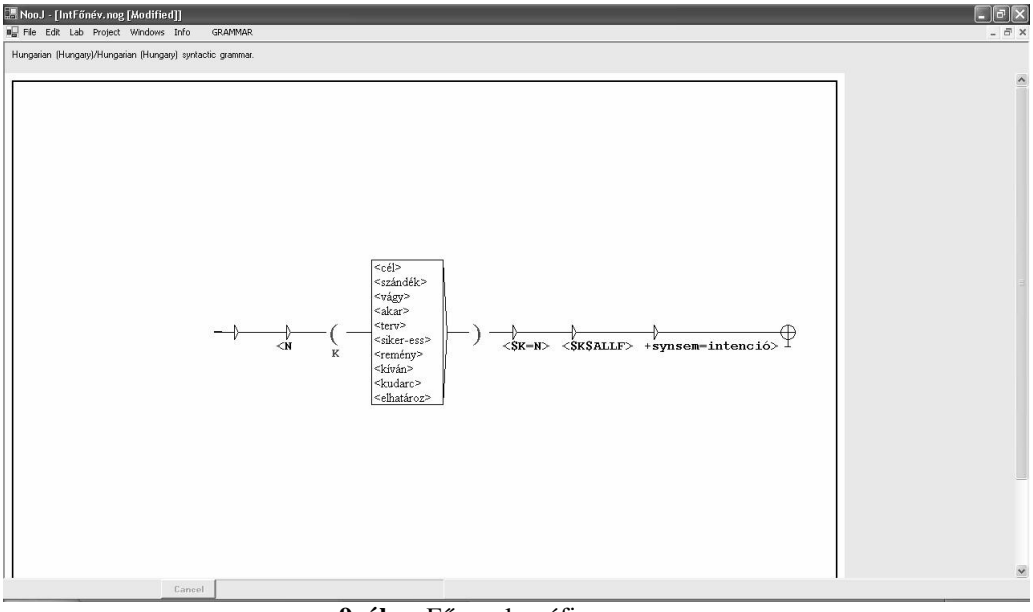
8. ábra Határozószók konkordanciája

A határozószavak azonosítása a szövegben megbízhatóan működik, a találatok az eddig vizsgált szövegeken, melyekkel jelenleg dolgozunk, megfelel a kézi lejelöléseknek. A nagyobb szöveganyagon statisztikailag is alátámasztható megbízhatóság ellenőrzése folyamatban van.

2.5 Főnevek

A jelenleg bemutatandó tényezők közül a főnevek csoportja zárja a sort, melyeket az alábbi gráf tartalmaz.

Ilma esetében ez úgy néz ki, hogy *Ilma vágya/ szándéka/ kívánsága, hogy víz helyett egy pohár finom bort igyon., Ilma akaratából kifolyólag folyóbort iszunk., stb.*.



9. ábra Főnevek gráfja

Ebben az esetben is az L-változó a listában szereplő szavak összes alakját tartalmazza, melyek közül azonban csak a főnevek, azoknak viszont minden ragozott formája engedélyezett.

Azt mondhatjuk, hogy ha a szövegben ilyen közvetlenül jelenik meg az intenció, mint téma, akkor az már a mi szempontunkból releváns, ezért fogadunk el minden megjelenési formát.

2.6 Konzekvenciák

A fentiekben leginkább az az irány körvonalazódott ki, amerre a kutatás tart. További gráfok segítségével méginkább árnyalni kell az egyes eseteket, illetve a cikkben és az előadásban nem szereplő tényezőket. Lényeges, hogy a találatokat a szövegben meg lehet jelölni (+synsem=intenció-ként szerepelt a gráfban és a konkordanciában, melyet ha leellenőriztünk, akkor a szöbegben is annotálhatunk, így láthatóvá válik majd a szöveg intencionális struktúrája), ami alapján a későbbiekben rá lehet keresni az egyes alakokra. A jelenlegi fázisban még pontos statisztikai adatok nem állnak rendelkezésre, de mégis megemlítenék pár kirajzolódni látszó tendenciát.

Jelen modulon leginkább vakokkal, drogosokkal és depressziósokkal készült interjúk szövegei alapján dolgoztunk. Itt a vártnak megfelelően a depressziósok szövegeiben úgy tűnik, hogy kisebb az intencionalitás mértéke. Ami már most látszik, az az, hogy ezen modul értelmezése szoros összefüggésben áll az aktivitás-passzivitás, illetve a későbbiekben kidolgozandó modalitás modullal, mely utóbbi a már fentebb emlegetett lehetségesség és kényszer megragadását tűzi ki célul. Elgondolásunkban aszerint lehet a későbbiekben kategóriákat meghatározni, hogy különböző csoportok az intencionalitás megjelenítését szolgáló eszközöket hogyan és milyen mértékben használják (pl. érdekes lesz az intencionalitáson belül leginkább a feltételes mód, illetve a gátoltság és a modalitás modul együttesének vizsgálata), illetve hogy ennek milyen pszichológiai háttere van.

Bibliográfia

1. László János: A történetek tudománya Új Mandátum Könyvkiadó, Bp. (2005)
2. Bakos Ferenc: Idegen szavak és kifejezések szótára Akadémiai Kiadó, Bp. (1984)
3. Pléh Csaba: Pszichológiatörténet Gondolat Kiadó, Bp. (1992)
4. Jerome Bruner: Valóságos elmék, lehetséges világok Új mandátum Könyvkiadó, Bp. (2005)

Az elbeszélések érzelmi aspektusának vizsgálata tartalomelemző program segítségével

Fülöp Éva[1], László János[2]

[1] PTE-BTK, Pszichológia Doktori Iskola
7624 Pécs, Ifjúság útja 6.
petymeg81@freemail.hu

[2] MTA Pszichológiai Kutatóintézet
1132 Budapest, Victor Hugo u. 18-22
laszlo@mtapi.hu

A tanulmány az NKFP 6/074/2005 számú pályázat támogatásával készült.

Kivonat. Egy szöveg érzelmi tartalma nagyon fontos információkat hordoz egy pszichológiai elemzést végző szakember számára. Jelen munka a NooJ tartalomelemző program segítségével különböző elbeszélések érzelmet kifejező nyelvi megnyilvánulásainak analízisét végezi. A program a lexikai elemek megragadásán kívül a szövegek morfoszintaktikai elemzését is lehetővé teszi. Lokális nyelvtanok alkalmazásával lehetőséget ad arra, hogy egy sokkal testreszabottabb, komplexebb, specifikusan a magyar nyelv egyedi kívánalmaihoz alkalmazkodó elemzés jöjjön létre. Ezeket a különleges szabályszerűségeket a programba illesztett gráfok létrehozásával kezelhetjük.

1. A szótár szerkesztésének alapjai

Nemzetközi vizsgálatok, főként angol nyelven, már végeztek hasonló tartalomelemzést az elbeszélések érzelmi vonatkozásainak kinyerésére. A James Pennebaker által használt érzelemszótárt [6] az érzelmi hívószavakra kapott asszociációk segítségével válogatták össze, így az egy előre meghatározott szerkesztési koncepciót mellőzve, a megkérdezett személyek által adott érzelem jellegű szavak eklektikus listáját jelenti. Ennek és más, eddig alkalmazott tartalomelemzőknek [2] a segítségével az elbeszélések lexikai analízise valósulhatott meg.

Magát a szótárt a Magyar Értelmező Kéziszótár szavainak felhasználásával hoztuk létre kiválogatva azokat a szavakat, melyek érzelmi jelentést hordozhatnak. Ezután két független bíráló ellenőrizte a válogatás helyességét. Külön csoportot képeznek a szótárban az érzelmet kifejező igék, melléknevek, határozószók, főnevek és idiómák. Az érzelemszótárban szereplő szavakat különböző kategóriákba soroltuk.

2. Kategórizáció

2. 1 Pozitív – negatív

Mivel a kellemes- kellemetlen dimenzió az egyik legalapvetőbb szintje a tapasztalásnak, elsőként elkülönítettük az érzelmeket azok valenciája szerint: pozitív, negatív és semleges/ kontextusfüggő osztályokba.

2.2 Affektus, érzés, érzelem, intenció

Az érzelmeket sokféleképpen lehet tárgyalni, a szakirodalomban [3,4,7,8,9,10] a leggyakrabban az érzelem, érzés-hangulat, affektus terminusokban jelennek meg. A pszichológiai kutatások szerint ezek különböző jelenségeket fednek le. A fentieket kiegészítve egy negyedik kategóriával, amely az intencionális jellegű érzelmi folyamatokat tartalmazza, besoroltuk az érzelemszótár szavait az affektus, az érzelem, az érzés és az intenció-motiváció kategóriákba. Ezen osztályok alkalmazásáról, elkülönítéséről úgy gondoljuk, hogy hasznosnak bizonyulhatnak a későbbi, a tartalmi elemzést követő értelmezési folyamatban.

Affektus alatt azokat az emocionális folyamatokat értjük, melyekben az átélőnél egy esemény, történés hatására ún. aktivációs kontúr (izgalmi szint) növekedés vagy csökkenés következik be. Ilyennek tekinthetők, pl. feldühödik, elbágyad. Az érzelmek osztályába azok az emóciót kifejező szavak tartoznak, melyek magukba foglalják az érzések kognitív kiértékelését, minősítését. Erre példa a szégyen vagy a büszkeség. Érzésnek (feeling) azok a kifejezések számítanak, melyek egy általános érzésről, hangulatról szólnak, pl. fájdalom, jókedvűség. Végül az utolsó kategóriát az intenció-motiváció tartalmú érzelmek képezik, így pl. vágy, epekedés, stb.. A 4 alapkategórián kívül külön kezeljük azokat a cselekvéseket, melyek közvetlenül implikálnak érzelmeket (pl. sír, elpirul)

2.3 Alapérzelem-szociális érzelem

Másik fontos osztályozás az alap- és a szociális érzelmek elkülönítése volt. Az alapérzelmek közé soroljuk azt a hat emóciót, melyek univerzálisan megjelennek minden kultúrában [1], melyek alapvetően meghatározzák intrapszichés és interperszonális működésünket. Eszerint az ‘öröm’, ‘bánat’, ‘undor’, ‘félelem’ és ‘meglepődés’ nyelvi kifejezései kerültek kigyűjtésre. S bár az érzelmek nagy része személyközi helyzetben jelenik meg, mégis vannak egyesek, melyek kimondottan társas közeghez kötöttek. Itt a társas érzelmek alatt a „másokkal létrejött valódi, elképzelt, elővételezett vagy felidézett találkozások által kiváltott emóciókat” [5]értjük. Ilyenek pl. a féltékenység, társas szorongás, sértettség, szégyen, zavar, büszkeség.

2.4 Közelítés-távolítás

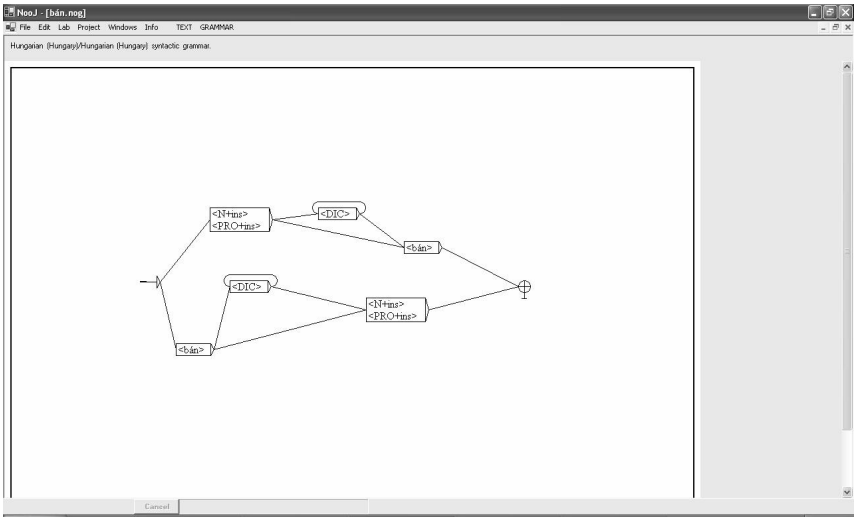
A társas kapcsolatok szempontjából igen jelentős a pszichés közeledés- távolodás szabályozása. Ennek szövegszintű megjelenítésének egyik eszköze a pozitív kapcsolódást, közelítést (pl. szeret) és az emocionális távolítást (pl. elhidegül) kifejező érzelmek használata. Ezeket is különválogattuk.

Az így megalkotott érzelemszótár szavai bekerültek a NooJ programba. Számos olyan helyzet adódik azonban, amelyre nyelvtani szabályt, megkötéseket kell szerkesztenünk az érvényes találatok érdekében.

3. Problémák

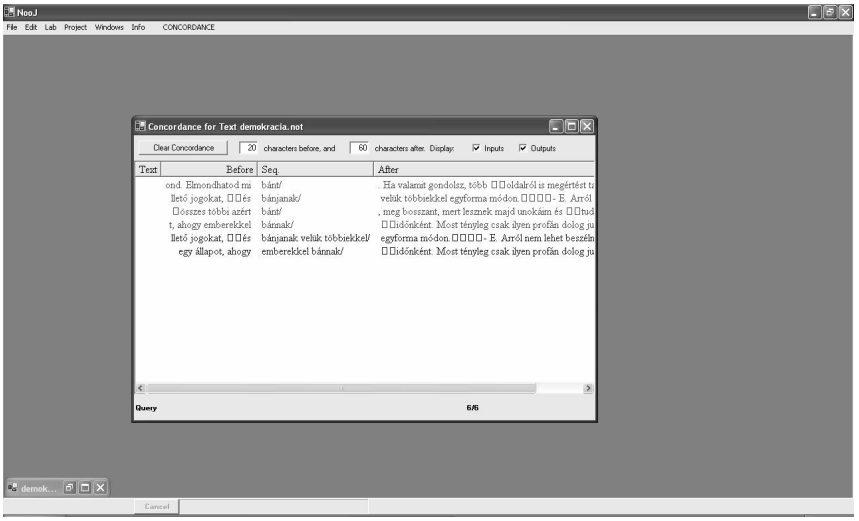
3. 1 Vonzatkeretek

A problémás esetek nagy részében egy nyelvi alak csak akkor minősül érzelemnek, ha bizonyos vonzattal szerepel. Ilyen pl. 'bír valakit/ valamit', 'fél valakitől/ valamitől', 'álmodik valakiről', 'bízik valakiben', stb. A következőkben bemutatott 'bán' ige is ilyen: kizárólag a 'bán valamit' értelmében fejez ki érzelmet. Ebben az esetben azonban gazdaságosabb a kizáró esetre készíteni egy gráfot, hiszen a 'bán valamit' lehet egy egyszerű tárgyrag ('bán vmit'), de lehet egy tárgyi mellékmondat is ('bánja, hogy..'), sőt némely esetben önmagában, vonzat nélkül áll (pl. 'most már bánom'). Tehát készítünk egy szabályt arra az esetre, ha –val, -vel rag követi ('bánik valakivel'), és ezeket az eseteket emeljük ki a találatok közül.



1. ábra. Példa a vonzatkeret problémájára: a 'bán' igére szerkesztett gráf.

Lefuttatva a gráfot megkapjuk a 'bán' ige összes alakja (az ábrán a felső négy találat) közül azokat, amelyek a 'bánik valakivel' értelemben szerepelnek (az ábrán az alsó két találat).



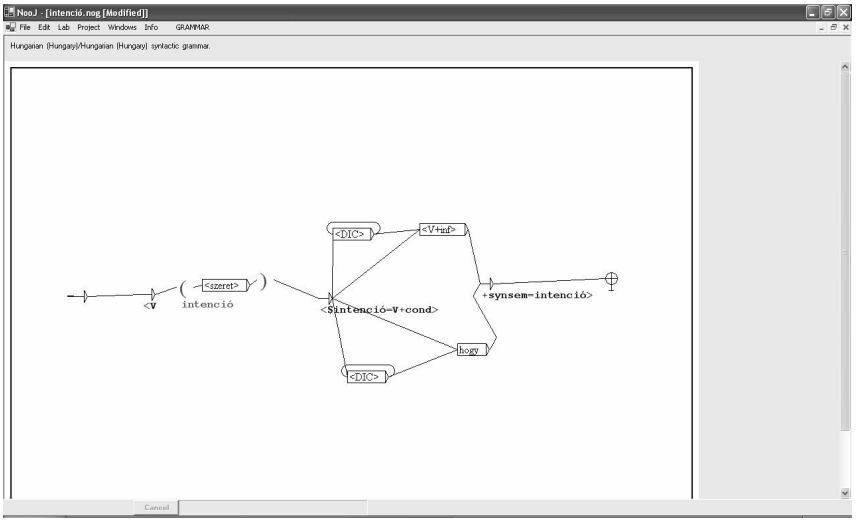
2. ábra. A ‘bán’ igére szerkesztett gráf konkordancieredményei.

Vannak olyan esetek, amelyekben az ilyen kitételek sem elegendők, viszont nincs más szabályszerűség, mely elkülönítené a különböző jelentésű alakokat. Pl. a ’bámul’ ige előfordulhat a tisztel és a hosszan néz értelmében is, de mindkettőnek lehet ugyanaz a vonzata. A rokon értelmű szavak nemcsak az érzelmek szempontjából releváns igéknél, de a főneveknél is előfordulnak, a különböző jelentésű alakokat még a kontextus sem különíti el, nincs specifikus velejárója pl. a ’méreg’, ‘keserűség ‘szavaknak a két különböző értelmében.

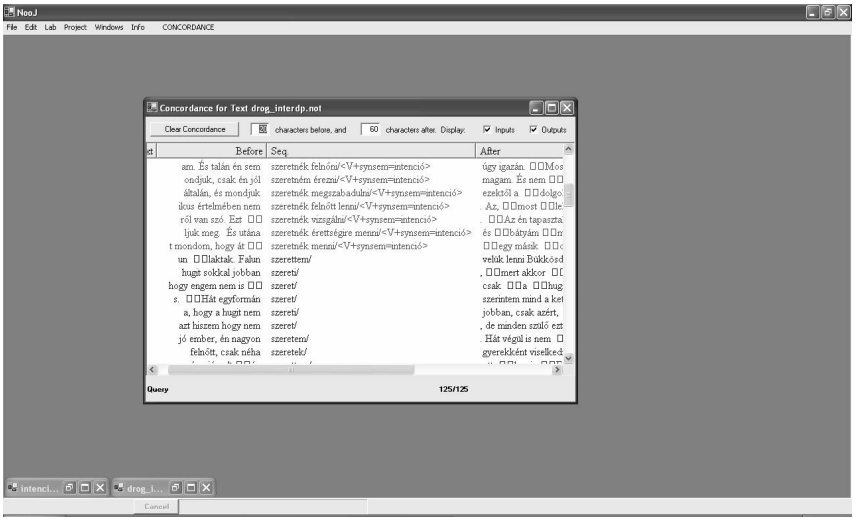
Mégis hasznunkra lehetnek ilyenkor az előfordulási gyakorisági adatok, hiszen bizonyos alakokkal feltehetőleg igen ritkán találkozunk.

Vegyük erre példaként a ’szeret’ igét, amely az egyik leggyakrabban előforduló érzelemszó. A ’szeret’ általában egy pozitív érzelmi viszonyulást jelent, néha azonban a cselekvő szándékát, intencióját fejezi ki, pl. ’ szeretne jó lenni’. Ahogy ez utóbbin látszik, intencionális jelentésben gyakran főnévi igenév áll utána, de nem minden esetben (pl. ’szeret olvasni=érzelem’). Kiköthetjük azt is, hogy a ’szeret’ feltételes alakjában intenciót fejez ki, különösen, ha főnévi igenév követi. De magánál a feltételes módnál is előfordulnak kivételek. Pl. a ’szeretné, ha ráfigyelne’ mondatban nem lehet megkülönböztetni a két jelentést (’azt akarja, hogy ráfigyeljen/ szeretné őt, ha az ráfigyelne’) vagy a ‘szeretne olvasni, ha…’ mondatnál sem. A tárgyi mellékmondattal kifejezett forma intencionalitásra utal (’szeretné, hogy szót fogadjon’). Azokban az esetekben, amikor a ‘szeret’ igéhez nem feltételes módban tárgy kapcsoldik főnévi igenév nélkül (‘szeret valakit’), mindig érzelemről beszélünk. Ha tárgy követi a ‘szeret’ ige feltételes módban álló alakját, akkor két eshetőség áll fenn. Az ige határozatlan tárgyas ragozásánál többnyire intencióról van szó (‘változást szeretne’), kivétel ha személyes névmásra vonatkozik (‘szeretne engem’=érzelem), de nem áll utána főnévi igenév (‘szeretne engem megölelni’=intenció). Amennyiben a ‘szeret’ határozott tárgyas ragozású alakjáról van szó, akkor főként az a mérvadó, hogy személyre (‘szeretné a fiút’= érzelem) vagy nem személyre vonatkozik (‘szeretné a süteményt’= intenció)- a ‘ha’ mellékmondat sem különít el jól . Itt is elméletileg főnévi igenév áll mögötte, csak nincs kitéve (‘szeretné a süteményt megkapni’). Az is gyakran megesik, hogy nincs semmi vonzat

az ige mögött. Egyéb eshetőségek is vannak még, de a sok előfordulási lehetőség ellenére mégis az esetek nagy részében fennáll a feltételes mód és az intenció összefüggés a 'szeret' igénél, különösen főnévi igenévvel, hiszen így sokkal gyakrabban fordul elő ebben az értelmében. Az alábbi gráf, amely kiszűri az érzelmek közé nem tartozó alakokat, megközelítő pontossággal működik.



3. ábra. Egy másik példa a vonzatkeret problémájára: a 'szeret' igére szerkesztett intencionalitást jelölő gráf



4. ábra. A 'szeret' igére szerksztett gráf konkordanciaeredményei.

3. 2. Kivételt képező alakzatok

A NooJ programba beépített érzelemszavak nagy részére érvényes, hogy minden morfológiai alakjukban emóciót fejeznek ki. Néhány szóalakot azonban ki kel iktatni a szótárból, mivel azoknak már más, nem érzelmi jelentésük van. Egyenlőre a program a szótőre visszavezetve belevesz helytelen formákat is. Ezen helyzetben nem érdemes gráfban gondolkodni, hiszen nincsenek nyelvtani szabályszerűségek, egyszerűen csak bizonyos alakok kiszűrésének szükségességéről van szó. Ezek persze élő szövegeken lefuttatva kerülnek leginkább elő.

Egyik kezelendő nyelvtani forma a visszaható alak. Vagyis a személyközi helyzetekben ilyenkor az ágensre csak visszaható formában vonatkoztatható érzelem (pl. 'felpaprikázódik', de 'felpaprikáz' nem).

A műveltetés ritkán, de szintén szerepet játszhat egy szó érzelemmé minősítésében (pl. 'megbotránkozik'-'megbotránkoztat').

Egy szó befejezett és folyamatos melléknévi igenévi formában is sokszor más-más jelentést hordoz. Így pl. a 'feldúlt', a 'lehangolt', a 'megtört' érzelmek, a folyamatos alakok, tehát a 'feldúló', 'lehangoló', 'letörő' nem.

Mint fentebb említettük, igék, főnevek, melléknevek és határozószók is szerepelnek a szótárban. Külön, minden szónál át kell gondolni, hogy mindegyik formában van- e keresnivalójuk a gyűjtésünkben. Pl. amiknek van: 'rettenet', 'szánalom', 'síró', 'rokonszenvező', amiknek nincs: 'rettenetes', 'szánalmas', 'siralmas', 'rokonszenves' (esetleg 'neki' vonzattal).

Az érzelemkifejezésére szolgáló ige, az 'érez' kikeresése sem mindig helyénvaló. Egyrészt bizonyos morfológiai alakok helytelensége miatt (pl. 'érzelmes'), másrészt szemantikai kitételek miatt sem (az 'érez' szenzoros vonatkozásában nem belső lelki állapotra utal).

A szenvedő szerezet főként az elváló igekötős alakok problémáját veti fel (pl. 'le van sújtva', 'el van ragadtatva', 'le van törve') , ezt az idiómánál szereplő gráfhoz hasonlóan egy szabály szerkesztésével egyszerűen meg tudjuk oldani.

Az igekötőkkel külön is kell foglalkoznunk. Mivel egyes igét nem jelentenek érzelmet minden igekötővel (a 'nevet' igen, de a 'kinevet' már nem implikál érzelmet), ezért külön szerepelnek a szótárban azok, amelyek valóban érzelmi tartalmúak. Ezekben az esetekben számolnunk kell azzal az eshetőséggel, amikor elválik az ige az igekötőtől. Ez egy egyszerű gráffal orvosolható. A szótárt szövegeken lefuttatva kapunk olyan téves találatokat is, melyeknél a szótárban igekötővel nem szereplő szavak fordulnak elő, mert egy elváló igekötő szerepel utánuk, de a program nem veszi őket egybetartozónak és mivel így már megváltozik a jelentése, ezért ezeket külön ki kell zárnunk (pl. a 'derül' érzelem, de a 'kiderül' 'derül ki' alakja nem.).

3. 3 Személy-nem személy elkülönítés

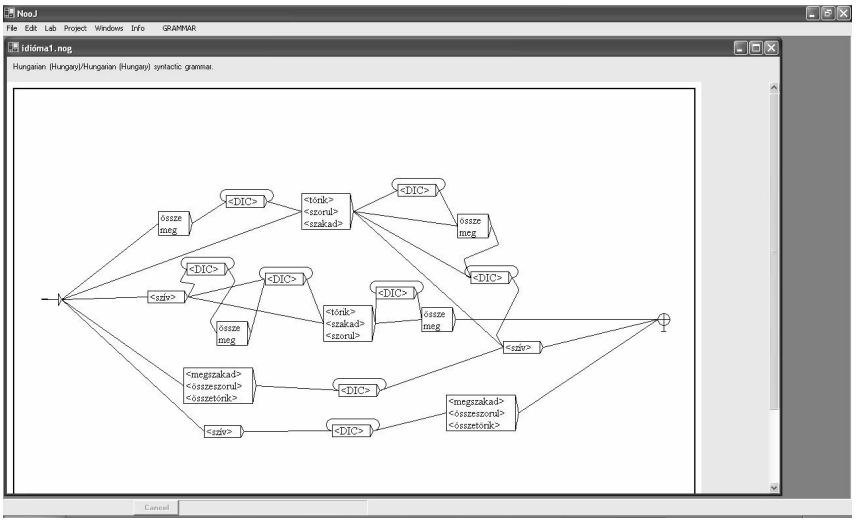
Érzelmek esetén sokszor az segít a szavak minősítésében, ha a szövegkörnyezetből kiderül, hogy élő személyre vonatkozik, nem tárgyra. Pl. 'szánt' (földet), 'szánt' - múlt időben (valakit). Ugyanígy az ágensnél sem mindegy, hogy személy-e. Pl. lehet 'derűs' az időjárás, de egy ember is, természetesen csak az utóbbi releváns a szöveg érzelmi tartalma meghatározásában.

3.4. Ágencia

Az érzelmek átélése szempontjából fontos tényező, hogy az elbeszélésben melyik emocionális állapot kire vonatkozik. Az érzelmi aspektus vizsgálatában elemzésünk fontos alapelve, hogy első körben mindig a narrátor perspektívájából tekintünk egy érzelmi helyzetre. Némely esetben egy másik személy kelti az adott érzelmet az átélőben, pl. 'valaki bosszant valakit'. Ilyen helyzetben maga a cselekvés (jelen példában a 'bosszantás') az elbeszélő szemszögéből természetesen még nem közvetít érzelmet. Ugyanez az aktus visszaható alakban azonban már igen ('valaki /valamit bosszant engem'). Ezen nyelvtani alakok egyikének- másikának használata a szövegben nagyon informatív lehet nemcsak az elbeszélő érzelmi történései, de a többi szereplő reprezentációja vagy akár az aktív –passzív átélés vonatkozásában is. Abban, hogy a személy milyennek látja a környezetében levő fontos emberek érzelmeit, különösen a felé irányulóakat, a nekik tulajdonított érzelmek elemzése segít. Mindkét fenn említett analízisnek fontos implikációi lehetnek, különösen a klinikai munkában. Konkrét példánkban így nem mindegy, hogy 'én bosszankodom' vagy 'engem bosszantanak'. Ugyanez a helyzet a 'megbánt', 'megfélemlít', 'felizgat', stb. szavaknál is, ezek a szótárban szintén külön jelzés alatt szerepelnek.

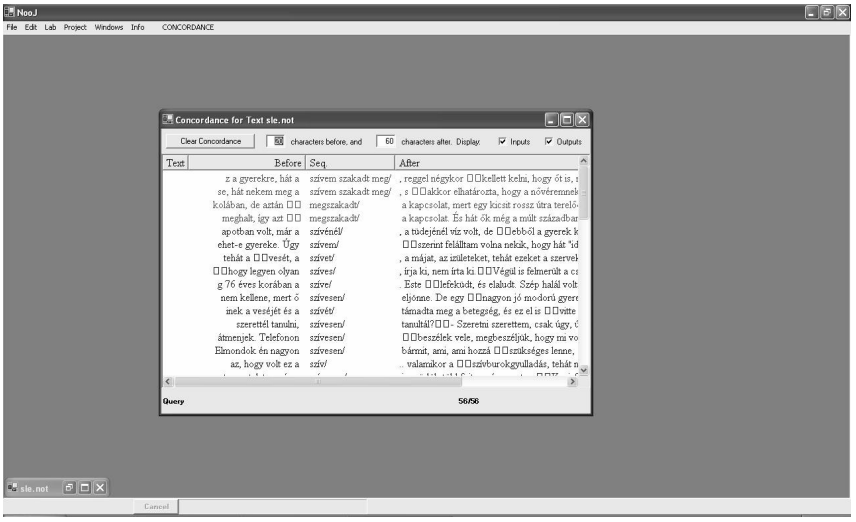
3.5. Idiómák

Érzelemszótárunk egyik csoportját képezik az idiómák. Ezeknél az okozza a nehézséget, hogy több szóból állnak, melyek azonban a magyar szórend lazasága miatt számos megfogalmazásban alkalmazhatók. Itt egy olyan idióma látható, amely ezen kívül még elváló igekötős igét is tartalmaz, ami tovább növeli az előfordulási lehetőségek számát.



5. ábra. Példa az idiómákra szerkesztett gráfra.

Ahogy a konkordanciatáblán látszik, a gráf megtalálta a keresett kifejezéseket (az ábrán az első két sor) és mást, hibás találatokat nem adott ki (a többi találat, mely a 'megszakad' és a 'szív' szavak külön beírásakor jön elő).



6. ábra. Az idióma példa konkordanciaeredményei.

Természetesen vannak ennél összetettebb, több tagot magukba foglaló kifejezések is, de szinte mindegyiknél megragadható egy olyan központi mag, mely alapján azonosítható az idióma,

pl. égnek áll a haja tőle = égnek+ <áll>+ <haj>+ PRO+abl/ N+abl

és ennek minden szórendi változata. Az idiómáknál a másik eshetőség az, hogy egy-egy szó, ami önmagában még emóciót fejez ki, kifejezésekben, bizonyos szókapcsolatokban már minősíthető érzelemnek.. Pl ha a 'tart' igére szerkesztünk egy vonzatkeretes gráfot, melyben kikötjük, hogy csak a 'tart valamitől/ valakitől' értelmében érvényes találat, akkor beleütközünk a 'távolinak tart magától', vagy 'távol tart valamitől', stb. kifejezésekbe, melyeket külön ki kell zárni.

Ugyanígy vannak rögzült szófordulatok, melyek szelektálásra szorulnak, hiszen így már nincs meg az érzelmi töltetük. Pl. ne haragudjon, de…, hála Istennek, azt szeretném mondani, stb.

Összességében elmondható a fentiekben leírt problémás esetekről, hogy sok kikötést igényelnek. Ugyanakkor nem elhanyagolandóak az előfordulási gyakoriságok sem. Elképzelhető, hogy bizonyos esetek egyáltalán nem fordulnak elő a hétköznapi beszédben. Az el nem dönthető esetekben pedig azt kell mérlegelni, szintén a gyakorlati tapasztalat alapján, hogy melyik definícióval, korlátozással válik be jobban a sok szövegen való ellenőrzés után egy- egy szó alkalmazása.

Az érzelmi kapcsolat nemcsak a pszichopatológia területén, de bármely tárggyal szembeni viszonyulás szempontjából sokatmondó lehet.

Az érzelemszótár kidolgozásának kezdeti lépései után további pontosítások, szabályszerűségek szerkesztése következik, de az első alkalmazások alapján már így is elég nagy találati arányra számíthatunk.

Bibliográfia:

1. Ekman, P.-Friesn, W. V.-Ellsworth, P. Emotion in the Human Face. New York: Pergamon Press. (1972)
2. Hogenraad, R., Daubies, C., & Bestgen, Y. Une théorie et une méthode générale d'analyse textuelle assistée par ordinateur. Le système PROTAN (PROTocol ANalyzer) (Version March 2, 1995). Louvain-la-Neuve, Belgium: Psychology Department, Catholic University of Louvain. (In French). (1995)
3. Izard, C. E. : Four systems for emotion activation: Cognitive and non-cognitive processes.Psychological Review, 100, (1993) 68–90.
4. Lazarus, R. S. . Emotion and adaptation. New York: Oxford University Press.(1991)
5. Leary, M. R.Érzés, megismerés és társas érzelmek in Forgács, J. (Forgas, J.P.) (szerk.): Érzelem és gondolkodás, Kairosz, Budapest (2000)
6. Pennebaker, J.W., Booth, R.J., & Francis, M.E.. Linguistic Inquiry and Word Count: LIWC 2006. Austin, TX: LIWC (www.liwc.net).(2006)
7. Plutchik, R.. The emotions: Fact, theories, and a new model (Rev. ed.). Lantham, MA: Universityof America Press.(1991)
8. Zajonc, R. B.. Feeling and thinking. American Psychologist, 35,(1980) 151–1275.
9. Zajonc, R. B. On the primacy of affect. American Psychologist, 39, (1984) 117–123.
10. Zajonc, R. B. Emotion and facial efference: A theory reclaimed. Science, 228, (1985) 15–21.

A kauzális kohézió vizsgálata az Intex számítógépes eszközzel

Mészáros Ágnes[1]

Papp Orsolya[2]

[1]SE Pszichiátriai és Pszichoterápiás Klinika
H-1083 Budapest, Balassa u. 6.
meszaros@psych.sote.hu
[2] PTE BTK Pszichológia Doktori Iskola
H-7624 Pécs, Ifjúság útja 6.
papporsi@lycos.com

1. Bevezetés

A tanulmány oksági kapcsolatok természetes nyelvi szövegekben történő automatikus azonosításának egy kísérletét mutatja be. Az MTA Pszichológiai Kutatóintézet 2004 tavaszán felvett élettörténeti szövegkorpuszát[78] az Intex számítógépes eszköz [9] segítségével elemeztük, annak a szemléleti és módszertani hagyománynak a talaján, mely az elbeszélő bizonyos személyiségbeli tulajdonságaira az elbeszélésbeli események leírásának szó- és mondatszinten megjelenő nyelvi-formai jegyeiből következtet [12, 3].

Schank és Abelson [8] óta a kognitív pszichológiában a világról és önmagunkról alkotott fogalomrendszerünk elsődleges szervező elveként tételezett oksági kapcsolatok vizsgálata és formalizált leírási kísérlete természetes módon illeszkedik bele az identitást az önmagunkról alkotott történetek, céljaink szerint változó összességeként leíró narratív pszichológiai kutatás azon törekvésébe, hogy az önéletrajzi elbeszélés egyes dimenzióit (például kauzális szerveződését) automatizált eszközök segítségével ragadja meg. A pszichológiai kérdésfeltevés tehát a következő elemekből tevődik össze: az okság típusainak konceptualizálása és pszichológiai mérőeszközeinek feltérképezése az egyik oldalon, a másikon pedig a narratívumok összefüggésrendszerét megteremtő oksági jegyek nyelvi operacionalizálása és automatikus felismerésükre tartalomelemző szoftver fejlesztése [1]. Tanulmányunkban a következő pontban az oksági kategóriákat Trabasso és van den Broek [11] modellje alapján tárgyaljuk, ezután a nyelvi markerek azonosítására és automatikus felismerésére teszünk javaslatot.

[78] A vizsgálatban résztvevő személyek hat személyes narratívumot idéztek fel: első emléküket, teljesítményt, veszteséget, félelmet, egy fontos személyhez kötődő jó és rossz kapcsolati élményt.

1.1 A kauzalitás egy modellje és mérési lehetőségei

Az oksági kapcsolatok kontextusfüggő, logikai-szemantikai kritériumok alapján végzett vizsgálatára szoros történetvezetésű, tanulságra kiélezett mesékben Trabasso és mtsai tanulmányai [12, 13] szolgálhatnak mintául. A narratívumok egyes eseményeinek fontosságát és felidézhetőségét az esemény manuálisan azonosított közvetlen oksági kapcsolatainak számával, valamint egy oksági láncon való elhelyezkedésével hozták összefüggésbe. Elemzési módszerük különböző korpuszokon történő alkalmazása azonban arra a következtetésre vezette őket, hogy nem lehet figyelmen kívül hagyni a vizsgált narratívum makrostruktúráját, a történetnyelvtanok által azonosított rekurzív szerkezeti elemek sajátos felépítési sorrendjét. Rekurzív tranzíciós hálózat modelljükben [11] ennek folyományaként a narratívumokat véges esemény–kategóriák (szetting, cél, elérési kísérlet, reakció, kimenetel) és oksági tipológiába rendezett viszonyok láncolataként írják le. Mivel korábban a narratív koherencia vizsgálatakor [6] amellett érveltünk, hogy csupán a történetstruktúra speciális szerkezeti elemeinek azonosítási képessége nyújthatna biztos kiindulópontot a narratív pszichológiai kérdések megválaszolásában, az ennek részét képező kauzális koherencia elemzésekor is egy olyan modellt preferálunk, mely a szó- és mondatszintű elemek mellett a narratívum makrostruktúráját is figyelembe veszi. Trabasso és van den Broek modellje ezek alapján megfelelő elméleti keretként szolgál különböző oksági viszonyok[79] felismeréséhez, nem nyújt azonban kidolgozott támpontokat szövegek automatizált, számítógépes elemzéséhez.

A Memphisi Egyetem Pszichológiai Kutatócsoportja által angol nyelvre kifejlesztett COH-METRIX szövegelemző program felépítése és általános működési elvei [2] ebből a szempontból úttörő szerepet töltenek be. A szoftver intencionalitás moduljától elkülönített kauzalitás modul szószinten vizsgálja az egyes (nem feltétlenül narratív) szövegek kohézióját egy olyan aránypárral jellemezve az elemzett szövegeket, melyben a számláló az oksági viszonyokat jelző partikulák (specifikus kötőszavak, például *so that*, *because,* valamint az oksági viszony jelenlétére általánosságban utaló igék és segédigék, például *cause, make*) számát tartalmazza, a nevező pedig a WordNet szemantikai hálóban oksági jeggyel jelölt igék (például *kill*) számát.

Magyar nyelvre fejlesztett, természetes nyelvű önéletrajzi narratívumokat elemző programunk távlati célja éppen az automatizált szóalapú és a logikai-szemantikai kiindulású, manuális elemzések közötti eltérések kezelése. Előadásunkban ennek első lépéseit mutatjuk be, az Intex számítógépes eszközzel kifejlesztett, Trabasso és van den Broek [11] egyes oksági típusai mentén összegyűjtött szavak és kifejezések bizonyos szintaktikai és szemantikai tulajdonságait is figyelembe vevő gráfok működésén keresztül.

---

[79] Az általunk végzett elemzéshez az eredeti tipológiából a Motivációs (célt kifejező szavak jelenléte), a Pszichológiai (belső állapotváltozást, reakciót leíró szavak, kifejezések), valamint a Fizikai/Eredmény típusú okságot (két cselekvést vagy egy cselekvést és egy állapotváltozást kifejező ige egymáshoz közeli jelenléte) használjuk fel. Az eredeti modell részletes leírását lásd Trabasso és van den Broek [11], a változtatások indoklását pedig a tavalyi kötetben [7].

1.2 A kauzalitás nyelvi megjelenése

Az oksági kapcsolat alapvetően propozíciók *közötti* viszonyra vonatkozik, ami a történetekben két esemény közötti relációt feltételez. Trabasso és van den Broek modelljében ez tagmondatpárok vizsgálatában jelentkezik olyan módon, hogy az egyik esemény (esemény1) az egyik, míg a másik (esemény2) a következő tagmondatban található. További megszorításokat is alkalmaz a nyelvészet az igei oksági kapcsolat jellemzésére [4]:

esemény1 *oka* esemény2-nek, ha

1. esemény1 implikálja esemény2-t,
2. esemény2 időben nem előzi meg esemény1-et,
3. és esemény1 jelenléte szükséges feltétele esemény2 bekövetkezésének (kontrafaktuális feltétel).[80]

Elemzésünk kettős irányú: a szövegkorpusz manuális vizsgálatakor, amely a számítógépes alkalmazás validálása egyben, a fenti három feltétel teljesülésekor azonosítunk oksági kapcsolatot. Azonban modulunk jelenlegi felépítése nem párok azonosítására alkalmas, csupán egyes szavakra vagy több szóból álló kifejezésekre kereshetünk vele. A logikai kritériumok érvényesülése pedig igepárokból álló lexikonok fejlesztésén keresztül valósulhat meg.

1.3 Az elemzés eszköze

Az elemzést az Intex számítógépes eszközzel végezzük [9], melyet természetes nyelvek formalizált leírására fejlesztettek ki. Az eszköz nagyméretű korpuszok valós idejű feldolgozását teszi elérhetővé: a gyors elemzés lehetőségét az nyújtja, hogy mind a bemeneti szöveget, mind a nyelvtanokat tömörített formában tárolja.

A programot lexikalista nyelvelemzésre fejlesztették ki, melynek alapjai az elektromos szótárak, és véges állapotú gráfokban ábrázolt nyelvtani szabályok. A szótárakban a szavak morfoszintaktikai és szemantikai jegyei egy szinten vannak kódolva, így a gráfokban mindezen információra egyidejűleg lehet hivatkozni. Az eszközt adottságai különösen alkalmassá teszik tartalomelemzési feladatokra, hiszen lehetőség van előre összeállított lexikonok elemeinek – akár szemantikai jegyek alapján történő – együttjárási vizsgálatára.

2. Az okság–modul felépítése

Az okságra vonatkozó elméleti modell kategóriáinak megfelelően három gráfot szerkesztettünk az egyes kauzalitás típusok, a történetben cél jelenlétét jelző Motivációs,

[80] Trabasso és van den Broek az oksági típusok manuális azonosításában használják ezeket a logikai feltételeket, elsősorban a kontrafaktuális tesztre támaszkodnak. Ezzel szemben például a WordNet szemantikai rendszerben a harmadik kritérium teljesülésétől bizonyos igepárok esetében eltekintenek (például *give–have*) [5].

belső állapotra vagy reakcióra utaló Pszichológiai, illetve két esemény közötti Fizikai viszony felismerésére. Jelen tanulmányban az utóbbi két gráfot mutatjuk be részletesen, az ágens intencionális hozzáállását jelző igei szerkezeteket és főneveket azonosító Motivációs okságot korábban [7] vizsgáltuk. Hangsúlyoznunk kell, hogy a modul kialakításának jelenlegi stádiumában kizárólag az adott korpuszban előforduló szavak és kifejezések feldolgozására törekszünk.

2.1 A pszichológiai okság azonosítása

A pszichológiai okság azonosítását az emocionális állapotokat leíró szavak felől közelítettük., melyek lehetnek önmagukban álló szemantikailag specifikus főnevek (például *fájdalom, katasztrófa*), melléknevek (például *szörnyű, boldog*), igék (*csalódik, szorong*), illetve az érzés/érez szó előtt álló melléknevek vagy határozószók.
Az alábbi példák ezen gráf találatainak természetes nyelvi környezetben való előfordulásait mutatják:

„Tehát például amikor a legjobb barátom 2002. december 21.-én öngyilkos lett.{S} Ez egy elég súlyos veszteség volt."

„Amszterdamban is elõfordult, hogy a lánynak beszólt egy fiú az utcán, egy tök ismeretlen férfi, és a lány elkezdett futni utána, a srác az pedig legyintett, hogy 'úristen, most mit csináljak, hát most akkor ezzel mit kezdjek, fussak utána, hát csináljon, amit akar', de közben folyamatosan szenvedett és dühös volt."

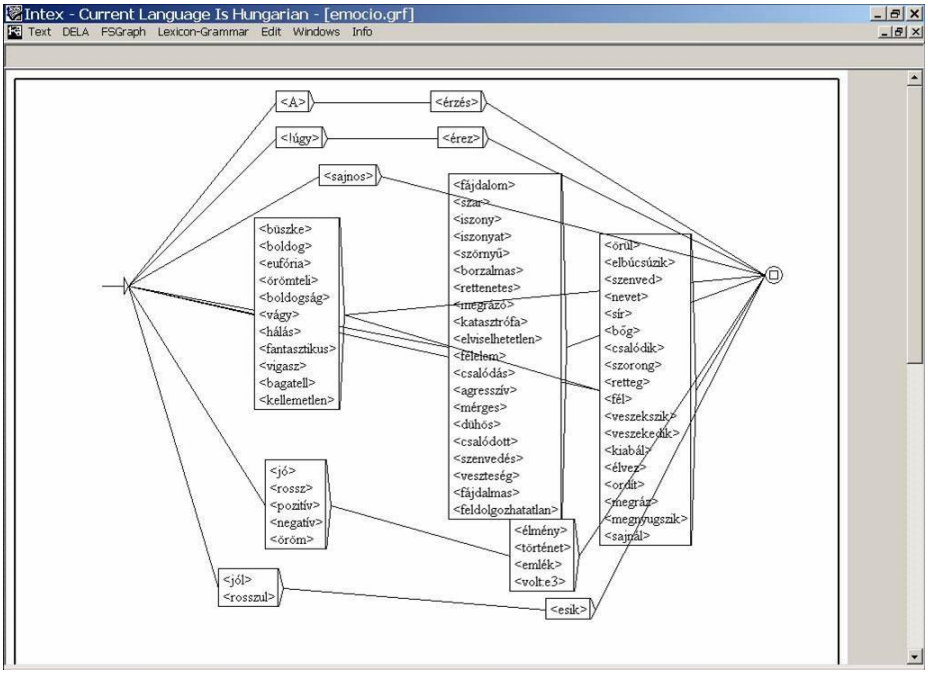
„És ez ilyen nagyon jó érzéssel töltött el, hogy bebizonyítottam a családomnak, hogy igen is, ha nagyon akarom, akkor meg tudom csinálni."

Ebbe a kategóriába soroltuk a történet egészére vonatkozó értékelő reflexiókat is, melyek érzelmi viszonyulást implikálnak.

„…és akkor még utána az orvos is tette a megjegyzést, és akkor ugye az volt a probléma igazából, nem maga, hát jó, maga a cselekedet, tehát maga az abortusz volt a baj, de az, hogy végül is nem sok időnk volt rá, eldönteni, és tulajdonképpen azt beszéltük, hogy ha nem most lett volna, hanem később, akkor akár igen. És ezért ez így rossz volt."

*„És én óvodás voltam és valami úton-módon nem szerettem az óvodában aludni. Állandóan fönt voltam, magyarul rossz voltam."

Az utolsó példa annak szükségességét mutatja, hogy a történet egészére vonatkozó értékelő reflexióknál a létigének csak E/3 alakját vehetjük figyelembe.



1. ábra. Érzelmi állapotokat azonosító gráf

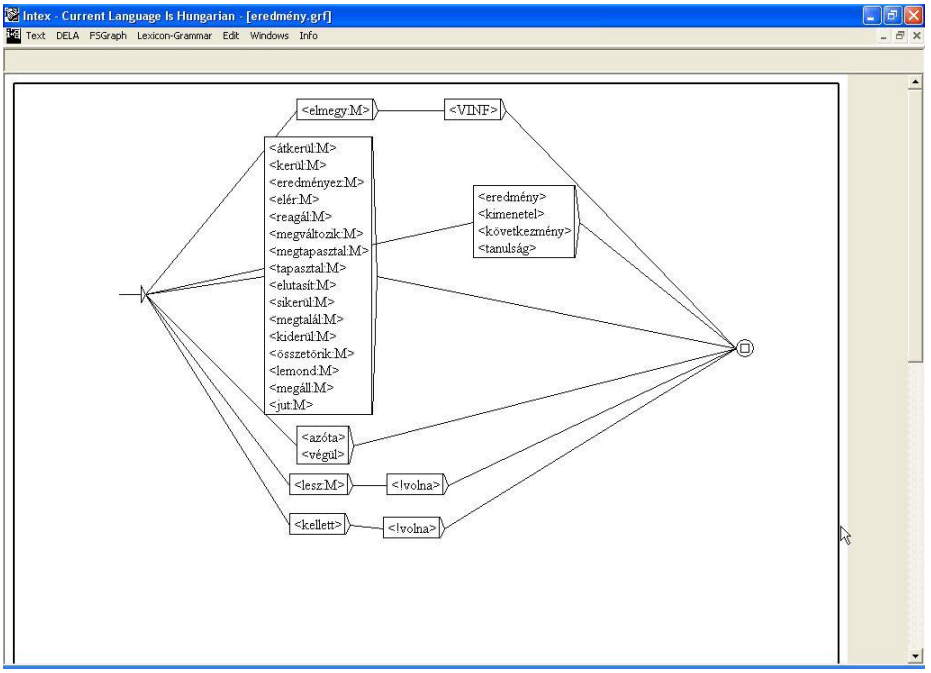
2.2 A fizikai okság azonosítása

Ennél az oksági kategóriánál szerencsésebbnek tartjuk a Schank és Abelson [8] által használt Eredmény okság elnevezést, mivel sokkal inkább utal az események közötti kontrafaktuális viszony természetére, mely nem csupán fizikai lehet.

Az Eredmény típusú oksági viszony azonosítását szolgáló gráf szavait a vizsgált korpuszon a kontrafaktuális teszt alapján végzett manuális elemzés szerint választottuk ki, és csupán azokat hagytuk benne a jelenlegi szólistában, melyek legalább 70 %-ban helyes találathoz vezettek. A keresett szavak túlnyomó részben igék, melyek múlt idejű, nem feltételes módú alakban fordultak elő a korpuszban (például *kiderült, sikerült*); emellett szemantikailag specifikus főnevek (például *eredmény, tanulság*), valamint néhány funkciószó (például *azóta, kellett*) jelölte az eredmény okságot.

„…nem vettek fel elsőre az orvosira, így elmentem dolgozni nővérként.”

„Talán az utóbbi időkből az egyik legrosszabb történet a fiammal kapcsolatban, amikor annak ellenére, hogy nem volt neki engedélyezve, az autót kivitték, elmentek furikázni, főtengelyesre hajtották, jó pár ezer forintba került, ez pár évvel ezelőtt fordult elő.”



2. ábra. Eredmény típusú oksági viszonyt azonosító gráf

Az 'elment'+főnévi igenév–szerkezet előfordulási jellegzetessége, hogy az eredmény típusú oksági viszony részeként egyes esetekben a kiinduló eseményt (2. példa), máskor a következményt (1. példa) jelzi. Az alábbi példa azt mutatja, hogy a rossz találatok strukturálisan a történet bevezető részében (szetting) fordultak elő.

*„Volt egy kirándulás, ami nagyon negatív eredménnyel zárult, ugyanis apáék elmentek vadászni, és a húgommal, ez télen volt, ez Szabolcs megyében történt, és ott nagyon ritka a domb, és volt egy nagy domb, és vittük a szánkót, és a húgommal szánkóztunk."

A 'kellett' segédige szövegbeli megjelenése –performativitása miatt– 85%-ban Eredmény típusú oksági viszonyt azonosított. A találatokat részletesen megvizsgálva azt találtuk, hogy a főnévi igeneves szerkezettel olyan eredményekről számolnak be a beszélők, amelyek külső hatásra következtek be.

„De például ez az ismerősöm kiállt mellettem abban, hogy ne higgyek másoknak, saját magam tapasztalataim alapján vonjak le következtetéseket és végülis így is tettem, és tényleg abszolút nem is kellett csalódnom."

Az idézett példa az ún. kettős kódolás kérdését is felveti, amennyiben az eredményként értékelhető reakció emocionális válaszként is felfogható.

A 'sikerül' múlt idejű alakja 82%-ban adott helyes találatot a vizsgált korpuszban az eredmény típusú okságra, téves azonosítás akkor fordult elő, amikor kognitív ige állt a közvetlenül előtte levő tagmondatban.

„Hát mondjuk a legutóbbi, az a Katinak az esküvõje, azt szervezgettem, és azok jól sikerültek."

*"És elkönyveltem magamban, hogy akkor ez most nem sikerült, aztán kiderült, hogy mégis."

Az *azóta* kötőszóval kapott találatok részletes elemzése szintén indokolt az eredmény típusú oksági viszonyok azonosításakor, hiszen funkciója szerint az egyik tagmondatban leírt, múltbeli eseményt köti össze egy másik tagmondattal, amelyben a jelenre gyakorolt hatásról van szó. Értelemszerűen oksági kapcsolat áll fenn a két esemény között, a múltbeli esemény az oka annak, ami a jelenben zajlik.

Korpuszunkban 31 találatot kaptunk. Az esetek túlnyomó részében (24) helyes, tehát eredmény típusú oksági viszonyt jelöl az azóta kötőszóval összekapcsolt két tagmondat. A jó találatokban az ige jelen idejű az eredmény oldalon, ami jól mutatja a jelenre gyakorolt hatást:

„...és megismerkedtem egy olyan fiúval, aki viszont árva volt, tehát õ is egyedül volt, és tizenkilenc éves koromban, és azóta tart a szerelem, az együttlét, mert õ a férjem,... "

A téves találatokban a kötőszót követő ige gyakrabban múlt idejű.

„És talán ilyen erõs félelmet azóta sem éreztem."

Közös jellegzetességük ezen felül, hogy nem két esemény között teremt kapcsolatot a kötőszó, hanem a múltbeli esemény jelentőségét hangsúlyozza a beszélő:

*„Szóval ekkor féltem életemben egyszer, elõször és talán azóta is utoljára. "

2.3 Kitekintés

Az Eredmény típusú okság azonosításában mindenképpen további cél egy olyan korpuszkiindulású igepár–lista felállítása, mely nem általános oksági szemantikájú, hanem tartalmas szavak mentén (pédául *lelőtte→meghalt*) eredményez találatokat. A pszichológiai irányú kérdésfeltevések szempontjából pedig a következő lépés a mentális folyamatra utaló 'kognitív' igék és kifejezések elhelyezése az események által kiváltott belső reakciók kategóriáján belül.

Bibliográfia

1. Ehmann, B., Kis, B., Naszódi, M., László, J.: A szubjektív időélmény tartalomelemzéses vizsgálata. Pszichológia, Vol. 25 (2), 2005, 133-142
2. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments and Computers (2004)
3. László, J., Ehmann, B., Péley, B., Pólya, T.: A narratív pszichológiai tartalomelemzés: elméleti alapvetés és elsõ eredmények, Pszichológia, Vol. 20, 2000, 367-390
4. Kiefer F. : Jelentéselmélet, Corvina, Budapest (2000)
5. Kuti, J., Vajda, P., Varasdi, K.: Javaslat a magyar igei WordNet kialakítására. III. Magyar Számítógépes Nyelvészeti Konferecia konferenciakötete, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2005)
6. Papp, O.: Narratív koherencia-elemző program helye a pszichológiai kutatásban. II. Magyar Számítógépes nyelvészeti konferencia, Konferenciakötet, Kiadó: Juhász Nyomda, Szeged (2004)
7. Papp, O., Mészáros, Á.: Oksági viszonyok azonosítása önéletrajzi narratívumokban. III. Magyar Számítógépes Nyelvészeti Konferecia konferenciakötete, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, 2005
8. Schank, R.C., Abelson, R.P.: Scripts, Plans, Goals and understanding: an Inquiry into human Knowledge Structures. Erlbaum, Hillsdale (1977)
9. Silberztein, M.:Intex Manual. Internetes elérhetőség: www.nyu.edu/pages/linguistics/**intex/**
10. Trabasso, T., van den Broek, P.: Causal Networks versus Goal Hierarchies in Summerizing Text. Discourse Processes, Vol. 9. Alex Publishing Corporation, New Jersey (1986) 1-15
11. Trabasso, T., van den Broek, P., Suh, S.Y.: Logical Necessity and Transitivity of Causal Relations in Stories. Discourse Processes, Vol. 12. Alex Publishing Corporation, New Jersey (1989) 1-25
12. Trabasso, T, Sperry, L. L..: Causal Relatedness and Importance of Story Events. Journal of Memory and Language, Vol. 24 (5), 1985, 595-611.
13. Trabasso, T., van den Broek, P.: Causal Thinking and the Representation of Narrative Events, Journal of Memory and Language, Vol. 24 (5), 1985, 612-630.

A személyközi közelítés-távolítás azonosítása lokális nyelvtanok segítségével[81]

Pohárnok Melinda

PTE BTK Pszichológiai Intézet, Pécs, Ifjúság útja 6.
pomelin@freemail.hu

Kivonat: A narratív pszichológiai tartalomelemzés az élettörténeti narratívumok inherens jellegzetességének tekintheti a szereplők egymás viszonyában való mozgását. (Pohárnok, 2004)[1] Feltételezzük, hogy e mozgás két alapvető irányultsága – közelítés és távolítás – jól megragadható nyelvi struktúrákkal rendelkezik, amelyek a számítógépes nyelvi feldolgozás eszközeivel elérhetőek. Ugyanakkor ezen nyelvi elemek megfeleltethetőek a kapcsolati szabályozás és a szelf-szabályozás pszichológiai változójának. Az előadás bemutatja a NOOJ nyelvészeti fejlesztőkörnyezetben az MTA Nyelvtudományi Intézet munkatársainak segítségével folytatott munka jelenlegi fázisát, amelyben a személyközi közelítés-távolítás azonosítására szolgáló szemantikai szótárakat és ezekre épülő lokális nyelvtanokat hozunk létre.

1. Bevezetés

Az elbeszélést a narratív pszichológia szemszögéből az egyéni pszichikum szempontjából alapvető információkat hordozó entitásként és működési- és élményszervezési-módként tekintjük. Ugyanakkor azt is feltételezzük, hogy az elbeszélésben jól megragadhatóak azok a szövegelemek, amelyek valamely pszichológiai jelenségre utalnak. Míg a korábbi kvalitatív vizsgálatokban az elbeszélés pszichológiai jelentései pusztán az intuitív megértés számára voltak hozzáférhetőek, mi azt tűztük ki célul, hogy a számítógépes morfológiai-szintaktikai szövegelemzés, és – feldolgozás eszközével mérhetővé tegyük ezeket a pszichológiai jelentéseket. Korábbi vizsgálatok (pl. Pohárnok és mtsai., 2005)[2] rámutattak, hogy a tárgykapcsolati elbeszélések – kapcsolati epizódok – alapvető strukturális jellegzetességének és a narratív pszichológiai tartalomelemzés módszerével megragadható dimenziójának tekinthetjük a kapcsolati tér változásait. Ugyanakkor vizsgálataink azt is megerősítették, hogy a személyközi közelítés-távolítás az elbeszélésekben összefüggésbe hozható az elbeszélő személy kapcsolati és érzelmi szabályozásának adekvátságával.

Munkánk alapvetése, hogy létezik egy olyan interperszonális vagy interaktív tér, amely mindig az én és a másik viszonya alapján szerveződik: a tér két végpontját az én és a másik adja meg, és egymás viszonyában való mozgásuk a kapcsolat alapvető sajátosságának tekinthető. Ez a tér, illetve az elbeszélő és a szereplők ebben való

[81] A tanulmány az NKFP 6/074/2005 számú pályázat támogatásával készült.

mozgása kiemelkedik az elsősorban kapcsolati témájú élettörténeti narratívumokban, és így vizsgálhatóvá válik a gépi szövegfeldolgozás eszközeivel. A személyközi közelítés-távolítás az elbeszélésekben egyrészt konkrét, fizikai térben való mozgásokként jelenhet meg. A felé (vele) – tőle (nélküle) irányairól van szó. Másrészt megjelenhet az elbeszélő és a számára fontos mások közti érzelmi viszonyulások „irányában", amely az én-elbeszélésekben az elbeszélő és a szereplők közti érzelmi viszony kifejezésében érhető tetten. A szeret - gyűlöl ellentétes „irányai" adják ki ezt a dimenziót.

A kutatás jelenlegi fázisában a PTE Pszichológia Intézet és az MTA Nyelvtudományi Intézet együttműködésében folyik. Ennek a munkának egy szeletét mutatja be az előadás.

2. A nyelvi korpusz

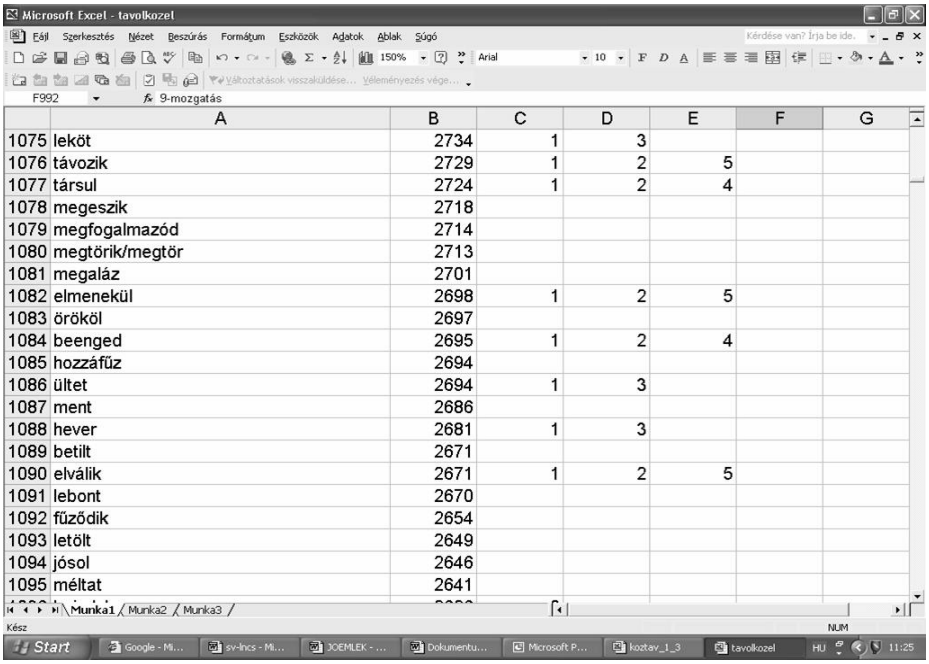
A személyközi közelítés-távolítás azonosítására szolgáló nyelvtanok felépítését két irányból kezdtük el. Egyrészt egy olyan szövegkorpusz felől, amelyben hangsúlyosan jelen vannak az én és a másik viszonyában zajló mozgások. Ebből a korpuszból még korábban kinyert minták alapján fogalmaztunk meg feltevéseket arra nézve, milyen szavak – szófajok -, szókapcsolatok felelhetnek meg a közelítésnek, illetve távolításnak. Másrészt a Magyar Nemzeti Szövegtár szófaji listáit használtuk fel, hogy azokat leszűkítve a közelítés-távolítás szempontjából releváns szótárakat építsünk.

Az a szövegkorpusz, amelyben korábban manuális kódolással azonosítottuk a személyközi közelítésre-távolításra utaló kifejezéseket, egy 120 671 szóból álló élettörténeti epizódokat tartalmazó, élőnyelvi szöveg („BPDinterjúk"). Mindennek a kutatás szempontjából több lényeges vonatkozása van. Egyrészt a vizsgálati személyektől olyan epizódokat kértünk, amelyek életük egy meghatározó személyével kapcsolatosak (pl. „Mondja el első párkapcsolatának történetét!"), ez lehetővé tette, hogy az elbeszélő én és a szereplő másik viszonyára bőséges utalást találjunk a szövegekben. Ugyanakkor megnehezíti a személyközi mozgások referencialitásának felismerését azzal, hogy a közelítés-távolítás forrása és célja legtöbbször rejtve marad a szövegben, hisz a vizsgálati személynek feltett kérdés utal rá. Így azt a döntést hoztuk, hogy nemcsak a referencia személy közvetlen megjelölését fogadjuk el (pl. „*az anyámat összecsókolgatom*"), hanem közelítésnek tekintjük az E/1 személyű mozgást, amelynek célja E/3 vagy T/3 ragozott névmás, vagy határozószó, vagy névutó (pl. „*éjjel bementem és befeküdtem mellé jük*"). Hasonlóan közelítésnek tekintjük az E/3 személyű mozgást, amelynek célja E/1 vagy T/1 ragozott névmás, vagy határozószó, vagy névutó (pl. „ *jött, meg szaladt hozzám*"). Végül a korpusz sajátosságához tartozik, hogy élőnyelvi szöveg révén nyelvtanilag rosszul formált, töredékes, így ez sokszor akadálya volt annak, hogy egyes nyelvtanok pontosan alkalmazhatóak legyenek.

3. A személyközi közelítés-távolítás azonosítására szolgáló szemantikai szótárak

A kidolgozandó lokális nyelvtanoknak kétféle irányú mozgást kell azonosítaniuk, és egyaránt számításba kell venniük a fizikai mozgást (jön-megy), és az érzelmi "mozgást"(szeret-gyűlöl). Mindehhez az eddigiek során háromféle szófaji szótárt dolgoztunk ki.

Az Ige-szótárhoz a Magyar Nemzeti Szövegtár (MNSz) 10000 szavas ige listáját használtuk fel. A listát négy független kódoló tekintette át, és a következő kategóriák mentén értékelte a szavakat:
1.humán/nem-humán mozgás
2.iránnyal bíró/iránnyal nem bíró
3.közelítő/távolító
4.aktív/passzív.



	A	B	C	D	E	F	G
1075	leköt	2734	1	3			
1076	távozik	2729	1	2	5		
1077	társul	2724	1	2	4		
1078	megeszik	2718					
1079	megfogalmazód	2714					
1080	megtörik/megtör	2713					
1081	megaláz	2701					
1082	elmenekül	2698	1	2	5		
1083	örököl	2697					
1084	beenged	2695	1	2	4		
1085	hozzáfűz	2694					
1086	ültet	2694	1	3			
1087	ment	2686					
1088	hever	2681	1	3			
1089	betilt	2671					
1090	elválik	2671	1	2	5		
1091	lebont	2670					
1092	fűződik	2654					
1093	letölt	2649					
1094	jósol	2646					
1095	méltat	2641					

1. ábra. Az igék szemantikai csoportokba sorolása

A kódolók közti egyeztetések után a nyelvtanokba beépíthető igék három csoportját tudtuk meghatározni:

1. Irány nélküli, vonzatfüggő humán cselekvés vagy állapot: *áll, vár, öltözik, marad, ül ,leszalad, elköltözik*
2. Iránnyal bíró közelítő/közeledő mozgás: *érkezik, kísér, támaszt, beenged, csókol*
3. Iránnyal bíró távolodó/távolító mozgás: *kizár, elhagy, eltűnik, kirak, elereszt*

A következőkben az MNSz határozószó (n=10000) és a névutó (n=88) szótárait tekintettük át, és kigyűjtöttük, illetve a közelítésnek és távolításnak megfelelő csoportokba soroltuk a szavakat. Így a következő csoportok adódtak a határozószók esetében.

1. Közelítő határozószók – téri: *hozzá, neki, közelben, négyszemközt*
2. Közelítő határozószók – érzelmi: *aggodalmasan, biztatóan, féltőn, megnyugtatóan*
3. Távolító határozószók – téri: *egyedül, külön, messzebbre, odébb, távol*
4. Távolító határozószók – érzelmi: *elutasítóan, gunyorosan, kiábrándítóan, megcsalatva.*

A névutók sok esetben átfedést mutattak a határozószókkal, így ott a szavaknak csak két kisebb csoportját tudtuk létrehozni.

1. Közelítő névutók: *alá, elé, elébe, felé, felől, köré, közé, között, mellé, mellett, mögé, mögött*
2. Távolító névutók: *aló, elől, innen, innét, kívül, közül, mellől, mögül, nélkül, túl*

4. A szótárokból felépített lokális nyelvtanok a NOOJ fejlesztőkörnyezetben

A kidolgozott szótárak segítségével az igecsoportok és a határozószók, illetve az igecsoportok és a névutók együtt járására terveztünk kidolgozni lokális nyelvtanokat. Az MTA Nyelvtudományi Intézete nyelvtechnológiai osztályának munkatársai által fejlesztett magyar nyelvű NOOJ fejlesztőkörnyezet morfológiai elemzője az igék és a névszók példányait meghatározott kimeneti jegyekkel látja el. A névszók esetén a ragnak megfelelő esetet, inflexiós toldalékokat – pl. többes szám, birtokos jelölve van - és képzőket használ. Az igék esetén többek között a szám, a személy, az igeidő és a mód morfológiai jegyek alapján meghatározott annotációja történik.

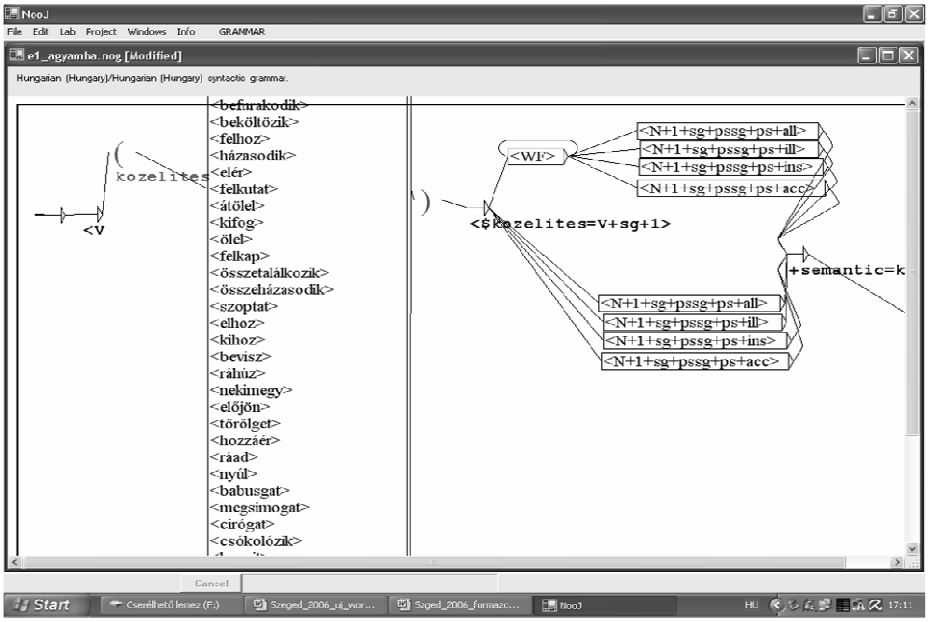
Mielőtt a szintaktikai nyelvtanok építésébe fogtunk, a Nyelvtudományi Intézet munkatársai rendelkezésünkre bocsátották az ún. „Jedlik-projekt" szövegkorpuszának annotált szótárát. Ez tartalmazta azt a korpuszt is, amelyen a korábbi manuális kódolással elvégeztük már a személyközi közelítésre – távolításra utaló szóalakok, szókapcsolatok azonosítását. Ugyanakkor néhány szóalakot ebből a korpuszból sem ismert fel a NooJ. Például a meghatározó személyközi viszonyok szempontjából lényegesek közül az „*anyámék*", „*anyukámék*", „*édesanyámék*", „*barátnőmhöz*" szóalakok voltak, amelyek annotációja a találati hibák elkerülése miatt nem történt meg. Módunk volt azonban ezeket a szóalakokat annotálni, majd beépíteni a korpusz szótárába.

Így végül lehetőségünk volt arra, hogy meghatározott számú, személyű igéket és meghatározott határozószókat, névutókat és névszók meghatározott esetű alakjait használjuk fel a nyelvtanok építéséhez.

4.1. Példák a személyközi közelítés azonosítására szolgáló lokális nyelvtanokból

Elsőként a közelítő igék azonosítására szolgáló egyszerű szintaktikai gráfot szerkesztettünk, azzal a céllal, hogy megtaláljuk azokat a közelítő igéket, amelyek szerepelnek a „BPDinterjúk"-ban. Így a kezdetben 596 szót tartalmazó teljes közelítő/közeledő-ige lista leszűkült 149 szóra, így könnyebben kezelhető vált a NooJ elemzője számára. Ezek után többek között a következő nyelvtanokba építettük be az igelistát:

1. Olyan szókapcsolat, ahol az E/1 vagy E/3 közelítő ige E/1 birtokos személyjellel ellátott, tárgyragos, vagy allativus, illativus, instrumentalis esetben lévő főnévvel áll együtt. Itt módosítani lehet, hogy a két szó közvetlenül egymás után, vagy bármely szóalakokat közrefogva van jelen a szövegben. Mivel a NooJ-ban az elemzés egységeként megadhatjuk a mondatot, ezért a mondat határai korlátozzák a keresett szóalakok közti szavak számát.



2. ábra. Ige-főnév közelítő gráf

Példák a gráf által azonosított szövegrészekből:

„és akkor *összetalálkoztam a barátommal* Miskolcon"
„oda *tartottam az arcomat"*
„soha nem *öleltem még meg anyámat*"
„hogy nem *fogom többet látni az apámat"*

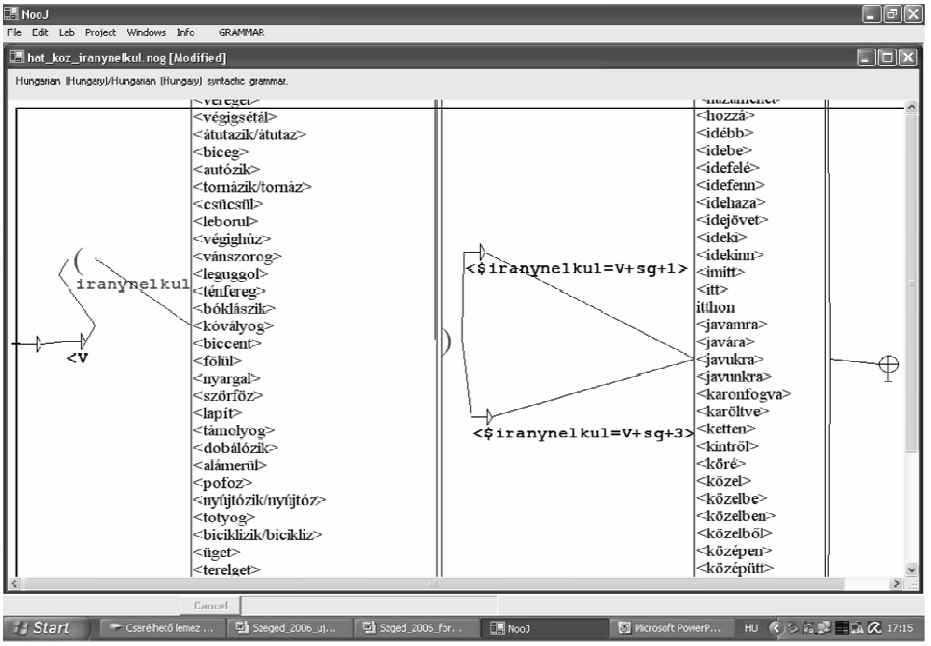
Amennyiben a gráffal E/3 közelítő igéket E/1 birtokos személyjellel ellátott meghatározott esetben ragozott névszókkal együtt keresünk, a következő példák adódnak:

„Aztán *beszaladt a hálószobámba* apám"
„*fogta a kezemet* , és akkor ez ilyen biztonságos érzés volt nekem"
„hogy rendőrség *hozott mindig vissza apukámhoz*"
„de addig fiú nem *fogta meg a kezemet"*

A megtalált kifejezések egy része megfelel a fizikai közeledés szemantikai kritériumának, ugyanakkor a példákból is kitűnik, hogy meg kellett oldanunk a tagadás kezelését. Erre egyelőre egy önálló negáció gráfot tudunk használni, amely megtalálja egy másik gráf által azonosított szövegrészek egyszerű tagadó formáit. A példák esetében kiemeli a „de addig fiú *nem fogta meg a kezemet*" és „hogy *nem fogom többet látni az apámat"* eseteket. Így ezek más szemantikai kategóriába kerülhetnek. A tagadás kezelése a személyközi közelítés-távolítás esetében azért fontos, mert a közelítés tagadása távolítást jelenthet - ahogy az előbbi példákból is kitűnhet.

A megtalált példák arra is felhívják a figyelmet, hogy a szavak azonos alakúság a beleszólhat az eredményekbe: a „fog"-ra nemcsak mint aktív igére, hanem a jövő időt kifejező segédigére is találatot kapunk.

2. Olyan szókapcsolat, ahol a közelítő, vagy irány nélküli humán cselekvést, állapotot kifejező E/1 és E/3 ige közelítő határozószóval áll együtt. A határozószó definíciójának bizonytalansága miatt ide kerültek azok az esetek is, amikor az ige után ragozott névmás áll (pl. *hozzám, vele*), mert ezeket határozószóként és névmásként is annotálták a szótárban.



3. ábra. Ige-határozószó közelítő gráf

Példák a gráf által azonosított szövegrészekből:

„Aztán meg este *bejött hozzám*"
„Tehát így *lefeküdtem mellé* a földre"
„és akkor így *vártam rá* , hogy elvisz"
„Aznap mikor *jöttem hazafelé*"

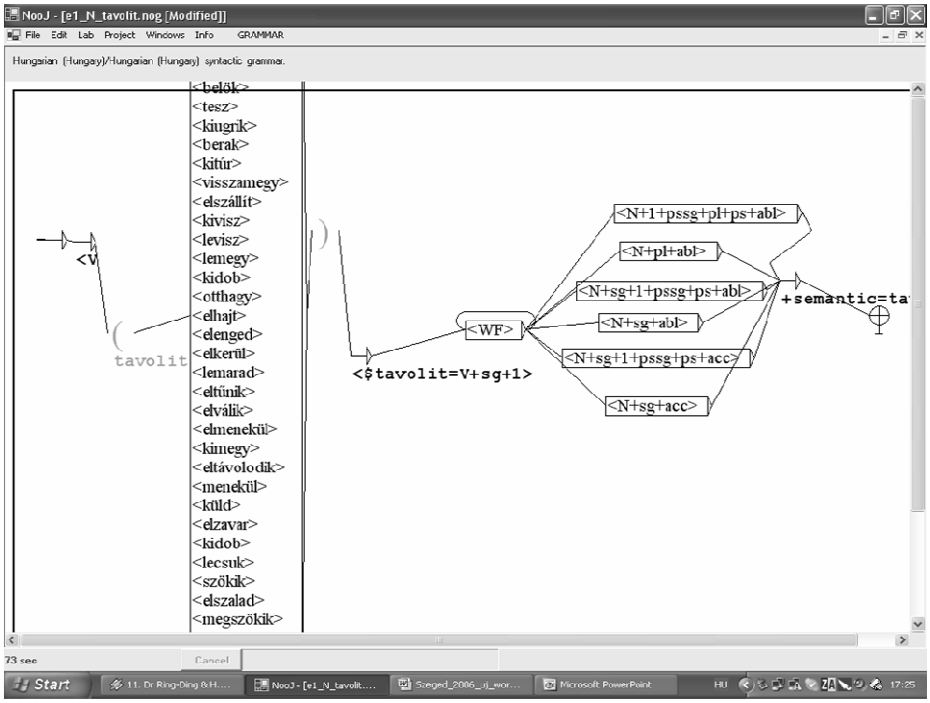
„Akkor *álltam neki* életemben először főzni"
„és nagyon sokáig mérges is *volt rám*"
„hogy ha az életben még *kapok rá* lehetőséget"
„Ő tudta, hogy *van neki* valakije"

Ahogy az idézett példák második fele is mutatja, ebbe a nyelvtanba további megszorításokat kell beépítenünk. Szükségesnek tűnik úgy megszűrni a határozószók, ragozott névmások szótárát, hogy az állandósult szókapcsolatokat és az elváló igekötőként is funkcionáló névmások egy részét (pl. *neki*áll, *rá*kap, *rá*jön) kiemeljük. Ugyanakkor további megszorításokat lehet elérni azzal, hogy ha kiemeljük a ragozott névmásokat. Ezeknek a számát, személyét később megadva ugyanis ki fogjuk tudni zárni azokat az eseteket, amelyekben azonos számú és személyű az ige és a cél is (pl. hogy ő bármit *adhat neki*), így nem az elbeszélő én és a másik viszonyára utalnak.

4.2. Példák a személyközi távolítás azonosítására alkalmas lokális nyelvtanokból

A távolítás azonosításához az előzőekhez hasonlóan a távolító illetve irány nélküli humán cselekvést vagy állapotot kifejező igéket, a távolító határozószókat és a ragozott névmások meghatározott csoportját használtuk fel. Így többek között a következő lokális nyelvtanok adódnak:

1. Olyan szókapcsolat, ahol az E/1 vagy E/3 távolítást kifejező ige E/1 birtokos személyjellel ellátott, ablativus, delativus vagy accusativus esetben lévő főnévvel áll együtt. Itt változtattuk a névszó és az ige egymásra következésének sorrendjét, és köztük lévő szóalakok előfordulási lehetőségét is beiktattuk.



4. ábra. Ige-névszó távolító gráf

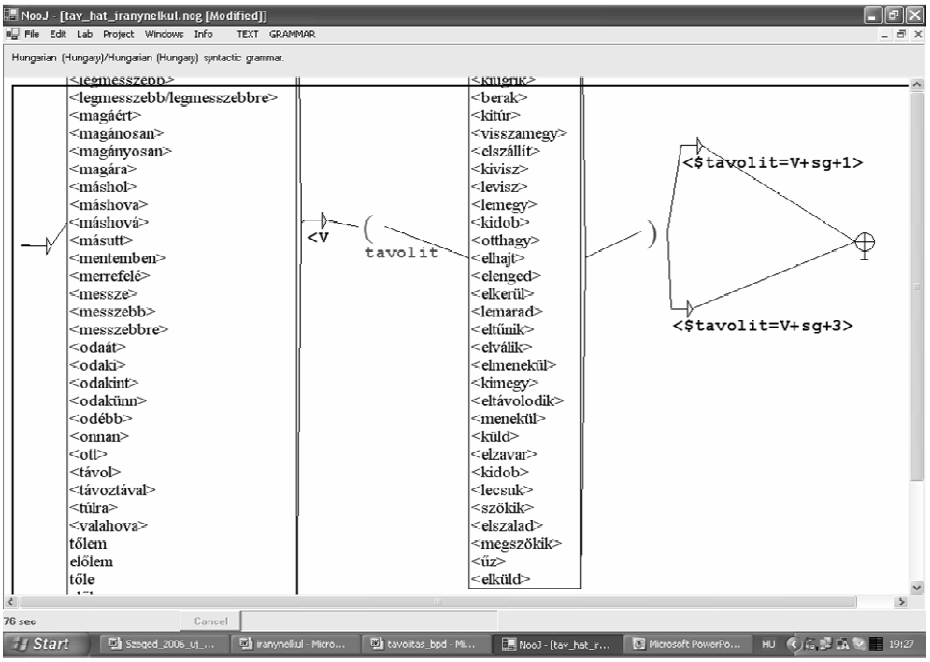
Az első lépésben itt is leszűkítettük az általános távolítás szótárt a korpuszban megtalálható szavakra az egyszerűbb kezelhetőség érdekében. Ennek következtében azonban olyan szemantikailag kevéssé differenciált igék maradtak csak benne a korpuszra illesztett szótárban, amelyek miatt sok érvénytelen találatot kaptunk a helyes találatok mellett.

„De akkor is én *hagytam ott a fiút*"
„Ezért *megyek egyik pasitól* a másik után"
„ha *elválok a férjemtől*"
„amikor már külön *mentem a férjemtől*"
„hogy a *mamámtól elkerültem*"

„nagyon sok *szenvedéstől menekültem* így meg."
„később francia *bokszot űztem*"
„hogy mindennap *elmentem venni egy csomó édességet*"
„de égve *hagytam a villanyt*"

A hibás találatok kizárása azzal oldható meg, hogy a nyelvtan névszói részét tovább szűkítjük azokra a főnevekre, amelyek jelentős másokra – szülők, családtagok, szerelmi partner, orvosok - vonatkoznak. Ezt a NooJ-ban két módon tehetjük meg. Egyrészt a korpuszban található szóalakokat egyenként visszük be a gráfba, vagy létrehozzuk a „jelentős másik" szemantikai kategóriát úgy, hogy ennek megfelelő annotációval látjuk el a korpusz szótárában az ehhez a kategóriához tartozó szóalakokat. A kutatás következő fázisában tervezünk létrehozni a fenti névszói kategóriákra alkalmazható szótárat, amely alapján új szemantikai annotációkkal és új szóalakokkal bővíthetjük a jelenleg használt NooJ szótárat.

2. Olyan szókapcsolat, ahol a távolító, vagy irány nélküli humán cselekvést, állapotot kifejező E/1 és E/3 ige távolító határozószóval áll együtt. A határozószó definíciójának bizonytalansága miatt ide kerültek azok az esetek is, amikor az ige után ragozott névmás áll (pl. *tőle, magától, előlem).* Az ige és a névszó sorrendjének itt is kétféle változatát használtuk. Az erre a szókapcsolatra kidolgozott gráfok egyik változatát mutatja az ábra.



5. ábra. Határozószó-ige távolító gráf

Példák a gráf által azonosított szövegrészekből:
„úgy éreztem biztos, hogy *magamra maradtam*”
„És akkor *megszöktem onnan*”
„hogy éjjel is *lökött magától*”
„a lányom, az már talán *eltávolodott tőlem*”
„Anyukám mindig sietett, *futott valahova*”

„hogy terhes *vagyok tőle*”
„Mert *volt külön* szobám”
„Hát sok jó élményem nem *maradt belőle*”
„és akkor utána *megy magától*”
„pár hónap alatt *lefut belőlem*”

A találatok arra hívják fel a figyelmet, hogy az iránynélküli humán cselekvésekhez sorolt létigék alkalmazása miatt adódik a legtöbb hibázás, így érdemes lenne a jövőben a létigékre és a hozzájuk kapcsolódó mind közelítést, mind távolítást kifejező szóalakokra külön algoritmusokat építeni. Ugyanakkor a távolítás kapcsán – ahogy korábban már említettük – különösen fontos lesz a tagadás rugalmas használata. Ennek kidolgozása is a további munka része.
A továbbiakban ezeknek a nyelvtanoknak a finomítását tervezzük, többek között a tagadás beépítésével, a sok téves riasztás adó igealakok –például a létigék – alkalmas előfordulásainak kiszűrésével és a már említett főnévi szemantikai kategóriák felhasználásával. Végül, mivel a jelenlegi nyelvtanok a fizikai mozgások azo-

nosítására alkalmasak, dolgozunk az érzelmi közelítésre-távolításra utaló szavak szótárának felépítésén, és szintaktikai nyelvtanokba való beépítésén.

Bibliográfia

1. Pohárnok M.: A térben való mozgás narratív dimenziójának vizsgálata borderline és depressziós betegek élettörténeti epizódjaiban. IN: Erős F. (szerk.): Az elbeszélés az élmények kulturális és klinikai elemzésében. Akadémiai Kiadó, Bp., (2004) 153-167
2. Pohárnok M., Nagy L., Bóna A., Naszódi M., Kis B., László J.: A kapcsolati mozgások számítógépes nyelvészeti vizsgálata élettörténeti narratívumokban. A LAS-vertikum közelítés-távolítás modulja. Pszichológia, (2005) 2:157-171.

A pszichológiai perspektíva modul fejlesztése[82]

Pólya Tibor

MTA Pszichológiai Kutatóintézet, Pf.: 398
1394 Budapest, Magyarország
polya@mtapi.hu

Kivonat: Az előadás a narratív perspektíva pszichológiai összetevőjének azonosítását végző modul fejlesztésének kezdeti fázisát mutatja be. A modul célja, hogy azonosítsa azokat a szövegrészeket, amelyek a szereplő pszichológiai perspektíváját érvényesítik, azaz a szereplő belső tudattartalmait mutatják be A modul fejlesztéséhez egyetlen 20 194 szavas élettörténeti narratívumot használtam. Az ebben a szövegben előforduló szóalakokból kialakítottam a pszichológiai perspektíva szótárat, amely mentális igéket, érzelem szavakat és szubjektív elemeket foglalt magában. A pszichológiai perspektíva modul képes arra, hogy az elemzésre kiválasztott szövegben téves találatokat szűrjön ki, illetve azonosítson olyan tagmondatokat, amelyek kifejtetten írják le a szereplő belső tudattartalmait.

1 A pszichológiai perspektíva fogalma

A narratív perspektíva fogalma két összetevőt foglal magában: a pszichológiai vagy fogalmi komponens, amely a szereplők belső tudattartalmainak bemutatását jelenti és a tér-időí komponens, amely az elbeszélő és a narratív elemek tér-időí lokalizációját jelenti [1]. Az itt bemutatásra kerülő munka célja olyan modul kidolgozása, amely alkalmas a szereplő tudattartalmait bemutató szövegrészek automatikus azonosítására magyar nyelvű narratívumokban.

2 A pszichológiai perspektíva pszichológiai relevanciája

A pszichológiai perspektíva pszichológiai relevanciájának azonosítását az alapvető attribúciós hiba fogalmát [2] felhasználva tervezzük elvégezni. A fogalom azt jelenti, hogy mások viselkedésének értelmezésekor alapvetően a személy diszpozícionális jellemzőire támaszkodunk. Ez a torzítás alapvetően a viselkedést megfigyelő személyekre vonatkozik, a cselekvő személyek ugyanis rendszerint szituációs tényezőkkel magyarázzák viselkedésüket. A pszichológiai perspektíva két alapvető formája – az elbeszélő és a szereplő perspektívája – között feltételezhetően ugyanazok a különbsé-

[82] A kutatást az NKFP 6/074/2005 pályázata támogatta.

gek jelentkeznek, mint a cselekvést megfigyelő és végrehajtó személy között. Azaz, ha a történet szereplőjének cselekvését a szereplő perspektívájából ismerjük meg, akkor viselkedését inkább szituációs tényezőkkel magyarázzuk, míg ha ugyanezen cselekvést az elbeszélő perspektívájából ismerjük meg, akkor diszpozícionális jellemzőkkel magyarázzuk azt. A szereplő és az elbeszélő perspektívájának különbségét két szempont mentén kívánjuk megragadni. Az első a felelősség fogalma. Azt feltételezzük, hogy ugyanazon cselekvés értelmezésekor nagyobb felelősséget tulajdonítunk a szereplőnek, amikor cselekvését az ő nézőpontjából ismerjük meg, szemben azzal, amikor az elbeszélő nézőpontjából ismerjük meg. A második szempont az empátia fogalmához kapcsolódik. Ezzel kapcsolatban azt feltételezzük, hogy a történet olvasói nagyobb empátiát mutatnak a szereplővel kapcsolatban, abban az esetben, amikor a cselekvését a szereplő perspektívájából ismerik meg, szemben azzal, amikor az elbeszélő perspektívája érvényesül a történetben. A feltételezés alapja mindkét esetben az, hogy a szereplő perspektívájának érvényesülésekor úgy látjuk, hogy a szereplő viselkedése szorosabban illeszkedik az adott szituáció jellegzetességeihez.

3 A pszichológiai perspektíva modul fejlesztése

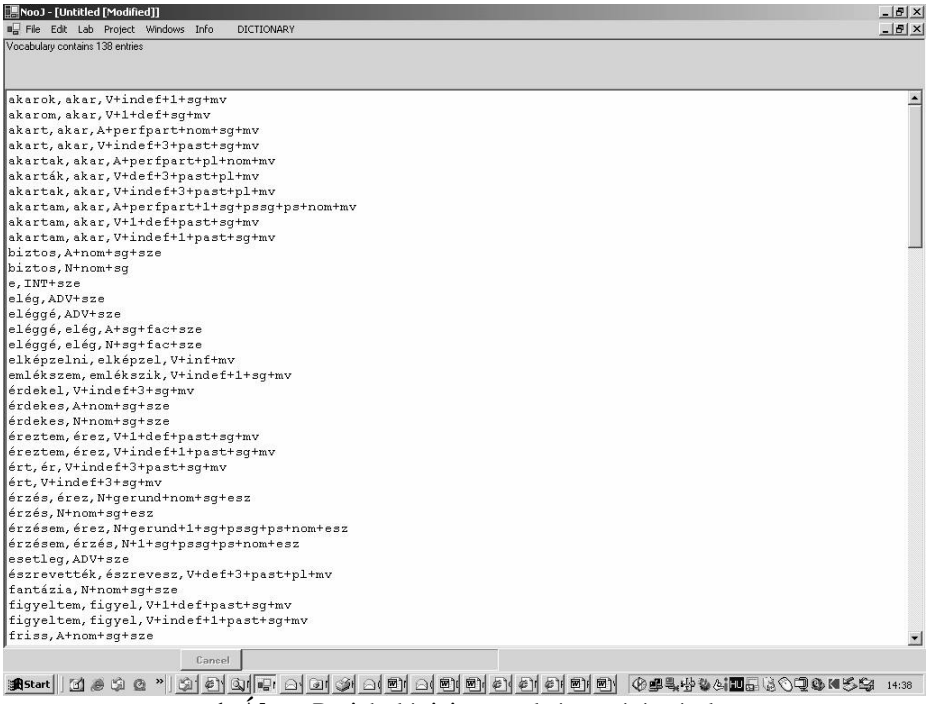
A perspektíva pszichológiai összetevőjének automatikus kódolására elsőként Wiebe [3] fejlesztett ki algoritmust, amely angol nyelvű történetek elemezését végzi. Ez az algoritmus két fő lépést foglal magába: először kiválasztja a szereplő perspektíváját érvényesítő szubjektív tagmondatokat, majd pedig meghatározza azt, hogy kinek a perspektíváját érvényesíti a szubjektív tagmondat. A pszichológiai perspektíva modul is ezt a két lépést foglalja magában. Az előadásban azonban csak az első lépésről lesz szó. A személyek azonosítását végző algoritmusok kidolgozására később kerül sor.

Az elemzéshez a NooJ számítógépes nyelvi elemző rendszert használtam [4]

3.1. A pszichológiai perspektíva nyelvi jegyei

Wiebe a szubjektív tagmondatok nyelvi jegyeinek három csoportját írja le.

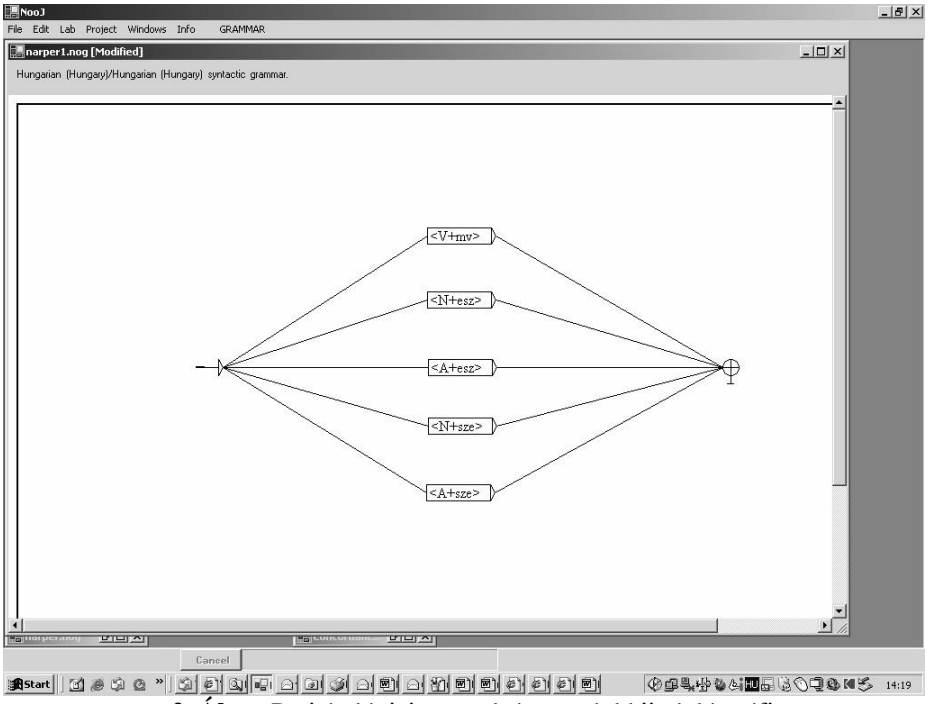
1. Mentális igék, amelyek a következő vagy az előző tagmondatban valamely szereplő tudattartalmait ágyazzák be. Például *Péter gondolta, hogy nem lesz ez így jó.* Vagy *Nem lesz ez így jó, gondolta Péter.*
2. Érzelmi szavak, amelyek a szereplők belső állapotára utaló igéket (például *sír*), főneveket (például *fájdalom*) és mellékneveket (például *szomorú*) foglalják magukba.
3. Szubjektív elemek, amelyek jelzik, hogy a leírás a szereplő belső állapotához igazodik. Ezen belül
   - Értékelést vagy ítéletet kifejező elemek (melléknevek: például *rossz*, határozószók: például *váratlanul*, kötelezettség kifejezései: például *kell*)
   - Tudás hiányának kifejezései (például *akárki* konkrét személyre vonatkozóan)
   - Bizonyosság kifejezései (például *nyilvánvalóan*)
   - Feltételes tagmondatok (például *mint* szóalak hasonlításhoz)



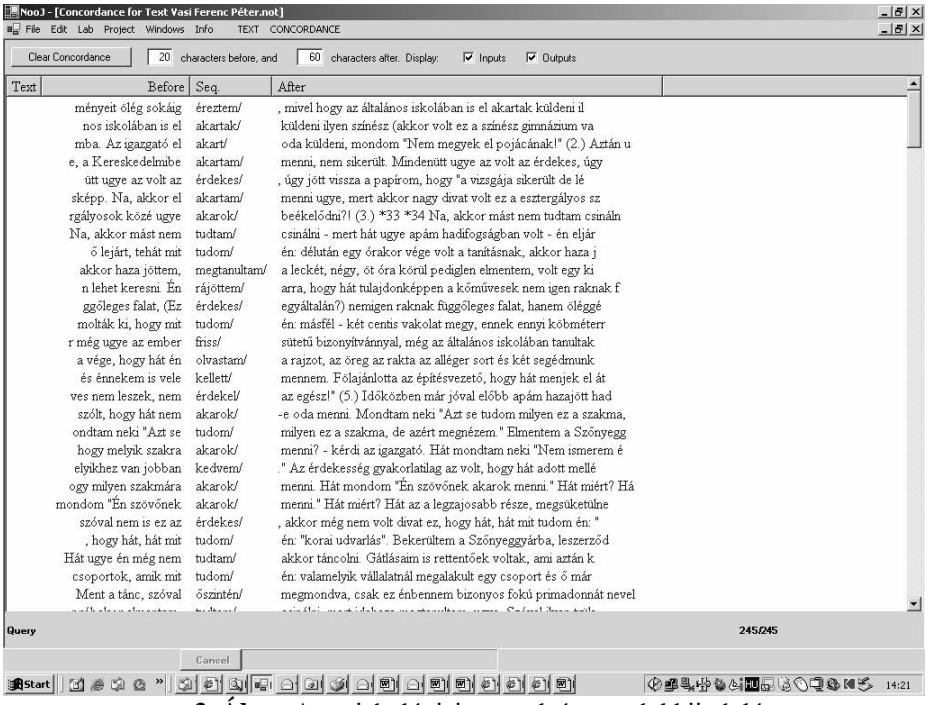
1. Ábra. Pszichológiai perspektíva szótár részlete

3.2. Pszichológiai perspektíva szótár

A modul fejlesztés kiindulásaként a pszichológiai perspektíva szótár kidolgozásába kezdtem bele. Első lépésként kiválasztottam egy 20 194 szavas egyes szám első személyű elbeszélőtől származó narratívumot [5] és az ebben a szövegben felismert szóalakok közül kiválasztottam azokat, amelyek besorolhatók voltak a mentális igék (mv), érzelmi szavak (esz) és szubjektív elemek (sze) kategóriába. Összesen 534 ilyen szóalak volt. A szóalakok kategóriába tartozását az annotációjukhoz hozzáadott szemantikus vonásokkal végeztem el (lásd 1. Ábra). A narratív perspektíva szótárba tartozó szóalakok azonosítására gráfot készítettem, ami a pszichológiai perspektíva modul fejlesztésének kiindulópontját adja (lásd 2. Ábra).



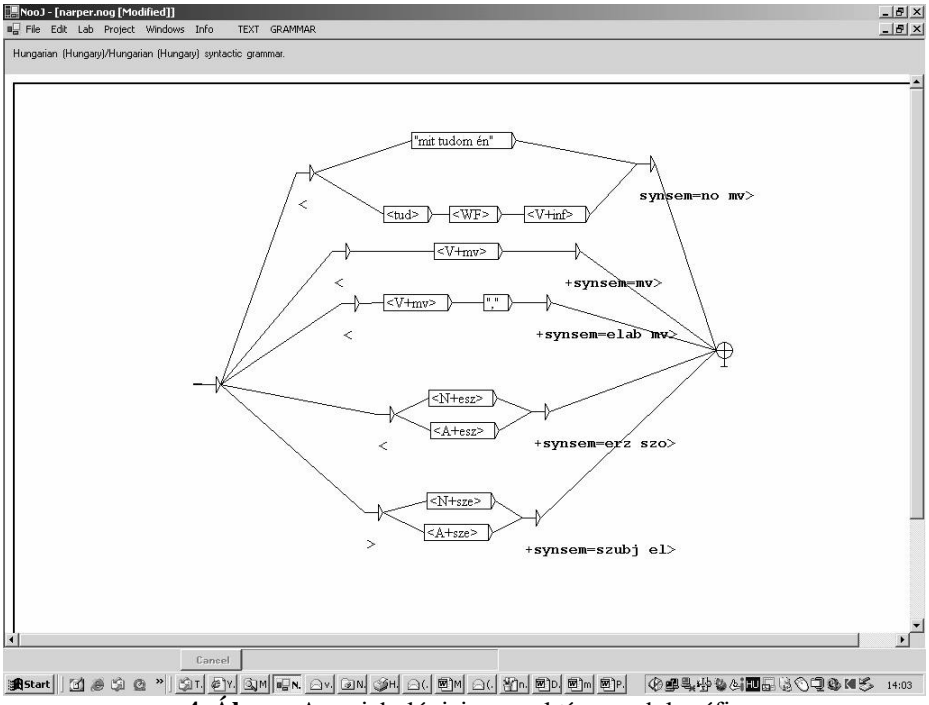
2. Ábra. Pszichológiai perspektíva modul kiinduló gráfja



3. Ábra. A pszichológiai perspektíva modul kiinduló gráfjának futtatásával kapott találatok

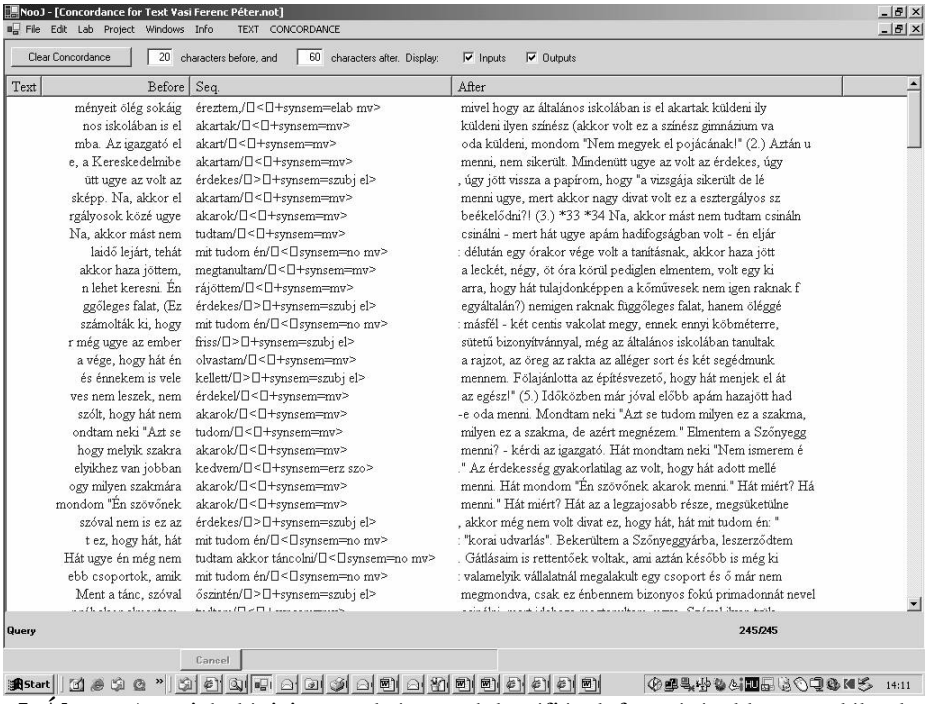
3.3. Pszichológiai perspektíva modul fejlesztése

A modul fejlesztése során két feladatot foglalt magában. Egyrészt pontosítani kellett a gráf futtatásával kapott találatokat. Az elemzett szövegben például gyakran előfordul a „mit tudom én” szófordulat, amely nem a szereplő belső tudattartalmaira vonatkozik, hanem az elbeszélő személy aktuális bizonytalanságát fejezi ki, ezért ennek előfordulását kiszűrtem a modul találatai közül (lásd 4. Ábra). Ugyancsak hibás találatokat adtak a mentális igeként kategorizált tud ige azon előfordulásai, amelyek azt fejezték ki, hogy valaki képes valamely cselekvés elvégzésére, például „tudtam csinálni”, így a tud ige ilyen használatait is kiszűrtem.



4. Ábra. A pszichológiai perspektíva modul gráfja

A mentális ige előfordulása önmagában képes arra, hogy utaljon a szereplő mentális állapotára, gyakran azonban a mentális állapot tartalmának kifejtett bemutatása is megjelenik a történetben. A modul fejlesztés során megoldandó másik fontos feladat a szereplők tudattartalmait kifejtetten bemutató szövegrészek azonosítása volt. Mivel a kifejtett tudattartalmak rendszerint önálló tagmondatban jelennek meg első közelítésként ezeket az eseteket úgy próbáltam azonosítani, hogy külön kiválasztottam azokat a mentális igéket, amelyek után közvetlenül vessző következik.



5. Ábra. A pszichológiai perspektíva modul gráfjának futtatásával kapott találatok

A pszichológiai perspektíva modul jelen állapotában öt csoport egyikébe sorolja be a pszichológiai perspektíva szótárban szereplő szóalakokat. A mentális igéket a szereplő mentális állapotára utaló mentális igék (synsem=mv), a szereplő mentális állapotára kifejtetten utaló mentális igék (synsem=elab mv) és a szereplő mentális állapotára nem utaló mentális igék alcsoportjába sorolja be (synsem=no mv). A negyedik és ötödik csoportot az érzelmi szavak (synsem=erzszo) és szubjektív elemek (synsem=szubj el) kategóriái adják.

3.4. Pszichológiai perspektíva modul fejlesztésének további lépései

Az előadás a pszichológiai perspektíva modul fejlesztésének első lépéseit mutatja be. A modul további kidolgozása során legalább négy területen kell feladatokat megoldani.

Elsőként a pszichológiai perspektíva modul által azonosított téves találatok kiszűrése a feladat. Az eddigi munka során csak a mentális igék helyes azonosítására alkalmaztam néhány szűrőt. Amellett, hogy a téves találatok kiválogatására további szűrőket kell alkalmazni a mentális igék esetében, az érzelmi szavak és a szubjektív elemek esetében is szükség van szűrésre.

A fejlesztés következő feladata a pszichológiai perspektíva szótár kiterjesztése. Az előadásban bemutatott szótár egyetlen narratív szövegben előforduló szóalakok elemzésére korlátozódott. Szükség van egy jelentősen nagyobb narratív perspektíva szótárra, amely újonnan elemzésre kerülő történeteken futtatva is felismeri a mentális igék, érzelmi szavak és szubjektív elemek nagy részét. Ezen a ponton hasznos lehet a

pszichológiai perspektíva szótár összekapcsolása a mentális igék szótárával, amelyet Vincze Orsolya dolgoz ki és az érzelmi szavak szótárával, amelyet Fülöp Éva fejleszt.

A fejlesztés során új feladatot jelent annak a személyeknek/szereplőnek az azonosítása, akinek pszichológiai perspektívája érvényesül a történetben. A szereplő sikeres azonosítása elsősorban az egyes szám harmadik személyű elbeszélőtől származó történetekben segíti a szereplő perspektíva érvényesülésének azonosítását. Ezekben a történetekben a szereplő perspektívájának érvényesítése szükségszerűen együtt jár azzal, hogy az elbeszélőtől különböző személy perspektívája érvényesül. Ez olyan téves találatok kiszűrését teszik lehetővé, amelyek például a „gondolom" szóalak esetében fordulhatnak elő. Ez ugyanis egyaránt leírhatja a szereplő és az elbeszélő mentális aktusát is, de csak az első esetben jelzi a szereplő perspektívájának érvényesítését. Ha az elbeszélő gondol valamit, akkor a megfigyelő perspektíva érvényesül, azaz a „gondolom" szóalak ilyen előfordulását ki kell szűrni a találatok közül. Ez a feladat válik megvalósíthatóvá a személyek azonosítása révén, mivel a szereplő perspektívájának érvényesülésekor az elbeszélő és szereplő személyében is különbözik, megfigyelő perspektíva esetében pedig azonos.

Végül a fejlesztés utolsó területét a perspektíva két komponensének összekapcsolása adja. A perspektíva tér-idői komponensének automatikus elemzésére kifejlesztett modul [6] összeépíthető a pszichológiai komponenst elemző modullal.

Bibliográfia

1. Pólya T., Kis B., Naszódi M., & László J. (2005). Az érzelmi tapasztalat minősége az élettörténeti elbeszélésben. A LAS-verticum perspektíva modulja. Pszichológia. 25(2), 143-155.
2. Ross L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L Berkowitz (Ed.), Advances in experimental social psychology. Vol. 10., pp. 174-221. New York: Academic Press.
3. Wiebe, J.M. (1991). Tracking point of view in narrative. Computational Linguistics. 20(2), 233-287
4. Silberztein, M. (2005). NooJ manual. Université de Franche-Comté.
5. <http://www.1956.hu/konyv/konyv201.html>
6. Pólya T. (2004). Élettörténeti narratív perspektíva és érzelemszabályozás. *II. Magyar Számítógépes Nyelvészeti Konferencia*. (278-281. o.), december, 9-10. Szeged.

Az aktivitás-passzivitás modul kidolgozása 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, Budapest
1132 Budapest, Victor Hugo u. 18-22.
laszlo@mtapi.hu

NKFP 6/074/2005 számú pályázat támogatásával készült

Kivonat: A narratívumok megszerkesztésének jellegzetességei a történetmondóra vonatkoztatható pszichológiai jelentéssel bírnak. Nem mindegy, hogy a világ történéseiről úgy számolunk be, mint amik aktív közreműködésünkkel történnek, amiknek alakulására cselekedeteinkkel hatással vagyunk – vagy mint amik passzív jelenlétünk mellett zajlanak. A pszichológiai relevanciával rendelkező tartalmak kigyűjtéséhez a tudományos narratív pszichológia eddig is használt tartalomelemző programokat. Jelen esetben kutatócsoportunk a morfoszintaktikai elemzésére is képes NooJ tartalomelemző program segítségével dolgozik, amely – az előbbiekkel ellentétben – arra is lehetőséget kínál, hogy ún. gráfokkal lokális nyelvtanokat létrehozva komplexebb, pontosabb elemzést végezzen.

1 Bevezetés

Az identitás bizonyos elméletekben úgy jelenik meg, mint társas közegben csiszolódó, folyamatosan újra- és újraszerkesztett élettörténet (Ricouer [6]; McAdams [3]; Gergen és Gergen [1]), melyből következtethetünk a történetmondó lelki folyamataira, viselkedésére.

Feltételezhetjük tehát, hogy a történetmondó aktuális nyelvhasználatának pszichológiailag releváns egységeit (szavak, kifejezések, nyelvtani formák stb.) tartalomelemző programok segítségével kvantitatív formába önthetjük. Számos empirikus kutatás irányult a szöveg lélektani vizsgálatára [2], s mivel az aktivitás-passzivitás szempontja is sok pszichológiai jelenségben játszik fontos szerepet, tartalomelemzéses mérése e jelenségek újszerű vizsgálatára adna lehetőséget.

A fejlődéslélektan felhívta a figyelmet arra, hogy már a csecsemők is aktív résztvevői egy kapcsolatnak, napjaink pszichológiájában teljesen elfogadott, hogy a gyerekek is aktívan közreműködnek személyiségük kibontakoztatásában. Erikson pszichoszociális fejlődéselméletében különböző fejlődési krízisek megoldásaként

jelenik meg az autonómia iránti igény, a kezdeményezés, illetve a teljesítmény igénye. A serdülőkori önértelmezés során a személyiség integrálja a korábbi fejlődési szakaszok énfogalmait, tovbbá összhangba hozza azt a társas közeg elvárásaival, a felé irányuló ítéletekkel. Az identitás tehát az interperszonális közeggel kölcsönhatásban alakul és formálódik [7].

A külső- és belső vezéreltség elmélete a személyiséget a kívülre és a belülre helyezett kontroll dimenziója mentén helyezi el. Egy belső kontrollos személyiség a történteket saját tetteinek következményeként éli meg, úgy érzi, hatással lehet élete alakulására; a külső kontrollos ember viszont a környezet történéseit nem magának, hanem a véletlennek, a sorsnak, vagy másoknak tulajdonítja, azaz mint hatókörén kívül eső dolgokat írja le őket [4].

2 Az igeszótár összeállítása

2.1 Igekategóriák

Jelen kutatás egy 'aktivitás – passzivitás' szótár kidolgozását mutatja be, mely munkát az igéknél kezdtük. Ennek során elsőként az igék csoportosítására alkalmas, pontosan körülhatárolt kategóriákat kellett létrehozni. Az ige által cselekvést, történést és állapotot fejezhetünk ki. A történést és állapotot kifejező igék általában passzívnak, a cselekvést kifejezők pedig általában aktívnak tekinthetőek – de ennél árnyaltabb megfogalmazásokra van szükség.

Aktivitásnak azt tekintettük, ha az ágens fesztelen tudattal rendelkezik; cselekvőképes és saját akaratából cselekszik, annak is tulajdonítva a történéseket (azaz belső kontrollos); cselekedeteivel hatással van környezete történéseire.

A passzív igéknek két csoportját különítettük el: Elsőbe tartoznak az állapotváltozás, történés igéi. Ide sorolhatók azon történések, amelyek a személyen kívül álló okokból – mint fizikai körülmények, transzcendens – következnek be, illetve változnak meg (pl.: adódik, adatik; folytatódik; butul, létesül)

Másik kategóriája a passzív igéknek az állapotot, folyamatosságot kifejező szavak. Ennek alkategóriái a létezést (van, létezik), állapotot (nő, lóg), birtoklást kifejező (birtokol, van valamije), illetve viszonyító, értékelő igék (hasonlít, ellentétben áll).

Mivel a fenti felosztás bizonyos igéket nem tud kezelni, további két csoportra volt szükség: Külön csoportba kerültek az „aktivációs kontúr" (Stern kifejezése) változásairól szóló igék, azaz olyan kifejezések, melyek az aktivitás növeléséről (pl.: fokoz) vagy csökkenéséről (pl.: lassít, leheveredik). De ide kerültek azon igék is, melyek saját tudatos elhatározással és aktív hozzájárulással egy állapotot fenntart (vár, fekszik, időzik). Továbbá külön csoportba kerültek a „STOP-GO" igék, melyek a cselekvés elkezdését, befejezését jelölik (pl.: abbahagy).

Az igék egy bizonyos részét kizártuk a csoportosításból: pl. a mentális (hisz, gondol, remél) és az érzelmeket kifejező (szeret, kedvel) igék többségét, melyek ugyan aktív belső tevékenységet jelölnek, viszont nem a külvilág megváltoztatására irányulnak. /A beszédaktusok [5] (pl.: ígér) kivételével, melyek az aktív alszótár részei lettek./ Nem kerültek csoportosításra továbbá a segédigék, illetve az olyan igék, melyeket általában a környezetünk eseményeinek leírására alkalmazunk (pl. hangutánzás igéi), vagy természeti állapotokra és azok változásaira vonatkoznak (pl. esteledik).

Végül kimaradtak a „fiziológiai" igék, amik a szervezet folyamatait jelenítik meg (pl.: remeg, kiizzad).

Az így összeállított alszótárakat végül két független bíráló is megvizsgálta, majd a NooJ program segítségével a találatokat kontextusban is ellenőriztük.

2.2 A megoldandó problémák

A fent megnevezett alszótárakban vannak átfedések, hiszen bizonyos igék két, akár három kategóriát is kaphatnak jelentésváltozataiktól, ragozásuktól, vonzataiktól, és a homonímia megjelenésétől függően.

Lehetnek egymást megengedő illetve kizáró átfedések is. Az aktivációs kontúr illetve a STOP/GO kategóriák kiegészítésként és nem kizárásként működnek. (Pl. az 'abbamarad' ige egyszerre kapott STOP/GO, illetve passzív -állapotváltozás, történés- kódot). Az aktivitás és a két passzivitás kód (főként az állapotváltozás, történés esetében) természetesen kizárja egymást, így ezek elkülönítésére szintaktikai gráfok készülnek.

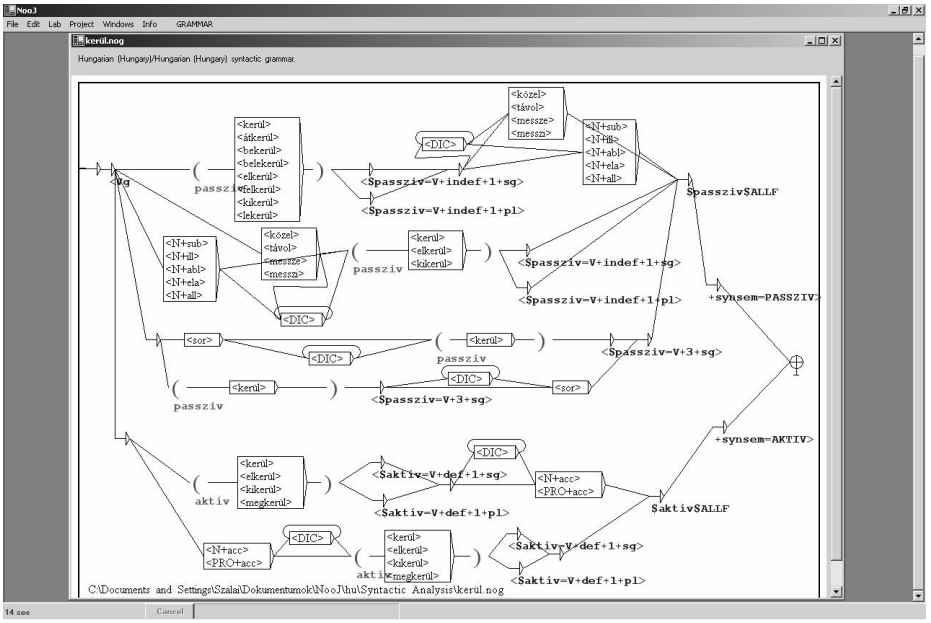
Kiindulópontunk szerint az aktivitás-passzivitás dimenzióját az elbeszélő, történetmondó szemszögéből vizsgáljuk, azaz a rá vonatkozó igéket vizsgáljuk meg. Ezek kiválasztásában segítségünkre vannak pl. a személyragok, a vonzatok, a határozott és határozatlan igeragozás különbségei, illetve az igekötők megjelenése.

A következőkben bemutatásra kerülő szintaktikai gráfok az alábbi nyelvtani jelenségeket próbálják kezelni:

- a vonzatkeret, az igekötők okozta jelentéskülönbségek;
- igék (jelen helyzetben főként kapcsolatok kezelésére utaló szavak), melyeknek kategóriája attól függ, hogy az elbeszélő alanya vagy tárgya/elszenvedője a cselekvésnek (az ágencia kérdése);
- az átvitt értelemben is használatos szavak;
- idiómák, félidiómák kötelékében álló igék.

2.3 A 'kerül' ige vonzatkerete

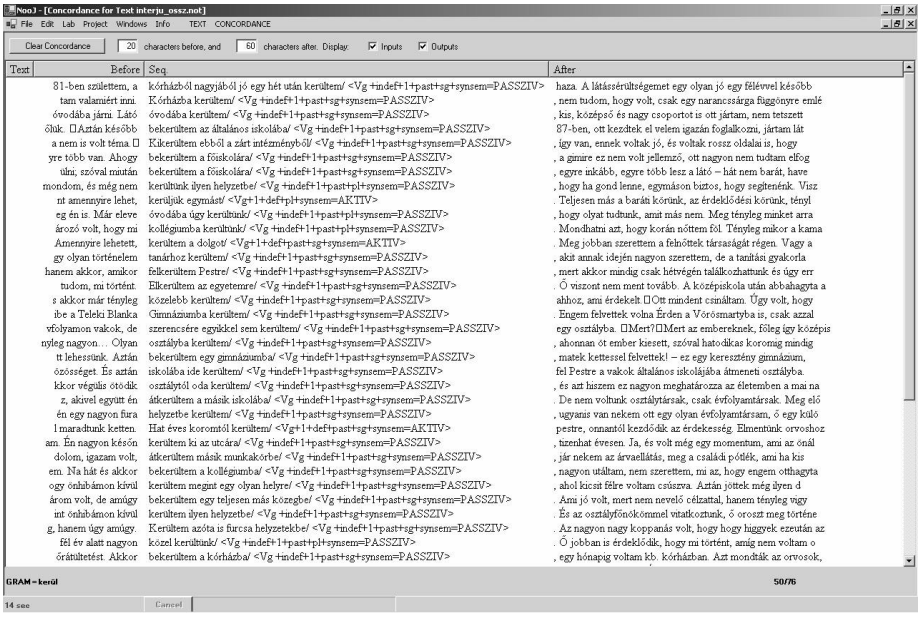
A 'kerül' igén kersztül szeretném bemutatni, hogy egy igének a ragozása, vonzatai, bővítményei, igekötői, illetve idiómán belüli megjelenése hogyan indokolhatja két kategóriába való bekerülését, illetve milyen megoldást kínálhat a besorolására egy gráf.



1. Ábra: Gráf a 'kerül' ige aktív és passzív formáira

Amennyiben a kerülni igét határozatlan ragozással használjuk, úgy passzív értelmű (pl. 'elkerül valahova', 'belekerül valamibe', 'elkerülök egy helyzetet', vagy 'kerülöm a találkozást valakivel', 'messze kerülök valamitől, ill. közel valamihez'), határozatlan ragozással pedig aktív értelmű lesz (pl. 'kerülöm a találkozást valakivel'). Mivel önmagában a határozatlan illetve határozott igeragozás nem tud különbséget tenni a két kategória között azonos megjelenési formái miatt, így feltételként a bővítményeket is tartalmazza a gráf. Helyet kapott benne továbbá egy idióma is a passzív oldalon ('sor kerül valamire'), mely megkötést tartalmaz a személyragokra nézve. Hiszen csak E/3 személyraggal lesz passzív, a 'sort kerítek valamire' alak már aktívnak számít.

A 2. ábra tartalmazza egy szövegen kapott találatokat. A mintán végzett manuális ellenőrzéssel egybevetve 74%-os pontosságot mutat az eredmény, 1,5%-os hibás találati arány mellett. A pontatlan találatot („Hat éves koromtól kerültem…") homonímia okozza, hiszen a 'hat' kifejezést mint tárgyragos főnevet ismeri fel. Egyes kifejezések jelen esetben két okból maradtak ki: a keresett szókapcsolatban ismeretlen - azaz a program alapszótárában nem annotált - szó állt (legtöbbször tulajdonnevek, pl. 'Bartókba kerültem'); illetve a mondat felszíni szerkezetében nem megjelenő vonzat miatt maradt ki a felismerés (pl. 'kerültem őt' helyett 'kerültem'). Míg ez előbbi probléma könnyen kiküszöbölhető, hiszen – véges számban – növelhető az alapszótár, addig az utóbbi megoldása jelenleg még kutatásunk részét képezi.



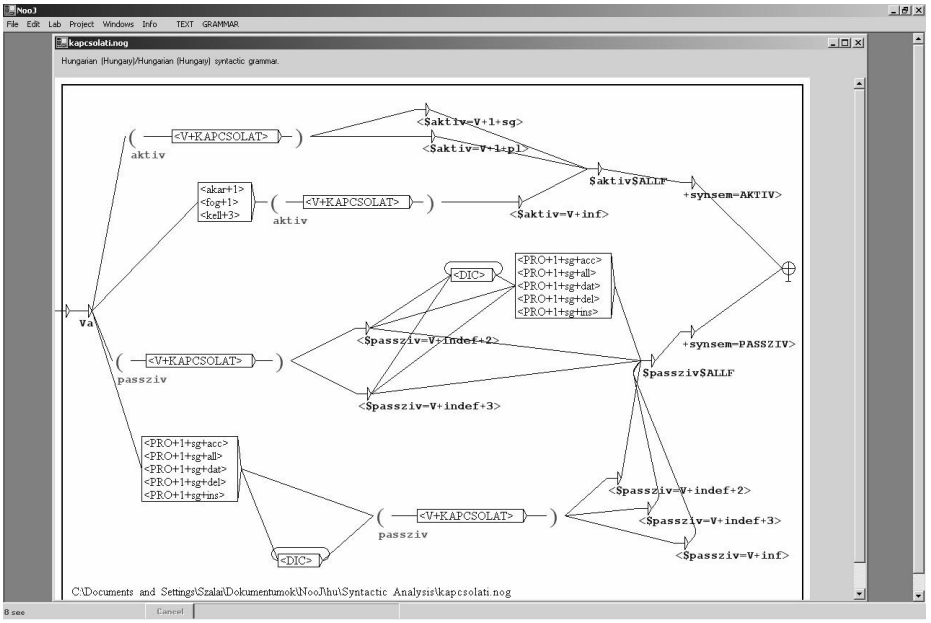
2. ábra: A 'kerül' ige gráfjának találatai a szövegben

2.4 Az elbeszélő mint alany és vonzat

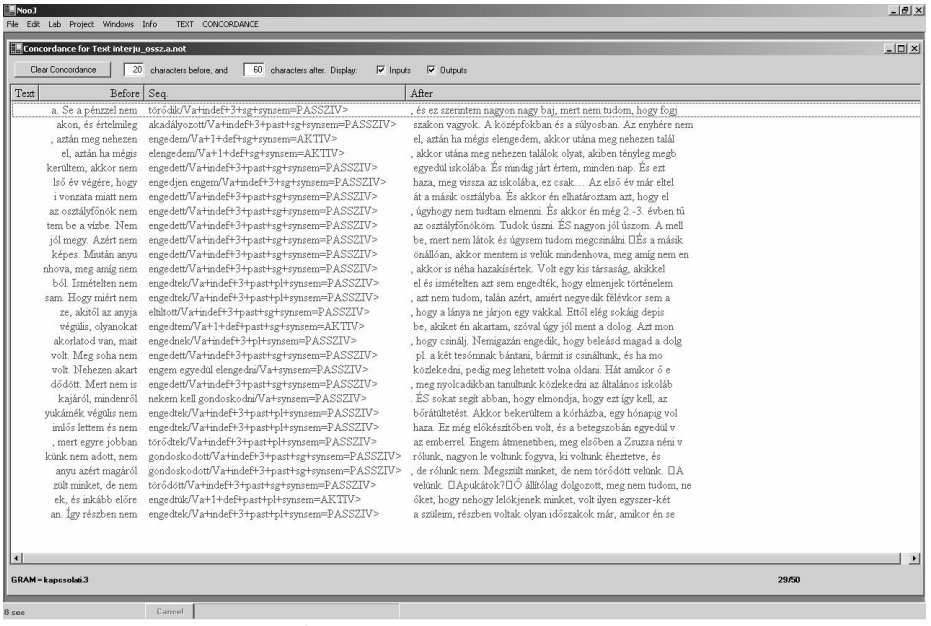
Sok ige egyszerre tagja az aktív illetve az állapotváltozás, történés passzív szótárának. Az elbeszélés menetében alanyként megjelenő történetmondó aktív értelművé teszi az igét, míg a vonzataként megjelenő pedig passzív értelművé (pl. 'én elhagyok valakit', vagy 'engem hagy el valaki'). Ezen tematika főként az igék azon csoportjánál lesz nagy jelentőségű, melyek egy kapcsolat vezetésére vonatkoznak, illetve az önállóság témaköréhez kapcsolódnak. (pl. elhagy, elenged, elhanyagol, eltilt, gátol, gondoskodik, kihasznál, kiközösít, kontrollál, megakadályoz, megcsal, szabályoz, törődik). Ezen igéket egy előkészítő gráffal 'kapcsolat' megkülönböztető jeggyel láttuk el a vizsgált szöveg annotációjában. A 3. ábrán látható gráf, mely az aktív és passzív megjelenési formákat próbálja meg elkülöníteni, csak az ezen igékre való utalást tartalmazza.

Ismét fontos megkülönböztető vonás volt a személyrag (pl.: 'én kiközösítek valakit', vagy 'ők engem közösítenek ki'), a határozott illetve határozatlan ragozás, továbbá a vonzatok megjelenése.

A vizsgált szöveg ezen igékre való manuális jelölésével a gráf 60%os egyezést mutat. Hibás találati aránya 7%. A találatokat a 4. ábra mutatja be.



3. ábra: 'Kapcsolati' gráf



4. ábra: A 'kapcsolati' gráf találatai

2.5 Idiómák

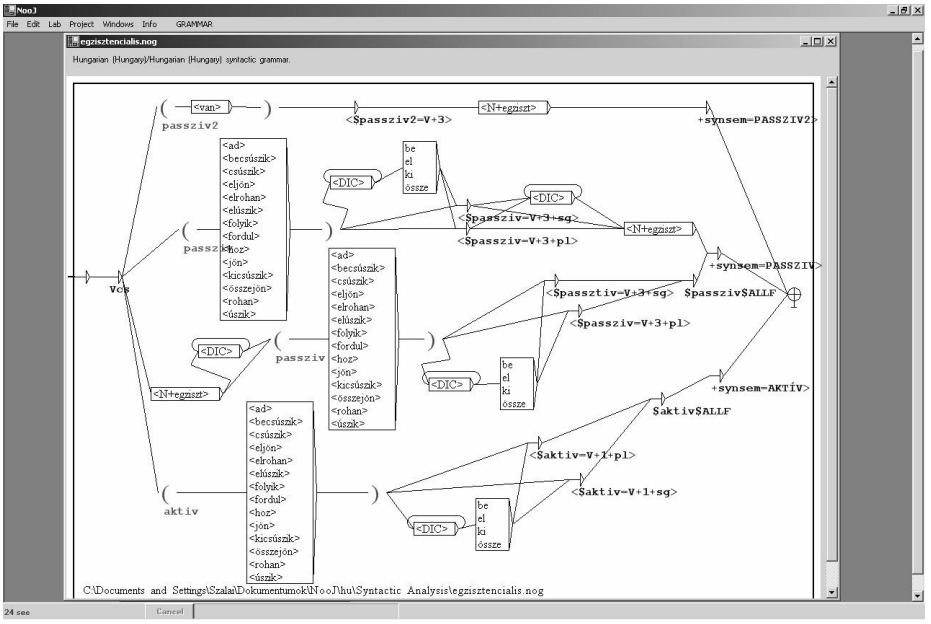
Mint fentebb látszik, gazdaságossági szempontból érdemes szabályszerűségeket felfedezni az igék megjelenési formáiban, és igecsoportokat felhasználni gráf szerkesztésekor.

Szintén egy csoportnak tekinthetők azon mozgást jelentő igék, melyek a következő absztrakt főnevekkel alkothatnak kifejezést: élet, idő, alkalom, helyzet stb. (Pl.: 'Jött egy lehetőség', 'becsúszott egy hiba', 'eljön majd az az idő' stb.) Aktívnak tekinthetők ezek az igék, ha saját mozgását fejezi ki vele az elbeszélő. A passzivitás állapotváltozás, történés kategóriáját kapja, ha a tárgyi/absztrakt világra alkalmazza azokat. Továbbá – a gráf egy újabb szálán – megjelenik a 'van' létige, mely megjelenése ezen absztrakt főnevekkel többnyire az állapotot kifejező passzív kategóriát képviselik.

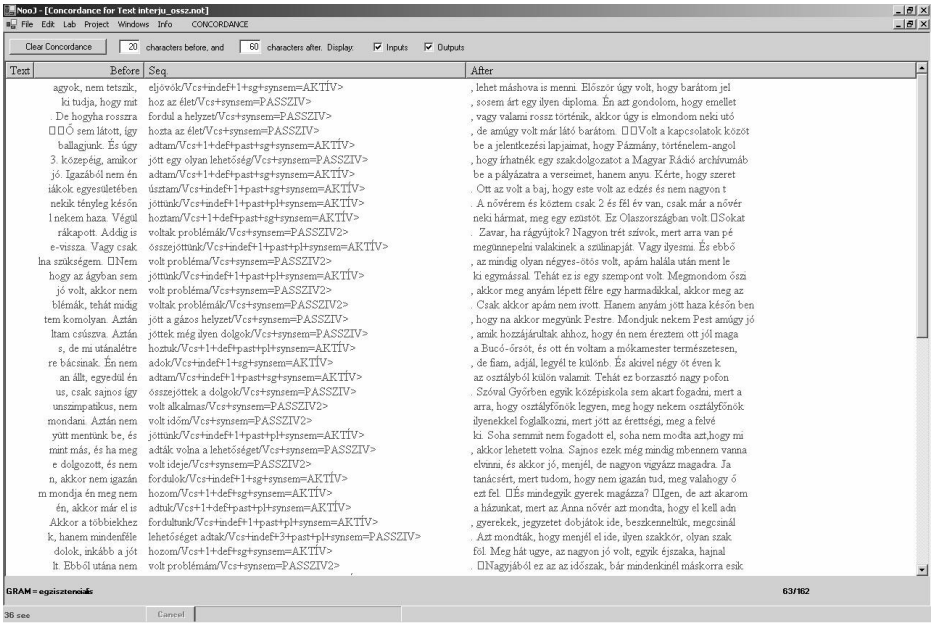
Egy előkészítő gráf segítségével a szövegen belül 'egziszt' kóddal láttam el vizsgálni kívánt absztrakt főneveket, így az 5. ábrán mellékelt gráfban nem kellett külön a felsorolást beilleszteni, csak e korábban már említett annotációt. (Ez jelentősen egyszerűsíti a gráfot, továbbá lehetővé teszi további alkalmazási lehetőségeit.)

Az általunk vizsgált szövegben – a gráfba foglalt esetekre nézve – a helyes találati arány megközelítőleg 80%. (Találatok a 6. ábrán láthatók.)

Hibás találatokat okozhat – a fent említetteken túl – a tagadás, főként pedig ha a mozgást jelentő igék átvitt értelemben is használhatók (pl.: 'összejön valakivel', 'rájön valamire', 'renbe jön egy betegség után').



5. ábra: Egzisztenciális kifejezések, idiómák



6. ábra: Az ‘idiómák’ gráfjának találatai (PASSZÍV kimenettel láthatók az állapotváltozás, történés passzív kategóriájának találatai, PASSZÍV2 kimenettel pedig állapotot, folyamatosságot kifejező passzív kategóriáé)

3. Befejezés

Természetesen az itt bemutatottakon kívül jóval több igére, igecsoportra készülnek gráfok a teljes szótárhoz. Jelen helyzetben igyekeztük bemutatni, milyen megoldható, illetve a jövőben majd megoldandó problémák merülnek fel munkánk során.
Célunk, hogy az általunk összeállított szótár minél szélesebb körű szövegeken alkalmazható legyen, minél pontosabb találati arány mellett. Ennek érdekében igyekszünk a magyar nyelv árnyalataira figyelve sokoldalú és rugalmas szótárt összeállítani.

Bibliográfia

1. 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.
2. László, J.: A történetek tudománya. Bevezetés a narratív pszichológiába. Új Mandátum Könyvkiadó, Budapest (2005)
3. 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.
4. Peck, D. – Whitlow, D. : Személyiségelméletek, Gondolat, Budapest. (1983) 97-106.
5. Reboul, A., - Moeschler, J.: A társalgás cselei. Bevezetés a pragmatikába. Osiris, Budapest (2000)

6. Ricoeur, P.: A narratív azonosság. In.: László j. – Thomka B. (szerk.) Narratívák 5. Narratív pszichológia. Budapest, Kijárat Kiadó (2001) 15-27.
7. Szakács F. – Kulcsár Zs.: Személyiségelméletek. Budapest, ELTE (2001)

A mentális igék szótára, valamint alkalmazása az automatikus tartalomelemzésben[83]

Vincze Orsolya[1], László János[2]

[1] PTE Pszichológia Intézet
orsolyavincze@hotmail.com
[2] MTA Pszichológia Intézet
laszlo@mtapi.hu

Abstract. Korábbi kutatásainkban a narratív pszichológiai tartalomelemzés olyan eljárásait dolgoztuk ki, amelyek segítségével a szöveg nyelvi-strukturális tulajdonságait figyelembe véve, releváns pszichológiai információk tárhatók fel. A módszer a minőségi tartalomelemzés értelmezési szabadságához képest az elemző számára kötelező, objektivált eredményeket nyújt [1]. Magyar és osztrák történelemkönyvekkel kapcsolatos vizsgálatunkban [2] az identitásközvetítés módjainak nyelvileg megragadható formáit azonosítottuk a mentális aktusokat megjelenítő igék kódolása révén, amelyek az érzelmi azonosulás mintázatairól tanúskodtak. A mentális aktusok kognitív szinten megjelenő cselekvések (gondol, dönt, lát, tud). Jelen kutatásunkban elkészítettük a mentális aktusokat tartalmazó igék szótárát a Magyar Nyelvtudományi Intézet Korpusznyelvészeti Osztálya által rendelkezésünkre bocsátott 10000 leggyakoribb igei listából. A cikkben a mentális szótár bemutatását, valamint az automatikus tartalomelemzésben való alkalmazását szeretném megmutatni.

1 Bevezetés

Mások szándékainak és gondolatainak értelmezési képessége humánspecifikus adottság. Az ember ezen képességét a pszichológia a mentalizáció terminusával illeti. Jelentőségét mutatja, hogy a saját és mások szándékainak figyelembevétele a viselkedések oki magyarázatában, már az egyedfejlődés korai szakaszában megjelenik [3], [4], [5].

A mentalizációs képességeknek a viselkedésmagyarázatban játszott szerepén kívül az önreflexivitásban és az érzelemszabályozásban is jelentős szerepe van. Saját mentális tevékenységeink hozzáférésével mentális állapotaink tudatos átélőjévé válhatunk. Kísérleti adatok azt mutatják, hogy a hároméves gyerekek már képesek megérteni a vágy, a kívánság valamint a vélekedés mentális állapotát, amelyet a szóhasználatukban megjelenő mentális terminusok is tükröznek (pl. „gondol”, „álmodik”, „tud”) [6], [7]. Saját és másoknak tulajdonított belső állapotok hozzáférése révén

[83] NKFP 6/074/2005 számú pályázat támogatásával készült

érzékenyebben tudjuk értelmezni saját állapotainkat és bejósolni mások jövőbeni viselkedését.

1.1 A mentális folyamatok nyelvi kifejezésének pszichológiai jelentősége

A mentális folyamatok nyelvi kifejeződésének elsősorban az elbeszélésekben van jelentősége. Ezeknek a nyelvi eszközöknek a révén kapunk betekintést a szereplők tudatába. A mentális kifejezéseknek nagy szerepe van az empátia és az azonosulás folyamataiban. A mentális kifejezések szótárba foglalását a LIWC (Linguistic Inquiry and World Count [8] is megvalósítja, azonban a LIWC alszótára csak kognitív elemeket tartalmaz és csak a szavak szintjén keres. Saját megközelítésünk valamennyi mentális folyamatot és állapotot leíró kifejezésre kiterjed, az igék mellett bevonja a főneveket, a mellékneveket és a határozókat is (a beszédaktusokat és az interpretatív igéket viszont nem). Ez a megközelítés nem csupán szavak szintjét, hanem a mondat szintjét is figyelembe veszi.

2 Mentális igék szótára

Jelen kutatásunkban elkészítettük a mentális aktusokat tartalmazó igék szótárát a Magyar Nyelvtudományi Intézet Korpusznyelvészeti Osztálya által rendelkezésünkre bocsátott 10000 leggyakoribb igei listából. Mentális akciónak minősült minden olyan ige, amely mentálisan végrehajtott cselekvést jelöl. A kódolást 7 független bíráló ellenőrizte.

A csoportosítást két szempontból végeztük. Elsősorban az olyan mentális igéket válogattuk ki, amelyek önmagukban hordozzák a mentális jelleget (n=308). Ezeket az igéket *abszolút mentális igéknek* neveztük el, amelyek közül néhány példa látható az 1. táblázatban.

1.Táblázat: példák az abszolút mentális igék listájából

ábrándozik	ábrándoz	aggodalmaskodik	aggódik
aggód	áhít	áhítozik	akar
akaródzik	akceptál	álmélkodik	álmodik
álmodozik	általánosít	dönt	analizál
asszociál	átél	átérez	átértékel
átértelmez	átgondol	átlagol	átszámít
átszámol	azonosít	azonosul	bánkódik
beazonosít	beképzel	beleél	beleért
belegondol	bizakodik	bizonytalankodik	csalódik
betanul	alultervez	elbizonytalanodik	elbízik
elhisz	elfeled	elégedetlenkedik	elfeledkezik
elfelejt	elfelejtkezik	elgondol	elgondolkodik
elles	elhatároz	elcsodálkozik	elképed
elképzel	elképzel	elgondolkozik	elmélázik
elméláz	elmélkedik	elmereng	elõrelát
fölidéz	fölismer	fölmér	föltételez
furcsáll	gondol	gondolkodik	gondolkozik

Ezen kívül kiválogattuk azokat az igéket is, amelyek csak bizonyos szókapcsolatban, vagy egy nyelvtani szerkezetben jelenítenek meg mentális akciókat. Ezeket az igéket *szabályalapú mentális igéknek* neveztük el (n=302) (2. táblázat)

2. Táblázat: példák a szabályalapú mentális igék listájából

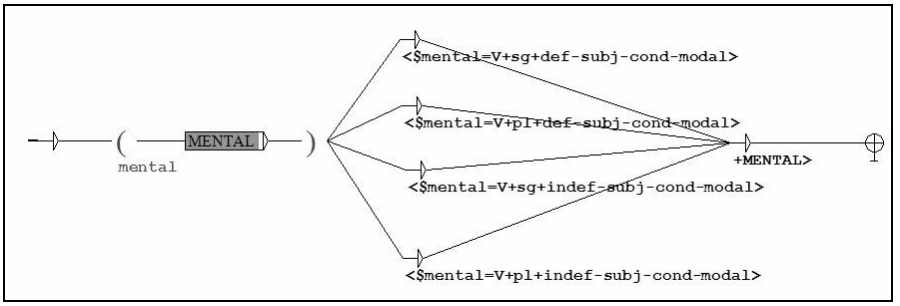
ige	gyakoriság	kitétel
általat	367	visszaható értelemben mentális
csiszol	738	elmét
átcsoportosít	1334	gondolatatot
felfrissít	305	felfrissítette az emlékezetét
átvesz	13390	gondolatot, ötletet
belát	5784	belát valamit
belemélyed	92	az emlékeibe, gondolataiba
forral	460	tervet, összeesküvést
csoportosít	791	szabályalapú
elakad	1217	szabályalapú miben;hol
elfogad	73604	nem tárgyat
elítél	7601	ha nem elitéltről van szó
elõretekint	140	ha tervez
eltekint	3809	eltekint valamitől
elvár	7028	valakitől valamit
elvet	2278	kivéve, ha magot vet el
kémlel	266	kémleli az eget
kifõz	185	kifőzi a tervet

A szabályalapú mentális igékre lokális nyelvtanokat a mentális főnevek és melléknevek szótárának elkészülte után lehet írni.

2.1 Lokális nyelvtanok az abszolút mentális igék esetében

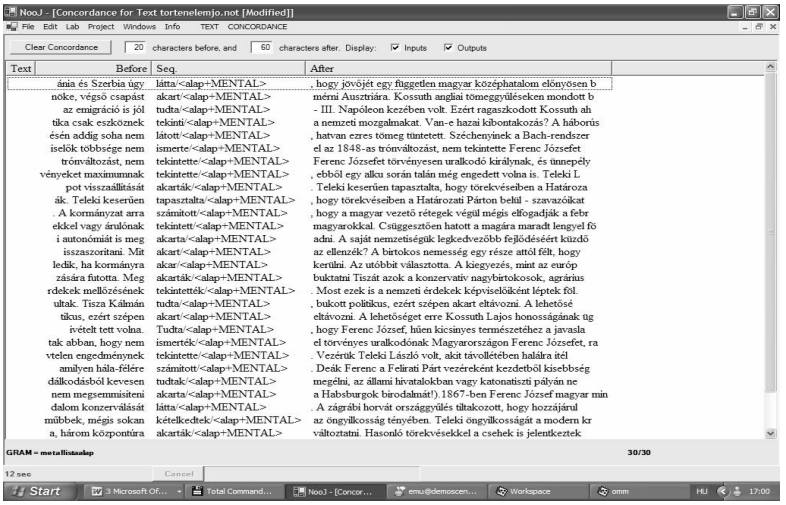
Az *abszolút mentális igék* csoportjára lokális nyelvtanokat írtunk a Nooj program segítségével, amelynek magyar nyelvre való adaptálását az MTA Nyelvtudományi Intézetének Korpusznyelvészeti Osztályának munkatársai végezték. Első lépésben egy olyan lokális nyelvtant alkottunk, amely két gráfból állt. Az első gráf tartalmazta az abszolút mentális igék listáját (MENTAL), amelyet egy következő gráfba ágyaztunk azzal a megszorítással, hogy személytől függetlenül egyes- valamint többesszámú igei alakokat keressen a program (1. ábra). Továbbá ki kellett zárnunk az igék feltételes módú ill. ható igeképzős alakját, mivel itt csupán az akció lehetséges jellegét kapjuk és nem magát az akciót. „*A magyar kormány szabadon dönthetett/döntene*” nem jelenti azt, hogy így is tett. Hasonló okokból kizártuk a felszólító módú alakokat is. Arra voltunk kíváncsiak, hogy a listában szereplő igék mennyiben támasztják alá azon elgondolásunkat, mely szerint az abszolút mentális igék nem igényelnek bonyolult nyelvtani megszorításokat, valamint nem szükséges hogy szó-

kapcsolatokban jelenjenek meg ahhoz, hogy a szövegben való előfordulásuk alapján valóban az abszolút mentális kategóriába sorolhassuk őket. Az elkészült gráfot előzőleg annotált történelmi szövegkorpuszon futtattuk.



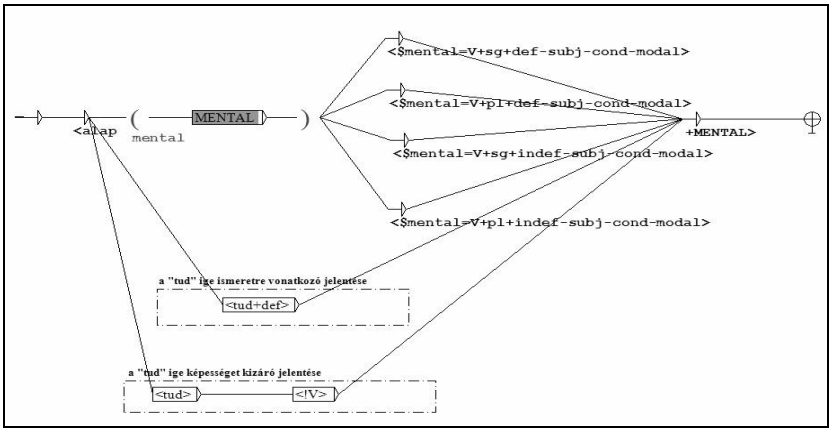
1. ábra Mentális gráf

Az eredmények igazolták feltevésünket (2. ábra). Azonban néhány ige esetében további megszorításokat kellett tennünk.



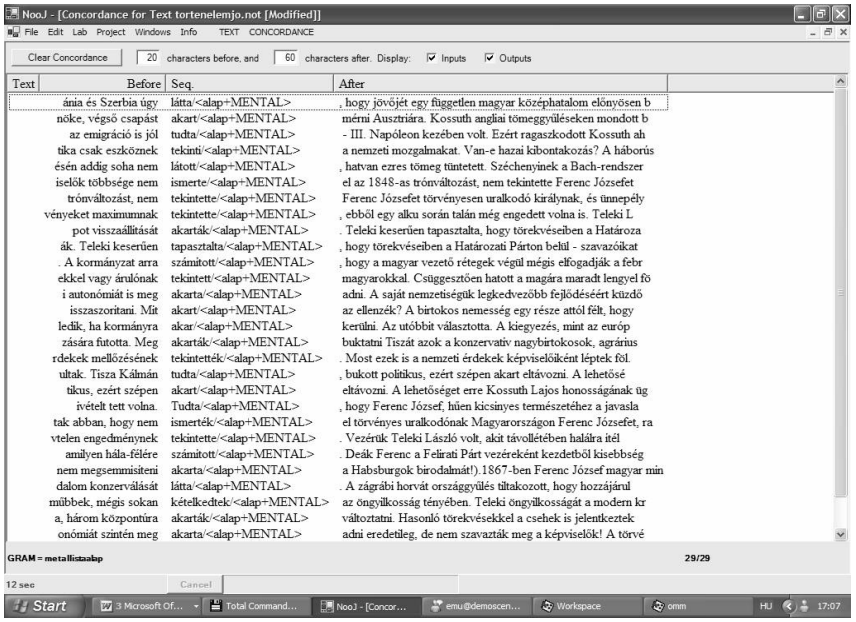
2. ábra A mentális gráf találatai

Mivel a „tud” ige csak abban az értelemben tekinthető mentálisnak, ha ismeretet jelent, egy további megszorítással ki kellett zárnunk azokat az eseteket, amelyekben a „tud” igét főnévi igenév követi, amely ebből következően képességre mutat rá (3. ábra).



3. ábra A mental gráf és a„tud" igére tett megszorítások

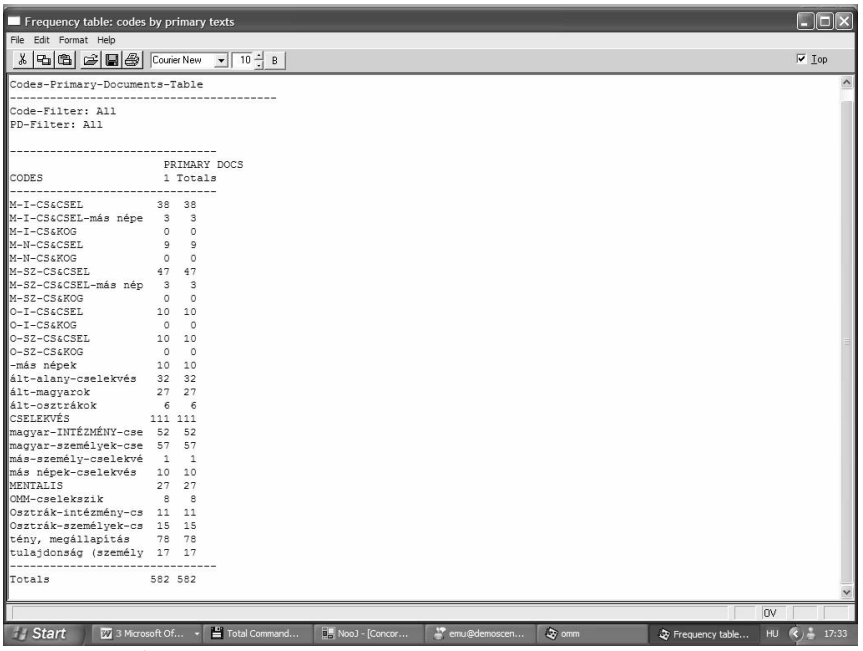
A 4. ábrán látható, hogy a gráf alkalmazta a „tud" igére vonatkozó megszorításokat és kiszűrte azokat az eseteket, amelyek képességekre vonatkoznak. Azonban egyenlőre nem képes kezelni az igékből képzett melléknévi igeneveket. A találatok közt szerepel az „*árulónak tekintett magyarok"* vagy a „*soha nem látott tömegek „* mondatrészek is.



4. ábra A mental gráf „tud" igei megszorításának eredménye

2.2 A mentális gráf összehasonlítása korábbi elemzéseinkkel

A „mental" gráfot összehasonlítottuk az Osztrák-Magyar Monarchiával kapcsolatos kutatásunk eredményeivel, amelyben az Atlas-ti program segítségével manuálisan kódoltuk a mentális akciókat. Az 5. ábrából jól látszik, hogy a manuálisan kódolt mentális akciók száma freq=27. Míg a Nooj programban alkalmazott „mental" gráf ugyanezen a szövegen freq=29 találatot adott. Azt a két esetet kivéve, amelyek a melléknévi igenevek problémájából adódtak, az automatikus és a manuális kódolás között nem volt eltérés.



5. ábra Az Atlas-ti programban manuálisan kódolt találatok a MENTALIS kód esetében

A „mental" gráf korábbi kutatási eredményeinkkel összehasonlítva megfelelően működött. Azonban ez nem jelenti azt, hogy nagyobb szövegkorpuszon futtatva nem jelentkeznek majd további problémák, amelyeket eddig nem vettünk figyelembe (pl. elváló/el nem váló igekötős igék). Azt is láttuk, hogy az igékből képzett melléknévi igenevek nehézséget okoznak a gráf helyes működésében. Úgy tűik, hogy ezt a problémát a melléknévi igenév mondatban betöltött hely és más szófajokkal való kapcsolatának meghatározásával sem lehet tökéletesen kiküszöbölni.

Jelenleg is folyik az MTA Nyelvtudományi Intézet által rendelkezésünkre bocsátott 40000-es főnévi listából a mentális főnevek kiválogatása, amelyek segítségével elkezdhetünk a szabályalapú mentális igékre lokális nyelvtanokat írni.

Végül a tervezett pszichológiai vizsgálatok (elsősorban empátia, identifikáció) szempontjából fontos lesz olyan nyelvtanok elkészítésére, amelyek a mentális kifejezéseket a mondatbeli szerepekhez, az alanyhoz (az elbeszélőhöz), a tárgyhoz vagy más szereplőhöz kapcsolják.

Bibliográfia

1. László, J.: A történetek tudománya. Bevezetés a narratív pszichológiába. Budapest, Új Mandátum Kiadó (2005)
2. Tóth J., Vincze O., László J. (2006): Osztrák és magyar középiskolai történelemkönyvek a monarchiáról. Szociálpszichológiai elemzés. *Educatio, 15,* 1, 174-182
3. Tomasello, M.: Gondolkodás és kultúra. Osiris Kiadó, Budapest (2002)
4. Leslie, A. M.: Pretending and believing: issues in the theory of ToMM. Cognition. 50. (1994) 211-238
5. Csibra G., Gergely Gy., (1998), A mentális viselkedésmagyarázatok teleológiai gyökere: Egy fejlődéslélektani hipotézis. In: Pléh Cs. (szerk.): Megismeréstudomány és mesterséges intelligencia, Akadémiai Kiadó
6. Kiss Sz..: Az „elmélet" elmélet és a szimulációs megközelítés szerepe a gyermek tudatelméletének magyarázatában. Pszichológia. 4. (1996) 383-396
7. Gopnik, A., Meltzoff, A. N., Kuhl: Bölcsek a bölcsőben. Typotex (2002)
8. Pennebaker, J. W., Francis, M. E., Booth, R. J.: Linguistic Inquiry and Word Count (LIWC): A Computerized Text Analysis Program. Mahwah NJ. Erlbaum Publishers (2001)

VIII. Poszterbemutatók

Simítás hasonlósági információ felhasználásával

Bíró István, Szamonek Zoltán, Szepesvári Csaba

MTA SZTAKI, Gépi Tanulás Kutatócsoport
1111, Budapest, Kende utca 13-17,
e-mail: szcsaba@sztaki.hu, zszami@elte.hu, ibiro@sztaki.hu

1. Bevezetés

Ebben a dolgozatban azt vizsgáljuk, hogy hogyan lehet a szavak egymáshoz való viszonyára vonatkozó információt kihasználva javítani a nyelvmodellek minőségén. Elviekben világos, hogy a szavak disztribúciós hasonlóságát kihasználva ugyannyi adat esetén jobb modelleket lehet építeni. Mivel azonban a disztribúciós hasonlóságra vonatkozó információ nem tökéletes, kérdéses, hogy az ebből adódó hiba ellenére is működhet-e egy a szóhasonlóságokra építő módszer.

A dolgozat fő eredménye az „SBS" (similarity based smoothing) algoritmus, amelyik képes kihasználni a szavakra vonatkozó hasonlósági információt, amennyiben ez az információ kellően pontos, míg ha az információ nem pontos, akkor az algoritmus addícionális vesztesége elhanyagolható.

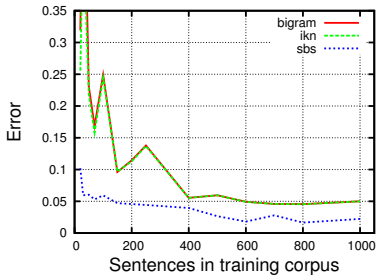
2. A javasolt módszer

A továbbiakban a módszert abban az esetben illusztráljuk, amikor bigram valószínűségeket ($\mathbb{P}(w_k|w_{k-1})$) akarunk modellezni.[1] A kiindulási pont az az észrevétel, hogy az egyes szókettesek valószínűségeit megfelelő ϕ_i bázisfüggvények választásával, a $\log \mathbb{P}(y|x) \propto \sum_{i\in I} \theta_i \phi_i(x,y)$ alakba írhatjuk és a tanulási feladatot így mint egy logisztikus regresszió feladatot is felírhatjuk. Egyéb információ hiányában minden információ-veszteséget elkerülő bázisfüggvényrendszernek „tartalmaznia kell" a $\phi_i(x,y) = \mathbb{I}_{\{i_1=x\}}\mathbb{I}_{\{i_2=y\}}$ függvényeket.[2]

A javasolt módszer lényege, hogy a szavak hasonlóságára vonatkozó információt a bázisfüggvények választásával visszük be az algoritmusba. Illusztrációként tekintsük azt az esetet, amikor a szavak csoportokba sorolhatóak és a $p(y|x)$ valószínűségek csak x csoportjától függ. Nyilvánvaló, hogy ebben az esetben a megfelelő bázisfüggvény választás $\psi_i(x,y) = \mathbb{I}_{\{i_1=y\}}\mathbb{I}_{\{c=C(x)\}}$, ahol $C(x) \in \mathcal{C}$ az x szó csoportja és $i = (i_1, c) \in \mathcal{W} \times \mathcal{C}$. Látható, hogy a bázisfüggvények száma nagyban csökkenthető, ha $\mathcal{C}$ számossága sokkal kisebb, mint $\mathcal{W}$ számossága.

[1] A bonyolultabb modellezési problémákra a bemutatandó elvek alapján triviális a módszert kiterjeszteni.

[2] Itt $i = (i_1, i_2) \in I = \mathcal{W} \times \mathcal{W}$.



1. ábra. A hiba a tanuló adat méretének függvényében.

Amennyiben bizonytalan, hogy a szócsoportok kellő információt tartalmaznake, akkor a fenti $\phi_\cdot$ és az itteni $\psi_\cdot$ bázisfüggvények együttesére kell építeni.Mivel az így kapott rendszer számossága már nagyobb lesz $|\mathcal{W}|^2$-nél, ahhoz, hogy ne veszítsünk a teljesítményből se akkor, ha relevánsak az osztályok, se akkor, ha nem azok, a regressziós irodalomban jól ismert regularizációt javasoljuk használni.Megmutatható, hogy az így kapott módszer valóban akkor is hatékony, amikor a szóosztályok relevánsak és akkor is, amikor nem.

Ha csak szavak közötti hasonlóságok állnak rendelkezésre, akkor a természetes megoldás a bázisfüggvények definiálására az ún. spektrális klaszterezés [1]. Az SBS algoritmus tehát egy a szavak felett definiált hasonlósági mátrixból kiindulva spektrális klaszterezés segítségével meghatározza bázisfüggvények egy rendszerét és erre építve a fent definiált regularizált logisztikus regressziós feladatot megoldva épít sztochasztikus nyelvmodelleket.

Az előzetes kísérletekben a módszer érzékenységét kontrollált környezetben, több szempontból is vizsgáltuk: a hasonlóságra vonatkozó információ minőségére, a módszer paramétereinek beállítására, illetve a rendelkezésre álló adatok mennyiségére nézve. Az 1. ábrán a tanult modell minősége látható a rendelkezésre álló minták számának függvényében.[3] A maximum norma hiba nyilván igen pesszimistán értékeli a modelleket (ezért nem látszik jobbnak az IKN mint a bigram az ábran), így igen biztató, hogy az SBS ebben a normában (is) jelentősen javít a korábbi módszerek eredményein. A valódi adatokkal végzett előzetes kísérletek eredmények szerint a módszer versenyképes a jelenlegi legjobb módszerekkel is.

Hivatkozások

1. Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD*, pages 269–274, 2001.
2. Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

[3] Összeasonlítási alapként a maximum-likelihood becslést ("bigram") és az IKN módszert használtuk. Az IKN [2] az átlagos egy state-of-the-art módszer, amelyiknek szintén az a célja, hogy a kis minták okozta minőségromlási problémákat megoldja.

Néhány nyelvstatisztikai módszerrel végzett elemzés összehasonlítása

Bujdosó Iván

ELTE BTK Alkalmazott Nyelvészeti Tanszék
bujdosxo@yahoo.com

1. Bevezetés

Az elemzés a szóstatisztikai elemzés körére korlátozódik. Az elemzési módszerek:

- egyszerű összeszámlálás és elemi matematikai műveletek
- normál eloszlási görbe, átlag és szórás számítása
- hatvány törvény alkalmazása
- mesterséges neurális hálózatok alkalmazása

A vizsgált nyelvek: az Európai Unió összes hivatalos nyelve és az eszperantó. A vizsgált nyelvek a kapott számszerű értékek szerint sorrendbe rakhatók. A vizsgálat végeredményének megelőlegezésével a nyelvek sorrendje:

1. finn; 2. észt; 3. magyar; 4. litván; 5. lett; 6. szlovák; 7. cseh; 8. lengyel; 9. szlovén; 10. máltai; 11. eszperantó; 12. görög; 13. dán; 14. svéd; 15. német; 16. olasz; 17. portugál; 18. spanyol; 19. francia; 20. holland; 21. angol

A vizsgált szövegek:

- Szerződés európai alkotmány létrehozásáról I-II. rész. A nyelvenként mintegy 35 oldalas anyag az összes felsorolt nyelven megtalálható az interneten.
- 283 regény angolul, eszperantóul, franciául, németül, olaszul és spanyolul; az internetről letöltve

Hipotézis: A nyelvek jellemezhetők egy, a vizsgálati módszertől függő számértékkel, és ezek alapján a nyelvek vizsgálati módszerenként sorba rendezhetők. Az így kapott sorok majdnem teljesen azonosak. Ezért az egyes vizsgálati módszerek a nyelv jellemzésére alkalmasak.

A vizsgálati módszerek közös jellemzője a szószintű vizsgálat. A szövegekből eltávolítottam a tagolásra szolgáló jeleket.

Egy szöveg szavai az alábbiakkal jellemezhetők: szóhossz, előfordulási gyakoriság, majd ezek alapján a szavak rangsora. Képzett értékek: a csak egyszer előforduló szavak száma (hapaxok), szókettősök, szóhármasok, szótávolság (ugyanazon szó következő előfordulása), Zipf egyenes jellemzői (meredekség, konstans, a regressziós együttható: R2), stb.

2. A négy különböző elvű vizsgálat:

Az első vizsgálatban az egyes nyelvnél a hapaxok számát arányítottam az összes szó számához. A két szélső értéket a finn (60%), illetőleg az angol (44%) adta.

A második vizsgálatban mind a 21 nyelvnek a szövegben előforduló szavát egy halmazként kezelve, megállapítottam egy "európai szóhosszúsági átlagot". Ebből kivontam az egyes nyelvekre jellemző értékeket (az egy betűs hosszúságú európai átlagból a kérdéses nyelv egy betűs szavainak a számát, ugyanígy a két-, a három-, stb-betűs szavak számát). Nyelvenként átlagot képeztem.

A harmadik vizsgálatnál minden nyelvnél megállapítottam a Zipf-görbe meredekségét. A két szélső értéket a finn (-0,76) és az angol (-1,11) adta.

A negyedik vizsgálatot egy amerikai egyetemen végezte el Manaris és munkatársai. Ők a mesterséges neurális hálók elméletét használták, a vizsgált 283 könyv szövegének terjedelme több nagyságrenddel nagyobb az alkotmánytervezet szövegénél. A nyelvek sorrendje azonban ugyanaz lett, mint az általam megállapított sorrend. Ezen túlmenően a kapott számszerű értékek és az általam végzett vizsgálat értékei jó egyezést mutattak, spanyol nyelv figyelembe vétele nélkül R2 = 0,98, spanyol nyelv figyelembe vételével R2 = 0,73.

A négyféle vizsgálat eredményeit láthatjuk az alábbi táblázatban a kapott számszerű jellemzőknek megfelelően. A táblázatban az egyes nyelveket az előző felsorolás számértékei jelentik, azaz a magyar nyelv az első vizsgálat szerint a 3., a második vizsgálat szerint a 9., a harmadik vizsgálat szerint a 3. helyen áll, a negyedik vizsgálatban nem szerepelt. A táblázatban a magyar nyelv kódját félkövérrel, az eszperantóét dőlt karakterrel szerepeltetem.

1	2	**3**	4	5	12	7	10	6	*11*	8	16	5	13	17	14	19	15	18	20	21
1	2	5	4	7	9	8	6	**3**	10	13	14	15	16	12	*11*	17	19	21	20	18
1	2	**3**	4	5	6	7	8	9	10	*11*	12	13	14	15	16	17	18	19	20	21
~	~	~	~	~	~	~	~	~	~	*11*	~	~	~	15	16	~	18	19	~	21

(Ahol ~ azt jelenti, hogy nincs vizsgált anyag erre a nyelvre, a vizsgálat Manaris szerint).

3. A fenti vizsgálatok eredményei:

1. Már egészen kis méretű korpuszból is következtethetünk arra, hogy a kérdéses szöveg milyen nyelven íródott.
2. A teljesen azonos tartalmú szövegek Zipf-együtthatóit sorrendbe állítva a nyelvrokonságról szóló ismereteinkkel összhangban levő képet adnak (a finnugor, a szláv, a germán, az újlatin nyelvek egymás mellett vannak, valamint a balti nyelvek a szláv nyelvek mellett). Ez az eredmény igazolja a kezdetben felállított hipotézisünket, azaz a kapott eredmény nem lehet a véletlen műve. Az eszperantó esetében pedig számszerű igazolását adja

Pennacchietti professzor 1981-ben kifejtett véleményének[84], amelyben az eszperantót tipológiailag a szláv és a germán nyelvek közé teszi, ellentétben azokkal a vélekedésekkel, amelyek agglutináló vagy izoláló jellegzetességeit hangsúlyozták.

Bibliográfia

1. Gledhill, C. 1998. The Grammar of Esperanto. A corpus-based description. München: Lincom Europa.
2. Manaris et al. 2006: Investigating esperanto's statistical proportions relative to other languages using neural networks and Zipf's law. Proceedings of the 2006 IASTED International Conference on ARTIFICIAL INTELLIGENCE AND APPLICATIONS (AIA 2006), February 13-16, Innsbruck, Austria.
3. Pennacchietti, F. 1981: Ne-hindeŭropaj trajtoj de la internacia lingvo, in: Sprachkybernetik, 1981, Paderborn, p. 95

[84] "La interna kohereco de Esperanto klariĝas do per tio, ke ĝi kapablas harmoniigi la postulojn de struktura simpleco, necesajn por vasta internacia uzo, kun la konservado de preciza *tipologia* stampo, nome tiu de la *ĝermanaj* kaj *slavaj* lingvoj de centra Eŭropo."

A kommunikációs fogalmak jelentésreprezentációjának egy modellje

Gyarmathy Zsófia[1], Szeredi Dániel[2]

[1] MTA Nyelvtudományi Intézet
gyzsof@nytud.hu
[2] BME – Média Oktató és Kutató Központ
daniel@szeredi.hu

Kivonat: Az alábbiakban bemutatjuk azt a koncepciót, amelyet létrehoztunk a kommunikációs igék jelentésreprezentációja során felmerülő problémák megoldásához. A komplex fogalmi terület leírásához három elmélet adott keretet: a konceptuális terek elmélete, Hobbs reifikációs eszköze, illetve Searle beszédaktus-elmélete.

A kommunikációs fogalmak példáján bemutatjuk az általános ontológiabeli fogalmak jelentésreprezentációjának egy módját. A munkát a MEO projekt [3] keretein belül végeztük.

A kitűzött cél a mindennapi tudat legalapvetőbb fogalmainak reprezentálása volt, amit konzekvensebben lehet elérni az egyes részterületek (*domainek*, szemantikai mezők, így például a kommunikáció) leírásával, mivel ezeken belül koherens, egységes rendszerben kezelhetők a fogalmak. Ehhez nyújt segítséget Gärdenfors konceptuális terekkel kapcsolatos elmélete [1], amely kognitív szemantikai elmélet szerint a fogalmak a fizikai térhez hasonlóan bizonyos dimenziók mentén szerveződnek. Másszóval az egyes fogalmak skálákba rendezhetők, és az egyes skálák által meghatározott térben egy pontosan kijelölhető (konvex) tartományt foglalnak el.

A jelentésleírás során tehát fontos szerepet kapott az egyes domainekben megjelenő skálák feltérképezése, és a domain fogalmainak ezen skálákon való elhelyezése. A kommunikációs domainben például a szóbeli megnyilatkozások fizikai megvalósulásai besorolhatók olyan skálákba, mint *hangerő*, *sebesség* és *tagoltság*. Ezek az általunk használt koncepcióban diszkrét skálák, pl.:

$$\mathbf{VelocitySpeechValues} = \{\mathbf{Rapid}_{\mathbf{Speech}}, \mathbf{Normal}_{\mathbf{Speech}}, \mathbf{Slow}_{\mathbf{Speech}}\} \quad (1)$$

$$\mathbf{VelocitySpeechScale} = \left\langle \mathbf{VelocitySpeechValues}, \prec_{\mathrm{Speech}} \right\rangle \quad (2)$$

Az egyes fogalmak jelentésreprezentációja ezek után egyszerűsíthető az egyes skálákon elfoglalt helyük meghatározásával:

Táblázat: három skála által feltárt fogalmak

fogalom	**hangerő**	**sebesség**	**tagoltság**
suttog	Quiet	-	-
mormol	Quiet	Obscure	-
motyog	Quiet	Obscure	Slow
hadar	-	Obscure	Rapid
darál	-	Clear	Rapid
kiabál	Loud	-	-
rikolt	Loud	Obscure	Rapid
ordít	VeryLoud	-	Rapid

A jelentésreprezentációhoz, melynek feladata az egyes nyelvfüggetlen fogalmak leírása, szükséges egy logikai nyelv. Ha ez a logikai nyelv gyenge, akkor jobban implementálható egyes alkalmazásokban; ha azonban erősebb, finomabban, pontosabban lehet a fogalmakat leírni. Ez utóbbi utat választottuk, mivel a pontosabb leírás igény szerint egyszerűsíthető egy gyengébb nyelv követelményeihez.

Az általunk használt nyelv tehát elsőrendű logika. A fogalmakat neo-davidsoni keretben írtuk le, valamint felhasználtuk a Jerry Hobbs [2] által bevezetett *reifikációs* eszközt (amelyet a predikátum neve utáni aposztróf jelöl). Ezáltal – a kvantoros formulákat kivéve – minden propozícióhoz rendelhető egy változó, amely a propozíció által jelölt tény fennállásának eseményszerűségét jelöli. Így például a 'Mari azt hiszi, hogy János beteg' fordítása

$$\exists e\ \mathbf{believe}(m,e) \wedge \mathbf{sick'}(e,j) \qquad (3)$$

A kommunikáció fogalmi rendszeréhez Searle egyik beszédaktusokról való elemzését [4] vettük alapul. Ez lehetővé tette az egyes kommunikációs fogalmak pragmatikai jellemzőinek egzakt leírását. Egy üzenetátvitelnek két pragmatikai feltételét különböztettük meg: az *előkészületi* és az *intencionális* feltételeket. Előbbi az adott beszédaktus megtörténtének külső feltételeit veszi számba, például felszólításkor többek közt a felszólítónak társadalmilag magasabb rangúnak kell lennie a felszólítottnál. Az intencionális feltétel az üzenetküldő szándékait írja le. A pragmatikai feltételek megadásában jól látszik a fenti reifikációs eszköz fontossága, mivel ezek a feltételek propozíciókra hivatkoznak, amelyek nem adhatók meg e nélkül az eszköz nélkül.

E feltételeknek, valamint a kommunikációs aktus tartalmának a leírása így ugyanolyan eszközökkel hivatkozhat a kommunikáció résztvevőire, propozicionális attitűdjeikre vagy a külső világra. Így például a parancsolás aktusának szemantikai tartalma mindig olyan ágenssel rendelkezik, amely második személyű; vagyis egy egyszerűbb modellben a befogadó:

$$\forall m,c,r\big[\ \big(\mathbf{deliverCommand}(m) \wedge \mathbf{contentOf}(c,m) \wedge \mathbf{recipientOf}(r,m)\big) \rightarrow \qquad (4)$$

$$\rightarrow \mathbf{agentOf}(r,c)\ \big]$$

Előkészületi feltétele ennek a kommunikációs aktusnak egyrészt az, hogy a forrás (a parancsoló) ténylegesen akarja a tartalom megvalósulását, másrészt, hogy a parancsoló társadalmi helyzete lehetővé tegye a parancs kiadását. Ebben az esetben tehát egy olyan összetett propozícióra kell tudni hivatkozni, amely két propozíció konjunkciója. Erre a célra Hobbs [2] bevezeti az *and* kétargumentumú predikátumot

(amelynek aposztrófos változatával tehát hivatkozhatunk a kívánt összetett propozícióra, mivel az argumentumaiban szereplő propozíciók együttes fennállásának tényét jelöli), amelyet axiómák segítségével kapcsol össze a konjunkció logikai konnektívummal (például biztosítja a kommutativitást). Így már leírható a parancsolás előkészületi feltétele:

$$\forall m,s,c,r,c_p\left[\left(\mathbf{deliverCommand}(m)\wedge\mathbf{contentOf}(c,m)\wedge\mathbf{sourceOf}(s,m)\wedge\right.\right. \quad (5)$$

$$\wedge\,\mathbf{recipientOf}(r,m)\wedge\mathbf{preparatoryConditionOf}\left(c_p,m\right)\Big)\rightarrow$$

$$\rightarrow\exists e_1,e_2\Big(\mathbf{and'}\left(c_p,e_1,e_2\right)\wedge\mathbf{wants'}(e_1,s,c)\wedge$$

$$\left.\left.\mathbf{superiorSociallyThan'}(e_2,s,r)\right)\right]$$

Intencionális feltétele a parancsolásnak pedig az a propozíció, miszerint a forrás célja, hogy a befogadó tudja meg azt (azaz magas meggyőzöttséggel rendelkezzen róla), hogy a tartalom megvalósulását a forrás akarja:

$$\forall m,s,c,r,c_i\left[\left(\mathbf{deliverCommand}(m)\wedge\mathbf{contentOf}(c,m)\wedge\mathbf{sourceOf}(s,m)\wedge\right.\right. \quad (6)$$

$$\wedge\,\mathbf{recipientOf}(r,m)\wedge\mathbf{intentionalConditionOf}\left(c_i,m\right)\Big)\rightarrow$$

$$\rightarrow\exists e_1,e_2\left(\mathbf{wants'}\left(c_i,s,e_1\right)\wedge\mathbf{believes'}\left(e_1,r,e_2\right)\wedge\right.$$

$$\left.\left.\mathbf{wants'}(e_2,s,c)\right)\right]$$

Ezzel a módszerrel az egy adott nyelven kifejezhető kommunikációs aktusok nagy részének szemantikai és pragmatikai feltételei leírhatóak. Emellett a teljes fogalomleíráshoz természetesen szükséges egyéb jellemzők formális modellezése is, úgy mint a szereplők (például forrás és befogadó) és a szintek (fizikai, nyelvi, jelentésbeli) elkülönítése, az aktusok többféle szempont (például a médium) alapján történő partícionálása. A kommunikációnak ezen aspektusaihoz azonban elégséges egy egyszerűbb leíró nyelv használata.

További kutatás egyrészt egy egyszerűbb, jobban implementálható logikai nyelv használa felé mutathat, ahol azonban bizonyos fogalmak már nem különböztethetőek meg, bizonyos jelentésreprezentációk már nem írhatóak le. Másrészt ezzel a formalizmussal további részterületek leírása is megadható, ahol a kommunikációs domainhez hasonló finomságú fogalmi megkülönböztetéseket lehet reményeink szerint elérni.

Bibliográfia

1. Gärdenfors, P. *Conteptual Spaces: The Geometry of Thought.* MIT Press Cambridge (2000)
2. Hobbs, J. R. *Discourse and Inference* (draft) (2003) `http://www.isi.edu/~hobbs/disinf-tc.html`
3. Magyar Egységes Ontológia projekt. `http://ontologia.hu`
4. Searle, J. R. *Az illokúciós aktusok szerkezete*. In Pléh Cs., Síklaki I., Terestyéni T. (szerk.) *Nyelv – kommunikáció – cselekvés.* Osiris Budapest (1997)

Automatikus tartalmi osztályozás és társítás kidolgozása az Igazságügyi Minisztériumba beérkező állampolgári levelekre

Kabai Dóra[1], Bigazzi Sára[2], László János[1,2]

[1] MTA Pszichológiai Kutatóintézet, 1132 Budapest, Victor Hugo u. 18-22.
{kabaidora, laszlo}mtapi.hu
[2] PTE BTK Pszichológiai Intézet, 7624 Pécs, Ifjúság útja 6.
bigazzisara@hotmail.com

Kivonat: Kutatásunk[85] célja egy olyan automatikus tartalomelemző módszer kidolgozása, amely az IM ügyfélszolgálatának munkáját segíti. A munkafolyamat összetettsége miatt két részfeladatot választottuk ki kutatásunk tárgyául, amelyek elképzelhetőek automatikus működésben. A levelek elsődleges szétosztása a főbb jogi területek mentén történik, a minisztérium szerkezeti felépítését követve. A jog logikájára támaszkodva kidolgoztunk egy kategóriarendszert, amiben kódoltuk a leveleket, majd csoportosítással vizsgáltuk a tematikus kategóriák differenciálhatóságát. A téma szerinti osztályozást tanulóprogramok segítségével végeztük. Mélyebb struktúrák feltárásához egy új módszertant dolgozunk ki. A természetes nyelven írt levelekben a témára és az eljárás stádiumára vonatkozó információk, események sorozata lefordítható a közigazgatás nyelvére. A szűk jogi területet érintő ügyek kifutási lehetőségei definiálhatók döntési utakként. Az utak leírásához sematikus forgatókönyvek alkalmazását javasoljuk, amelyek események sorozatát és köztük lévő relációkat is képesek magukban foglalni [2], [3], [4]. A levélben megjelenő események láncolata szűkíti a releváns forgatókönyvek halmazát, így lehetővé válik a megoldási lehetőségek automatikus társítása.

1 Levelek elosztása

Az ügyfélszolgálat komplex munkafolyamatának feltérképezése után elkülönítettünk olyan feladatrészeket, amelyek elképzelhetőek emberi jelenlét nélkül. A levelek beérkezésük után egy elsődleges osztályozáson esnek át, ahol az érintett jogi területek szerint továbbküldik őket a megfelelő osztályra. Az ügyfélszolgálat tapasztalataira és a jogszabályokra támaszkodva rögzítettünk egy kategóriarendszert, amelybe a leveleket besoroltuk. Klaszterező eljárással vizsgáltuk a korpusz klaszterekbe sorolhatóságának mértékét. Eddigi eredményeink azt mutatják, hogy sem az entrópia, sem a tisztaság értéke nem tér el jelentősen az 50%-tól. További futtatások szükségesek

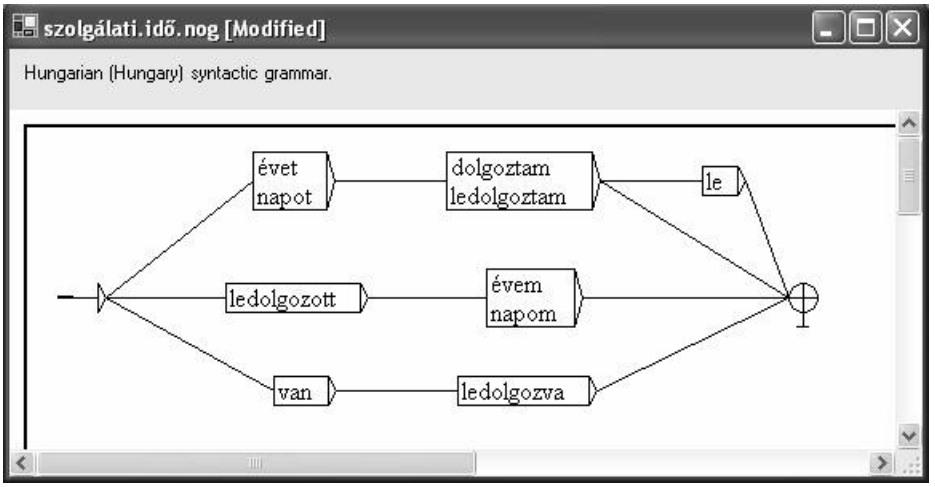
[85] Kutatásunkat az NKFP 6/074/2005 pályázat támogatja.

ahhoz, hogy végső következtetéseket vonjuk le, például a minta növelésével javulhatnak az eredmények.

A klasszifikáció eddigi eredményei eltérést mutatnak a téma és más dimenziók esetében, ami alátámasztja azt az elgondolásunkat, hogy a tematikus jegyek elemzése szavak, kifejezések szintjén elégséges lehet. A találati arány javuló tendenciát mutatott a tanuló-, tesztadat arány változtatásának függvényében, azonban további fejlesztések, a korpusz növelése szükséges, hogy a találati arány elfogadható hibaszázalékkal működjön.

2 Társítás

A egyedi esetkehez való megoldási lehetőségek automatikus társításának feltétele a nyelvek átjárhatósága és a probléma pontos beazonosítása. A természetes nyelven írt levelekben megjelenő információk, események lokális szintaktikai, szemantikai szabályok segítségével lefordíthatók rögzített közigazgatási kifejezésekre[86]. Az 1.ábrán láthatunk egy gráfot, ami a szolgálati időre vonatkozó kifejezéseket kódolja. Ha a levélben elmesélt történet minden pontját lefordítjuk jogi nyelvezetre, tulajdonképpen egy jegyzetet készítünk a levélből, amely a jog fogalmi rendszerében értelmezhető.



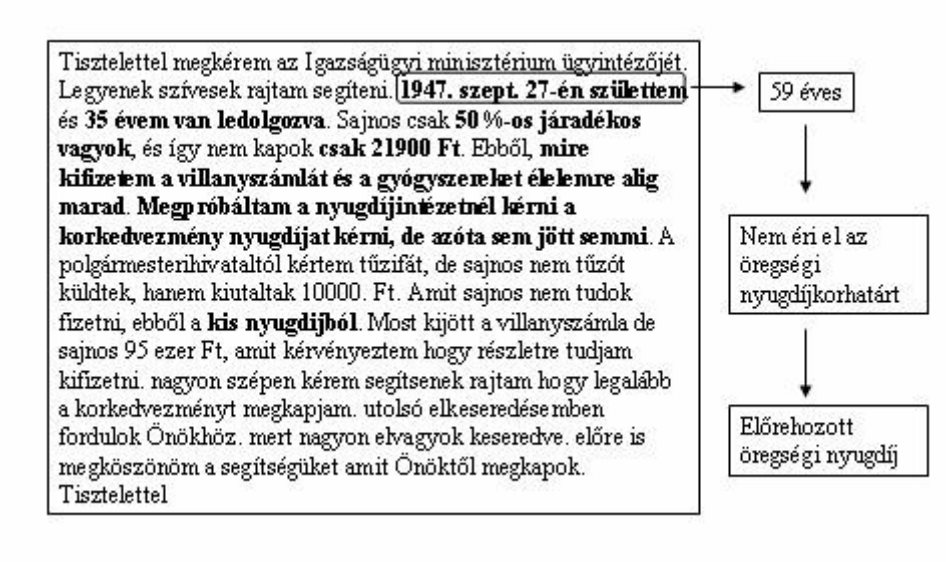
1. ábra. A szolgálai időre vonatkozó kifejezések gráfja.

A probléma pontos beazonosításához az érintett jogi területet döntési utak formájában felvázoljuk. Az utak események kötött sorrendjéből állnak, amelyekhez egyértelműen társíthatóak kifutási lehetőségek. Minden út leírásához egy sematikus forgatókönyvet használunk, amely jogi kifejezéseket tartalmaz. A levél fordításából készült eseménysor összevethető a forgatókönyvekkel, amelyek az egyezés mértéke szerint sorrendbe

[86] A Nyelvtudományi Intézet fejlesztése és együttműködése révén nyílt lehetőségünk a Nooj nevű program használatára.

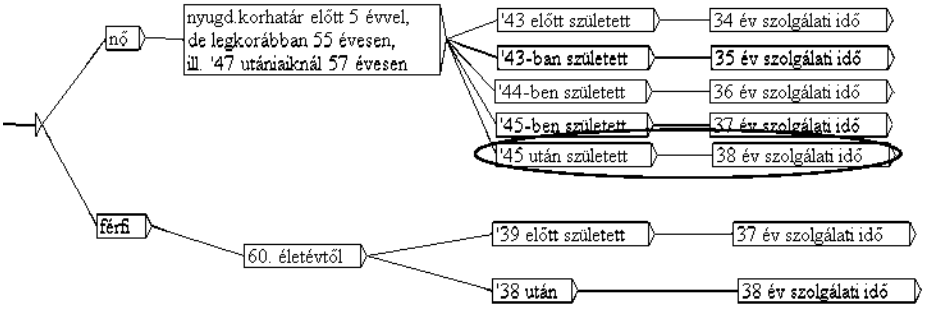
rendezhetők. Ezzel a módszerrel a megoldási javaslatok listája automatikusan társítható az állampolgárok leveleihez.

A következőkben egy példán keresztül szemléltetjük a társítás folyamatát. A nyugdíjjal kapcsolatos panaszok, kérések úgy tűnik jól elhatárolható csoportot alkotnak. A beérkező levél nyelvi markereinek detektálása révén szűkíthetjük a vonatkozó jogi területet és a releváns forgatókönyvek halmazát. Az 2. ábra egy panaszos levelet tartalmaz, amelyben nyelvi markerek utalnak a nyugdíj típusára („1947. szept. 27-én születtem", „35 évem van ledolgozva", „50%-os járadékos vagyok"), az eljárás folyamatának állapotára („Megpróbáltam a nyugdíjintézetnél kérni a korkedvezmény nyugdíjat kérni, de azóta sem jött semmi").



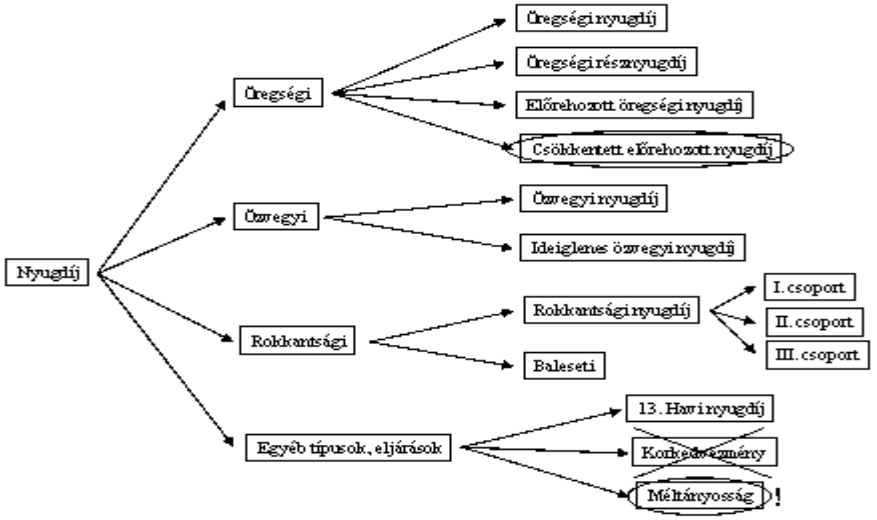
2. ábra. Egy nyugdíjjal kapcsolatos állampolgári levél. A nyelvi markerek detektálása szűkíti a vonatkozó jogi területet.

A születési évszám arra vonatkozó információt tartalmaz, hogy az állampolgár elérte-e a nyugdíjkorhatárt. A jogszabályok egyértelmű feltételeket szabnak az életkor és a szolgálati idő tekintetében. A 3. ábrán látható gráf mutatja az előrehozott öregségi nyugdíj feltételi hálózatát. A példában szereplő eset adatai alapján nem jogosult előrehozott öregségi nyugdíjra, mert nem rendelkezik az előírt minimális szolgálati idővel.



3. ábra. Az előrehozott öregségi nyugdíj feltételei egy döntési fával modellálva.

A levélből kinyert információk és a jogi adatbázis oda-visszaható szűkítése és bővítése révén zárhatunk ki, illetve vethetünk fel megoldásként újabb kifutási utakat. A 4. ábra szemlélteti, hogy a példában felmerültek alapján az előrehozott öregségi nyugdíj kizárásával a nyugdíj típusát illetően meghatározható a csökkentett előrehozott nyugdíj. Míg az egyéb eljárások tekintetében kizárható a korkedvezmény, azonban felmerül a méltányossági kérelem beadásának lehetősége.



4. ábra. A nyugdíjhoz kapcsolódó jogi terület ábrázolása egy döntési fa formájában.

Bibliográfia

1. Abonyi J.: Adatbányászat. A hatékonyság eszköze. Gyakorlati útmutató kezdőknek és haladóknak. ComputerBooks, Budapest (2006).

2. Graesser, A.C., Pomeroy, V., & Craig, S. (2001). Psychological and computational research on theme comprehension. In W. van Peer and M.M. Louwerse (Eds.), Thematics in psychology and literary studies (pp. 19-34). Amsterdam: Benjamins.
3. László J.: A történetek tudománya. Bevezetés a narratív pszichológiába. Budapest: Új Mandátum Könyvkiadó (2005)
4. Schank, R. C., Abelson, R.: Scripts Plans Goals and Understanding – An Inquiry into Human Knowledge Structures. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.
5. Schank, R. C.: Dinamikus emlékezet. A forgatókönyv-elmélet újraértelmezése. Budapest: Vince Kiadó. (2004)

Anaforafeloldás magyar nyelvű szövegekben

Lejtovicz Katalin Eszter[1], Kardkovács Zsolt Tivadar[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
H-1117 Budapest, Magyar Tudósok krt. 2.

1 Bővített kivonat

Napjainkban az anaforafeloldás problémájának algoritmikus megoldása egyre inkább döntő fontosságú, és sürgető feladat, mivel számos alkalmazásban szükség van az utalószavak gyorsan és helyesen történő feloldására. Gépi fordítás esetén fontos, hogy a személyes névmási anaforákat - amennyiben a célnyelv megkülönböztet nemeket - a célnyelv megfelelő nemű személyes névmási anaforájára fordítsuk. Tény-kinyerést megvalósító algoritmusok implementálásakor figyelembe kell venni, hogy hosszabb szövegek esetén a szöveg elején elhangzott központi témára később sokszor utalószóval hivatkozunk. A keresők, illetve indexelők esetében a találati pontosság 15-20%-os hibájának egyik legmeghatározóbb oka, hogy a szövegben található anaforikus elemek különböző kifejezésekre képződnek le, ezért a szógyakoriságon alapuló indexelés eredő hibája és bizonytalansági tényezője relatív magas. Általánosságban tehát azt mondhatjuk, hogy a helyes anaforafeloldás bármilyen szövegfeldolgozással kapcsolatos területen kulcskérdés.

A feloldásra használt algoritmus kiválasztása esetén figyelembe kell venni, hogy milyen fajta anaforák fordulnak elő a magyar nyelvben, és hogy az egyes típusok milyen gyakorisággal jelennek meg a szövegekben. Az anaforák hat legfontosabb típusa a következő: névmási, zéró, határozói, NP, igei, rész-egész (lásd 1. táblázat). A felsorolás egyben tükrözi az egyes típusok előfordulási gyakoriságát is, balról jobbra haladva az egyre csökkenő gyakoriságú anaforákat tüntettük fel.

Az anaforafeloldási algoritmusok két fő típusba sorolhatóak, azaz megkülönböztetünk tudás alapú és tudásszegény algoritmusokat. A tudás alapú rendszerek emberi munkával előfeldolgozott bemeneten dolgoznak, a tudásszegény módszerek azonban automata parszolást végeznek. A tudás alapú feldolgozás jóval megbízhatóbb eredményeket ad mint a tudásszegény, viszont az emberi munka felhasználásából következően lassabb és költségesebb is annál. Ezidáig anaforafeloldást megvalósító algoritmusok főként angol nyelvre születtek. Magyar nyelvre a poszterben is bemutatásra kerülő program használható. Ez az algoritmus a tudás alapú kategóriába tartozik, és ezen belül is az úgynevezett CT (Centering Theory) elvet használja.

A CT a következőképpen foglalható össze:

1 Egy diskurzusban mindenegyes megnyilatkozásnak pontosan egy központi témája van.
2 Egy anafora nagy valószínűséggel a központi témára utal vissza.

A diskurzus során az egymást követő megnyilatkozások általában az előző mondat központi témáját folytatják.

~ anafora	Magyar nyelvű példa mondatok
névmási	*Péter* azért nem jött el a moziba, mert **ő** már látta a filmet.
zéró	A *lány* elment a boltba, de **(ő)** nem vitt magával pénzt.
határozói	Mi elmegyünk a *vendéglőbe*, és veled majd **ott** találkozunk.
NP	*Bush* felszólalt a szenátusban. **Az elnök** beszédében…
igei	A *kislány* énekelt, és a testvére is **így tett**.
rész-egész	Bár *Svájcban* 1,2€ a benzin litere, **Zürich**ben jóval drágább.

1. táblázat. Az anaforák típusai.

Az egyik legjobban ismert CT alapú algoritmus, Brenan, Friedman és Pollard 1987-es (röviden BFP) algoritmusa is a középpontba helyezés elvét használja.

A szakirodalomból ismert algoritmusok közül azért a BFP-t érdemes a magyar nyelvre alkalmazhatóvá tenni, mivel ez az algoritmus figyelembe veszi nyelvünk sajátosságait. A BFP algoritmus előnye, hogy jó találati aránnyal működik mind az izoláló típusú angol nyelvre, mind az agglutináló magyarra. Tehát jól működik a mondatrészek sorrendjét kevésbé megkötő magyar mondatok esetében, és a sorrendet jobban megkötő angol mondatok esetében is. A BFP algoritmus magyar nyelvre történő adaptálásával a leggyakrabban előforduló anaforák, vagyis a névmási, zéró és határozói anaforák feloldása valósítható meg. A program bemeneteként a Szeged Korpusz CD-n található nyelvtanilag elemzett mondatok szolgáltak. A CD-n található magyar nyelvű szövegek szintaktikailag és morfológiailag is elemzettek. Az algoritmus anafora-antecedens párok képzését, a biztosan rossz jelöltek kiszűrését és a megmaradtak közül a mondatok közötti átmenetek rangsorolása után kapott, legnagyobb valószínűséggel helyes pár kiválasztását végzi. A programot leginkább már csak nyelvészeti területen kell fejleszteni (szűrésben szereplő feltételek szigorításával, új információ-régi információ szerinti tagolással). A program tesztelése a Szeged Treebank 2.0 CD-n történt.

A kapott eredmények:

- A szövegben levő összes anaforának a 37%-át találja meg a program.
- A szövegben levő azon anaforákból, melyek megtalálását a programban megkíséreljük, 39,6% megtalálása sikeres
- A program 21%-ot old fel helyesen az összes anaforából
- A program 23,8%-ot old fel helyesen azokból az anaforákból, amelyek feloldását megkíséreljük

A magyarra megvalósított BFP algoritmus találati arányát (39,6%) összehasonlítva az angolra készült BFP algoritmuséval (59%) azt állapíthatjuk meg, hogy az angol nyelv esetében kb. 20%-kal értek el jobb eredményeket. Nagy valószínűséggel a már említett továbbfejlesztések hatására a magyarra írt program feloldási aránya el fogja érni az angolnál tapasztaltakét.

Fogalmi hálózat természetes nyelvű szövegek feldolgozásához

Németh Bottyán[1]

[1]BME - TMIT; AITIA International Rt.
bottyan@tmit.bme.hu

Kivonat: A számítógépes beszédfeldolgozás egyik legnehezebb kérdése, hogy hogyan rendelhetünk jelentést egy természetes nyelvű szöveghez. Erre egy lehetséges megoldás, a szöveg jelentését ábrázoló fogalmi reprezentáció létrehozása. A szöveg fogalmi reprezentációjának létrehozásához két dologra pedig feltétlenül szükség van, ugyanis kell egy reprezentációs nyelv, és kellenek szabályok, amikkel a természetes szöveghez hozzárendelhetjük a vele azonos jelentésű fogalmi struktúrát. A cikkben javaslatot teszek a szöveg fogalmi reprezentációját, illetve a „megértéshez" még szükséges háttér-információkat ábrázoló fogalmi rendszer felépítésére. Ennek a rendszernek érdekessége, hogy a hagyományos szemlélettel szemben a fogalmi hierarchia alapját nem osztályok képezik, hanem prototípus objektumok, amik rugalmasabb ábrázolást tesznek lehetővé. Végül megismerkedhet az olvasó a javasolt rendszer egy egyszerű alkalmazásával egy beszélgető-robot alkalmazásban.

1 A prototípus alapú fogalmi hierarchia

Napjainkban egyre aktuálisabb téma a természetes nyelvű szövegek számítógépes feldolgozása. Ezen belül a szöveg szemantikai értelmezése még egy viszonylag kiforratlan terület. A gyakorlatban használt eszközök nem foglalkoznak a szöveg szemantikájával, hanem „egyszerűbb", statisztikai alapokon, vagy logikai szabályrendszereken nyugszanak. A szemantika egyik lehetséges értelmezése a számítógép számára, a szöveg tartalmának valamilyen belső fogalmi hálózathoz kapcsolása. Hogyan épüljön fel azonban a szöveg jelentését ábrázoló rendszer? A fogalmak strukturált ábrázolásával, már Arisztotelész is foglalkozott. Ő osztályozta a világban fellelhető dolgokat, és az osztályokat hierarchiába rendezte. Ez a szemlélet az uralkodó napjainkban, és ilyen alapokra épülnek a különböző ontológiák is. A filozófusok ezt a szemléletet először a 19. században kritizálták, később Wittgenstein alakította ki a prototípus alapú szemléletet, aminek alapja a szigorú osztályba tartozással szemben a „családi hasonlóság", a családokat pedig prototípusokkal lehet jellemezni. Ez a szemlélet és sokkal rugalmasabb és sokkal inkább megfelel az ember mindennapi életben használt fogalmi hierarchiájának. Leírhatók vele olyan fogalmak, amik hagyományosan nehezen osztályozhatóak, illetve egyes példányok más-más nézőpontból más-más családba tartozhatnak [1]. Egy az arisztotelészi megközelítéssel nehezen osztályozható fogalom a „játék". Például vannak játékok, amit a nyereményér játszunk (lottó), de van

olyan is ahol nincs győztes és vesztes (babázás). Valahol a szerencsén van a hangsúly (lóverseny fogadás) valahol pedig nem számít a szerencse (sakk). A résztvevők száma is igen változatos lehet. Milyen tulajdonságokkal jellemezhetjük hát a „játékot". A prototípus alapú szemlélettel könnyen megoldhatjuk ezt a problémát a következőhöz hasonló definícióval. Játék például a foci, a babázás, a táblajátékok és a szerencsejátékok, és az ehhez hasonlóak.

Ugyan a legtöbb gyakorlati rendszer még az arisztotelészi szemléleten alapszik, vannak prototípus alapú fogalmi rendszert használó alkalmazások is, ezek egyike a Cougaar multiágens architektúra [2], ami főleg a modell könnyű bővíthetőségét használja ki **(fig. 1.)**. Az általam javasolt modell is a prototípus alapú szemléleten nyugszik. Ezt a döntést egyrészt a rugalmassága és egyszerűsége indokolta. Alkalmazásakor nem kell külön osztályokkal és példányokkal foglalkozni. A fogalmi modellt természetes nyelvű szövegek fogalmi reprezentációjához kezdtem fejleszteni [3], de később egy prototípus beszélgető-robot alkalmazás keretein belül implementáltam. Sajnos a megvalósíthatóság miatt néhány helyen kompromisszumokra volt szükség, ami gyengít a modell erején. Itt főleg arról van szó, hogy a családhoz tartozás meghatározásához szükséges hasonlósági mutató hiányzik az implementációból, emiatt a családhoz tartozást explicit kellett jelölni. Ennek ellenére megmaradt a rugalmasság és könnyű bővíthetőség. Fontosnak tartom azt is, hogy ez a leírás közelebb áll az emberi gondolkodáshoz, mint az arisztotelészi, ezért közvetlenebb lehet a kapcsolat a nyelv és a fogalmi reprezentáció között. Gondolok itt arra, hogy a „játék" definíciója itt nagyon hasonlít arra, ahogy egy gyereknek elmagyarázunk egy számára új fogalmat.

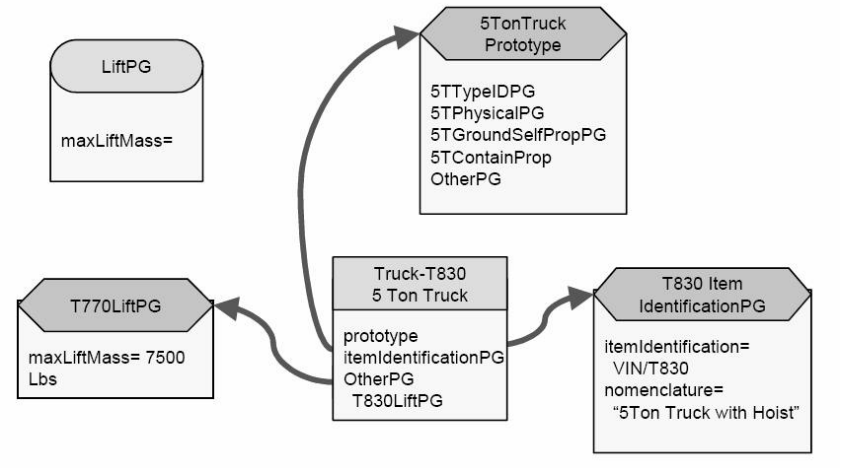


Fig. 1. A Cougaar rendszerben a fogalmi rendszer bővítése, ha az 5 tonnás teherautóra felszerelnek egy darut, és eddig nem volt darus teherautó a rendszerben.

2 Megvalósítás

Az implementáció első lépése egy teljesen általános adatstruktúra, egy irányított gráf. A gráf élei és a csomópontjai is címkézve vannak. A címkén és néhány szerkesztést segítő kiegészítésen kívül csak egy egyediséget jelző flag tartozik pluszban a csomó-

pontokhoz. Az egyedinek jelölt fogalmak azon a fogalmaknak felelnek meg amik az ember számára közvetlenül tapasztalhatóak alapvető tulajdonságok, és így egyértelműen megkülönböztethetők minden mástól **(Fig. 2.)**.

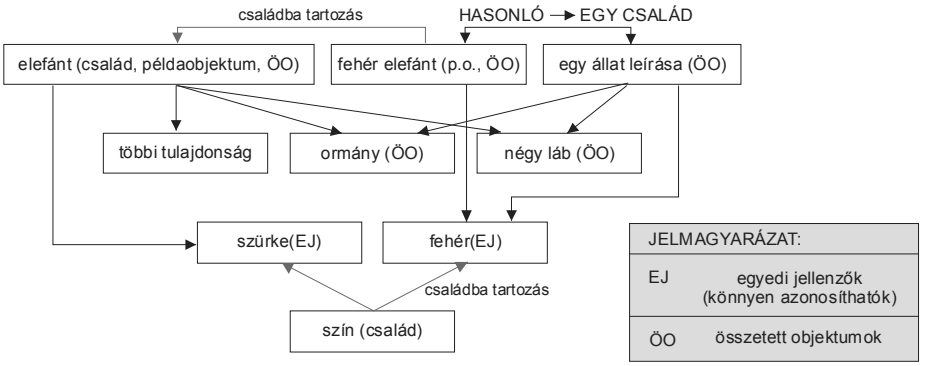


Fig. 2. Egy példa arra, hogyan használhatók az egyedi címkék (színek), és hogy lehet a hasonlóság a családba tartozás alapja (Ismeretlen állat a hozzá leghasonlóbb prototípus családjába tartozik.).

Ahhoz hogy ez a címkézett gráf értelmezhető legyen a számítógép számára szükség van bizonyos szabályokra. A szabályok speciálisan címkézett élekre vonatkoznak. A fogalmi rendszerhez az öröklődést jelölő szabály kapcsolódik legszorosabban. Ez a gyakorlatban annyit jelent, hogy egy objektum egy speciális címkéjű éllel kapcsolódik a prototípusához, amit a program külön kezel ((**Fig. 2.)** fehér „elefánt" és „ fehér elefánt" esete).

Létrehoztam még néhány szabályt, amik az előbbihez hasonló értelmezési szabályok, de már szorosabban kapcsolódnak a megvalósított beszélgető-robot alkalmazáshoz. A szabályokkal és az egyedi fogalmakkal együtt a fogalmi gráf már jelentéssel bír a program számára, befolyásolja annak működését. Ezek a szabályok tulajdonképpen egy minimális kiindulási tudást jelentenek. Ilyen kiindulási tudás, hogy az egyes eseményeknek lehetnek feltételeik, az események egymás után következnek illetve a lezajlott események képesek módosítani a világot. A kiegészítésekkel már megadható egy egyszerű működési modell a beszélgető-robothoz:

1. Az input érzékelése.
2. Az input értelmezése. A kapott mondathoz hozzárendelünk egy a világmodell alapján lehetséges kijelentést.
3. A következmények mérlegelése. A kijelentésnek megfelelően változtatjuk a világmodellt és a felhasználó szándékairól alkotott képet.
4. A világ és a beszélgetőtárs modellje alapján kiválasztjuk az értelmes szándékainkat.
5. Kiválasztjuk a szándékainknak és a világmodellnek megfelelő kijelentéseket.
6. Ezek közül választunk egyet.
7. A kiválasztott kijelentés végrehajtása.
    a. A világmodell frissítése (következmények)
    b. A kijelentés szöveggé alakítása és kiírása

VISSZA AZ ELEJÉRE

Röviden szólnék még arról, hogyan kapcsolódik a konkrét szöveg és a fogalmi hálózat egymáshoz. A kapcsolat úgy valósult meg, hogy a fogalmakhoz mintákat definiálok, amiket a program megpróbál illeszteni a bemenetre, és ha ez sikerül, akkor a fogalom aktivizálódik. Ezek után a rendszer az aktivizált fogalmakat megpróbálja egységgé kovácsolni és egy kijelentésként értelmezni. A minták az ember érzékleteinek feleltethetők meg, természetesen egyszerűsített módon. Egy minta legegyszerűbb esetben egy szó, vagy egy szövegrészlet, de lehet összetettebb esetben bármilyen 0-1 kimenettel rendelkező rutin.

3 Összefoglalás, értékelés

Bemutattam egy arisztotelészitől eltérő fogalmi reprezentációt, ami közelebb áll a természetes emberi gondolkodáshoz, ebből kifolyólag alkalmasabb természetes nyelvű szövegek reprezentálására. A modell működőképességét bizonyítandó bemutattam annak konkrét alkalmazását egy beszélgető-robot alkalmazásban.

A vázolt alkalmazás egyik nagy problémája, a tudásbázis feltöltése. Ebből kifolyólag érdekes a „beszélgetve tanulás”, vagyis az, hogy a program a felhasználóval folytatott beszélgetésekből tanuljon.

Másik lehetséges kutatási irány a fogalmi gráf és a szöveg kapcsolatának részletesebb vizsgálata. Ide tartozik többek között a kontextus kezelése, vagy a metaforák és hivatkozások értelmezése.

Bibliográfia

1. Antero Taivalsaari: Classes vs. Prototypes, Some Philosophical and Historical Observetions, Nokia Research Center, P.O. Boksz 45, 00211 Helsinki (1996)
2. Cougaar.org, BBN Technologies: Cougaar Architecture Document 11.4 (2004)
3. Németh Bottyán: Természetes nyelvű szövegek elemzése szövegkönyvek segítségével, BME MIT, TDK, Budapest (2004)
4. Wallace, Dr. Rich (2002): The Anatomy of A.L.I.C.E.; A.L.I.C.E. Artificial Intelligence Foundation, Inc. 2005, http://www.alicebot.org/documentation/
5. Traum, David R., Schubert, Lenhart K., Massimo Poesio, Nathaniel G. Martin, Marc Light, Chung Hee Hwang, Peter Heeman, George Ferguson, James F. Allen: Knowledge Representation int he TRAINS-93 Conversation System, The University of Rochester, Computer Sience Department, New York, Technical Reaport 663
6. Loebner Prize Home Page (2005), http://www.loebner.net/Prizef/loebner-prize.html

Az alacsony szintű beszédfelismerés mesterséges feljavítása magasabb szintű modellellenőrzéshez

Németh András[1,2], Balázs László[1], Gyepesi György[1]

[1] Alkalmazott Logikai Laboratórium,
Hankóczy J. u. 7. 1022 Budapest,
{xandrew, bazsi, ggyepesi}@all.hu
[2] Budapesti Műszaki és Gazdaságtudományi Egyetem
Számítástudományi és Információelméleti Tanszék,
Magyar tudósok körútja 2. 1117 Budapest,
xandrew@cs.bme.hu

Kivonat: Jelen cikkben ismertetünk egy módszert, mellyel a HMM+GM felépítésű klasszikus beszédfelismerő rendszer teljesítményét feljavítjuk a rendelkezésre álló szöveges átirat segítségével. A feljavítás mértéke az eredeti rendszer teljesítményétől a szinte tökéletes fonéma felismerés szintjéig folytonosan változtatható. Így mérhetővé válik, hogy a különböző, a fonéma felismerő rétegre épített magasabb szintű modellek milyen alacsony szintű hiba esetén milyen teljesítményt nyújtanak. Az itt ismertetett kutatásban a hibatűrő szókeresés és a fonéma n-gramm modellek viselkedését vizsgáltuk meg.

1 Bevezetés

A beszédfelismerő rendszerek jellegzetes felépítési módja, hogy az emberi beszédet több, egymásra épített statisztikai modellel jellemzi, majd az így kapott teljes beszédmodell segítségével történik a felismerés. A legalsó szintet a vizsgált nyelv fonémáinak akusztikus modelljei alkotják. Általánosan elfogadott a beszédfelismeréssel foglalkozók körében, hogy tisztán fonéma szinten jó minőségű beszédfelismerés nem készíthető (erre különben az ember sem képes). Ezért a fonéma szint fölé a konkrét feladattól függően különböző magasabb szintű nyelvi modelleket, pl. fonéma n-gramm statisztikákat, szótárakat, nyelvtanokat illesztenek.

A különböző megoldások hatékonyságát a teljes összetett modell hatékonyságával szokás mérni. Arra ugyan van lehetőség, hogy csak az alacsonyabb szintű modellek használatával végezzünk méréseket, de a magasabb szintű modellek önálló értékelése a klasszikus validációs módszerekkel nem megoldott.

Jelen cikkben ismertetünk egy módszert, melyben a magasabb szintű modellek értékelése az alacsony szint teljesítményének függvényében történik. Így a különböző, magasabb szintű modellekre vonatkozó kísérletek eredményei összehasonlíthatóvá válnak akkor is, ha más fonéma szintű modellel dolgoznak. A beszédfelismerési technológiáknál minden esetben nagyon fontos a felhasználási területhez igazítás, így

a legalkalmasabb magas szintű modell is alkalmazásfüggő. Ha a különböző magas szintű modellekhez a fenti típusú értékelő függvények a rendelkezésünkre állnak, akkor megalapozottabban választhatjuk ki az adott körülmények között (az adott helyzetben elérhető fonéma felismerési hiba ismeretében) a legalkalmasabb magasabb szintű modelleket.

2 A HMM+GM beszédfelismerő architektúra és feljavítása

A felismerés bemenetét egy feature vektor sorozat adja, melynek minden eleme a beszédjel egy kis szakaszának valamilyen akusztikai jellemzőit tartalmazza. A fonémák akusztikai modellje a legtöbb esetben egy Hidden Markov modell, melynek állapotaihoz a feature vektorok valamilyen folytonos eloszlását rendeljük. Leggyakrabban Gauss-mixture típusú eloszlásokat használunk.

A rossz minőségű modellek esetében is nagyon jól működő Viterbi align eljárással minden egyes feature vektorra megmondható, hogy ott a modell mely állapotban van a legnagyobb valószínűségű, az ismert elhangzó szövegnek megfelelő trellis mentén.

Generáljunk egy új feature vektor sorozatot úgy, hogy minden pozícióban az align szerint ott található állapot eloszlásából „húzunk” egy vektort, azaz véletlenszerűen válasszunk az adott eloszlásnak megfelelően. Ezzel lényegében azt érjük el, hogy a kapott feature vektor olyan lesz, mintha a modellünknek tökéletesen megfelelő beszédből nyertük volna ki. Ha ezt a feature file-t adjuk a felismerő bemenetére, akkor érthető módon nagyon jó, méréseink szerint kevesebb mint 5%-os fonéma hibájú felismerést tapasztalunk.

Ha az így kapott feljavított feature file és az eredeti feature file különböző együtthatós, egy összegű lineáris kombinációját tesszük a felismerő bemenetére, akkor a hiba az együtthatók függvényében folytonosan változtatható a kiindulási hibaarány és a szinte tökéletes felismerés között. Az így kapott állítható minőségű alacsony szintű felismerésre ráépítve a vizsgálni kívánt magasabb szintű modellt megkapjuk a bevezetésben előrebocsátott tulajdonságú értékelő függvényt.

3 Hibatűrő keresés és az N-gramm modellek teljesítménye a fonéma hiba függvényében

A hibatűrő keresés a keresett szó és a felismert fonémasorozat edit distance távolsága ill. a felismert sorozatban előforduló fonéma-trigrammok halmazának és a keresett szó trigramm halmazának összehasonlításával történik ([2]). A két módszer nagyon hasonló eredményt ad, és mindkettő esetében a teljesítmény közel lineárisan függ a fonéma hibától.

Az n-gramm modellek esetében egy nagy nyelvi korpuszból (pl. [1]) kinyert fonéma trigramm előfordulási valószínűségek segítségével irányítjuk a felismerést a nyelvre jellemző hangsorozatok preferálása felé. (Lásd pl. [3]).

Ebben az összefoglalóban példaként megadjuk a keresési teljesítmény függését a fonéma hibától. A teljes cikk mindkét modell részletes leírását és értékelését tartalmazni fogja.

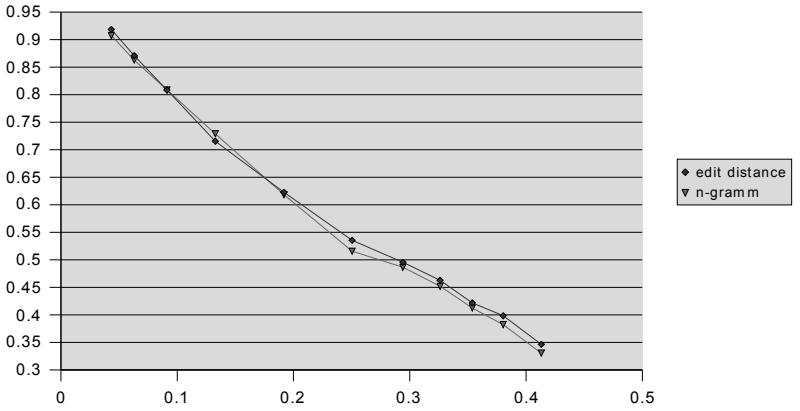


Fig. 1. Az edit distance alapú és a trigramm alapú keresés teljesítménye f-measure-ben a fonéma felismerési hiba függvényében

Bibliográfia

1. P. Halácsy, A. Kornai, L. Németh, A. Rung, I. Szakadát, V. Trón: Creating open language resources for Hungarian. In Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004), 2004
2. K. Ng: Subword-based Approaches for Spoken Document Retrieval Ph.D. Thesis,MIT, 2000
3. C. D. Manning, H. Schtze (ed.): Foundations of Statistical Natural Language Processing. MIT Press, 1999

A KAPU tartalomelemző program narratív pszichológiai alkalmazásának lehetőségei és a program bemutatása

Puskás László[1], Karsai Barna[2]

[1] Pécsi Tudományegyetem Bölcsészettudományi Kara, Pszichológia Doktori Iskola, laszlopuskas@gmail.com
[2] Eötvös Loránd Tudományegyetem Természettudományi Kara, programozó matematikus szak, karsaib@gmail.com

Kivonat: A KAPU tartalomelemző program az első olyan számítógépes elemző rendszer, mely a lejegyzett szöveg jellemzőinek vizsgálatán túl, azzal párhuzamosan, az elhangzott beszédszakaszok hangtani sajátosságait is elemzi. Ez olyan új lehetőséget adhat a narratív pszichológiai tartalomelemzéssel foglalkozó szakemberek számára, mely bővíti, pontosítja, sőt bizonyos értelemben át is értelmezi a tartalomelemzésről kialakult hagyományos gondolkodásmódot. A szöveg akusztikus paraméterei közül a következők vizsgálatára hagyatkozunk a hagyományos tartalomelemző programok által vizsgált jellemzők mellett: időparaméter; dallam; dinamika; az adott szöveg statisztikája; szünetek száma és hossza; fonációs szakaszok száma; frekvenciacsúcsok száma; az elbeszélő hangjának jellemzői.

1 Bevezetés

Poszterünk célja, hogy bemutassuk a 2005 óta fejlesztett programmal kapcsolatos terveinket, elképzeléseinket és az eddig elért eredményeket. Reményeink szerint hamarosan magát a működő programot is be tudjuk mutatni a szakmai közönség számára. Nemcsak a program működését, kezelői felületét, de a program fejlesztői környezetét is be kívánjuk mutatni a konferencián.

A program WinXP (és annak kompatíbilis verziói) és Linux operációs környezetben fut. A KAPU hangelemző program képernyő- és kezelői funkciói és különböző megjelenítési funkciói java Swing technológiával megvalósított GUI felületre integráltak.

2 A hangok kezelése

A hangelemző modul hagyományos WAV fájl-formátumot használ, amely lehetővé teszi hang-fájlok lejátszását és felvételét már a program első verziójában is. Jelenleg kétcsatornás, sztereo hangformátumot használunk:

1. Wav formátum
2. Más tömörített formátumban való tárolás a jövőben (mp3)

3 Alkalmazói modulok

Az alkalmazói modulok több funkcióból és alkalmazási ablakból állnak:

1. A hangvezérlő elemek
2. Akusztikai modul (Spectrum, Signal, Megjelenítések)
3. A fonetikai modul
4. Szövegmodul
5. Tartalmi modul
6. Elemek

4 Program modulok

A módszerek és algoritmusok egy külön programblokkja gondoskodik az alkalmazás kompatibilitásáról más programokkal, alkalmazásokkal.

1. Média: felvétel, lejátszás
2. Digitális hangfeldolgozás: szűrés, spektrum-elemező, hangelkülönítő csatornák
3. Tartalomelemző: Összekapcsolja a szöveget és fonetikai paramétereket. Kezeli a szövegadatbázist.
4. Időzítő: az alkalmazás szinkronizálása modul összehangolja a vizsgált tartomány alapján a modulokat (spektrum, fonetika stb.)

5 Digitális hangfeldolgozás

1. A digitális hangfeldolgozás alapfunkciói lehetővé teszik a gyakorlati alkalmazást.
2. Szűrő: a beszédhangokat elkülönítése a zörejektől.
3. A különböző beszélők hangjának elkülönítése.
4. A személyek hangjának spektrum-elemzése.

6 A program szótárának adattárolása és a szöveg fonetikai jelekké történő átalakításának algoritmusa [1]

Bibliográfia

Olaszy G. - Németh G. - Olaszi P. - Kiss G. - Zainkó Cs. - Gordos G.: Profivox - a Hungarian TTS System for Telecommunications Applications. International Journal of Speech Technology. Vol 3-4. Kluwer Academic Publishers. 2000. 201-215.

A HunNER Korpusz

Simon Eszter[1], Farkas Richárd[2], Halácsy Péter[1],
Sass Bálint[3], Szarvas György[2], Varga Dániel[1]

[1] BME MOKK
{daniel, halacsy}@mokk.bme.hu, esimon@cogsci.bme.hu
[2] Szegedi Tudományegyetem, Informatika Tanszékcsoport
{szarvas, farkas}@inf.u-szeged.hu
[3] MTA Nyelvtudományi Intézet
joker@nytud.hu

Kivonat: Cikkünkben egy folyamatban lévő projektet mutatunk be, melynek célja egy nagyméretű, manuálisan tulajdonnév-annotált korpusz létrehozása. A tervezett korpusz jól használható lesz gépi tanuláson alapuló tulajdonnév-címkézők tanítására és szabványos kiértékelésére, miközben elő- és utófeldolgozó eszközöktől független. A projektet a BME MOKK, a Nyelvtudományi Intézet és az SZTE közösen indította. A projekt fontos mellékterméke egy olyan klasszifikációs útmutató magyar nyelvre, amely időtálló, és a fenti intézmények közötti konszenzuson alapul. Az elkészült korpusz a konzorcium döntése alapján szabadon hozzáférhető lesz kutatási célokra.

1 Annotációs séma és útmutató

A projekt egyik fontos célja kialakítani egy egységes annotációs útmutatót. A konzorciumi tagok által eddig használt útmutatók között lényeges eltérések vannak. Ezek szabályait közös munkával konzisztens rendszerré ötvöztük.

Az annotálás során mindig szem előtt tartandó elveink a következők:

- Névnek nevezzük azt a kifejezést, ami unikusan, vagyis egyedi módon referál a világ valamely entitására. Tehát nem annotálunk olyan frázisokat, amelyek ugyan a világnak valamely egyedi részére utalnak, de nem teljesen egyértelmű módon.
- Nem annotálunk egymást átfedő vagy egymásba ágyazott neveket. Vagyis minden annotációnak be kell fejeződnie, mielőtt egy másik elkezdődik.
- Mivel a nevek nem kompozicionálisak, tehát jelöletük nem a részeik jelöletéből áll össze, ezért a neveket nem bonthatjuk részekre az annotáláskor. Például a *Kossuth Lajos utca* egy névként jelölendő, hiába van benne egy személynév. Mindig a leghosszabb nevet (a legkülsőbbet) jelöljük a jelölhetők közül.
- Az inflexiókról hagyományosan azt szoktuk gondolni, hogy nem vagy csak elhanyagolhatóan minimális mértékben változtatják meg a tulajdonnév "jelentését", vagyis ugyanarra utalnak, mint a toldalék nélküli alakok. Ezért ha az azonosított tulajdonnév ragozott formában szerepel a szövegben, a raggal

együtt, a teljes alakot annotáljuk. A képzők közül viszont csak néhányról gondoljuk ezt, ezért a képzett alakokat nem jelöljük, kivéve a földrajzi névből *-i/-beli* képzőkkel képzett melléknéveket.

A leginkább vitatott kérdéseink megegyeznek a tulajdonnév-klasszifikáció kapcsán nemzetközi szinten is felmerülő kérdésekkel. Ilyen például a *tag-for-meaning* elve, melyet követve a tulajdonneveket aktuális szövegbeli kontextusuk alapján osztályozzuk. Ezzel kapcsolatban problémák elsősorban a metonímiák esetében állnak elő, amikor valamely típusú entitással egy másik típusú entitásra utalunk; illetve az olyan neveknél, amelyek több dologra: épületre, emberi közösségre és intézményre is utalhatnak, mint a múzeumok, iskolák vagy színházak. A nevek metonimikus használatára a legjellemzőbb példát a szervezetre referáló helynevek, illetve a helyre referáló szervezetnevek adják. Az *A János kórházban sok a macska* példamondatban a *János kórház* egy intézménynek a neve ugyan, de ebben a kontextusban helyet jelöl. Ugyanígy: a *Washington Moszkvával tárgyal* mondatban mindkét helynév valamilyen szervezetre referál. A tag-for-meaning elvet követve a *János kórház*at helynévként, *Washington*t és *Moszkvá*t pedig szervezetnévként kellene annotálnunk. Egy másik lehetséges annotálási mód szerint egy típusú nevet kontextustól függetlenül, eredeti jelöletének megfelelően kell jelölni (*tag-for-tagging*). Ennek a konfliktusnak a kiküszöbölésére bevezettünk két új alkategóriát, melyekkel jelölni tudjuk a metonimikus használatot, és egyben lehetővé tesszük minden újrafelhasználó számára a neki tetsző elv követését.

A kialakításnál figyelembe vettük azt a szempontot is, hogy az általunk használt annotációs séma kompatibilis legyen nemzetközileg elfogadott tulajdonnév-klasszifikáló sémákkal. Ezek közül a számunkra legfontosabbak a Szeged NER korpusz [1] építéséhez már adaptált CoNLL [2,3], valamint a Linguistic Data Consortium által alkalmazott [4] sémák. Ezek alapján a korpuszunkban jelölendő típusok:

a személynevek (PERSON),
az embereknek valamely szervezetnél betöltött szerepét jelölő frázisok (ROLE),
a cím- és rangjelölő szavak (RANK),
a szervezetnevek (ORGANIZATION),
a helynevek (LOCATION),
a szervezetre referáló helynevek (ORG:LOC),
a helyre referáló szervezetnevek (LOC:ORG),
a márkanevek (BRAND/PRODUCT),
a műcímek (TITLE) és
egyéb tulajdonnevek, vagyis amelyek nem tartoznak a fenti kategóriák egyikébe sem, de tulajdonnevek (MISC).

2 Külső annotáció

A korpusz minden feldolgozási lépésében (tokenizálás, mondatra bontás, tulajdonnév-címkézés) külső (*standoff*) annotációt használunk. Ennek lényege, hogy az eredeti dokumentumokat sima szövegként rögzítjük, és az annotációkat nem beágyazott markupként, hanem egy külső fájlban jelöljük úgy, hogy megadjuk, hogy az eredeti szöveg melyik karaktertartományára vonatkozik a címkézés, és hogy milyen címkét

kap a szövegrészlet. A külső annotáció előnye, hogy az annotálást teljesen különválasztja a használt feldolgozó eszközöktől, és minden formai információ elérhető a feldolgozottság minden fázisában.

Páros számú annotátorral dolgozunk, és az annotátorok közötti egyetértést mérjük:

$$\frac{2 * |\text{ugyanúgy jelölt entitások}|}{|\text{A annotátor által jelölt entitások}| + |\text{B annotátor által jelölt entitások}|}$$

Az annotátorok mindig csak a rögzített útmutató alapján dolgozhatnak, amit a menet közben felmerülő problémás kérdések megvitatásával folyamatosan fejlesztünk, egész addig, amíg az egyetértés 95% feletti nem lesz.

3 A korpusz forrásai

A korpusz méretének lehetővé kell tennie, hogy az azon tanított statisztikus tulajdonnév-felismerő modellek általános szövegen is megállják a helyüket, és specifikus szövegen is jól tudjanak működni. Számításaink szerint legkevesebb félmillió szövegszó címkézése teszi lehetővé, hogy a korpuszban megfelelő méretű részkorpuszok legyenek.

A korpusz elsődleges forrása magyar nyelvű valódi hírek teljes szövege. A korpusz téma szerinti eloszlása a következő: gazdaság (100 ezer szövegszó), sport, belföldi politika, nemzetközi politika, törvények/rendeletek, tudomány/technika, fórum/blog, szoftverkézikönyvek, filmszövegek/szépirodalom (50-50 ezer). A gépi fordítási alkalmazásokra tekintettel a szövegeket úgy választjuk ki, hogy minden műfajban legalább 1/5 rész angolból magyarra fordított szöveg legyen.

4 Jogok

Alapvető cél, hogy a létrejött korpuszt bárki teljesen szabadon használhassa, és kiegészíthesse további standoff annotációkkal.

Bibliográfia

1. György Szarvas, Richárd Farkas, László Felföldi, András Kocsor, János Csirik: A highly accurate Named Entity corpus for Hungarian. Proceedings of International Conference on Language Resources and Evaluation 2006.
2. Nancy Chinchor, Erica Brown, Lisa Ferro, Patty Robinson: Named Entity Recognition Task Definition. MITRE and SAIC. 1999.

3. Erik F. Tjong Kim Sang, Fien De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003, pp. 142-147.
4. http://projects.ldc.upenn.edu/LCTL/Specifications/SimpleNamedEntityGuidelinesV6.5.pdf

MEO ontológiamodell

Szakadát István[1], Szőts Miklós[2], Gyepesi György[2], Varasdi Károly[4], Ungváry Rudolf[3], Simonyi András[1,2], Gyarmathy Zsófia[4], Szaszkó Sándor[5], Szeredi Dániel[1,4]

[1] BME SZKT MOKK, Budapest Stoczek utca 2.
i@syi.hu, dani@szeredi.hu, andras.simonyi@gmail.com
[2] ALL, Budapest Hankóczy u. 7.
{szots,gyepesi}@all.hu
[3] OSZK, Budapest Budavári Palota,
rudi@hungary.com
[4] MTA NYTI, Budapest Benczur u. 33.
varasdi@nytud.hu, gyzsof@budling.nytud.hu
[5] BME TMIT, Budapest Magyar tudósok krt. 2.
szaszko@tmit.bme.hu

Kivonat Cikkünkben először röviden értékeljük a MEO ontológia-infrastruktúra építő projekt eredményeit, majd bemutatjuk azt a keretrendszert, ontológia modellt, amely – saját ontológiai elkötelezettségeink szerint – mindenféle más ontológiaépítő munka alapjául, keretéül szolgálhat. Ennek során foglalkozunk a tárgy- és metaszintű fogalmak, illetve a nyelvi és fogalmi réteg elkülönítésével. [6]

Kulcsszavak: ontológia, fogalom, reláció, osztály, attribútum, metafogalom, modell, fogalmi modell, nyelvi modell

1. Bevezetés

A Magyar Egységes Ontológia (MEO) projekt 2004 decemberében indult és 2006 végén zárul le. A projekt célja között szerepelt egy csúcsontológia és egy szakontológia felépítése, egy szakterületi ontológia támogatásával működő alkalmazás kifejlesztése, illetve az ontológiák építésére, menedzselésére alkalmas ontológia-infrastruktúra felépítése.

A projekt lezárultával egy demoalkalmazás segítségével szemléltetni tudjuk, hogy miként lehet hasznosítani az ontológiai tudást a közönségszolgálati tevékenység támogatására a távközlés területén, de a projekt legfontosabb célja mégsem az volt, hogy az ontológiai tudás gyakorlati hasznosításának lehetőségét bizonyítsa. Sokkal inkább arra törekedtük, hogy az ontológiaépítés elméleti és gyakorlati teendőit felfedezzük, leírjuk és elérhetővé tegyük más (későbbi) projektek számára. A MEO-projekt eredményeivel kapcsolatban ki kell emelnünk,

[6] A MEO-projekt a KPI NKFP 2/042/04. sz. támogatásával jött létre.

hogy – a nyílt forráskódú kezdeményezések mintájára, azok általánosításaként létrejött Creative Commons mozgalom szellemében – a projekt eredményeinek döntő hányada szabadon hozzáférhető. Úgy gondoljuk, hogy a számítógépek számára biztosítandó szemantikai, ontológiai tudásbázist – anyanyelvünkhöz hasonlóan – csak hosszútávon és közös, egymáshoz igazodó erőfeszítések eredményeként tudjuk csak felépíteni és fenntartani. A MEO ehhez a közös munkához próbálta meg letenni az alapokat.

Az ontológiai tudást a konkrét alkalmazásokban szakontológiák segítségével lehet igazán hasznosítani. A MEO-projekt mégis – története során mindvégig – a csúcsontológia felépítésére koncentrált. Bár a csúcsontológiáknak is lehet gyakorlati haszna akkor, amikor egymástól elkülönülten fejlesztett informatikai rendszereket kell kooperációra késztetni, ám figyelemkoncentrációnk és fókuszunk a csúcsontológia irányába inkább annak a szándéknak tulajdonítható, hogy ,,ezen a terepen" lehetett a legjobban megtanulni, megérteni és megoldani, kezelni remélni az ontológiaépítés problémáit.

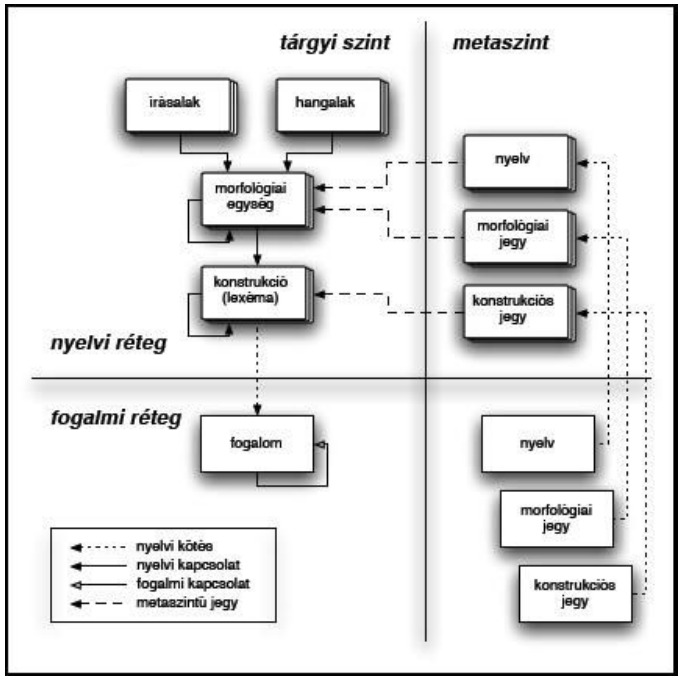
Az ontológia a világ dolgairól szóló tudásunkat tárolja nyelvfüggetlen formában. A MEO projekt kiemelt célja volt az ontológiák tényleges nyelvfüggetlenségének biztosítása, ezért az ontológia modellezése során elkülönítettük a nyelvi és fogalmi rétegeket egymástól. A következőkben röviden bemutatjuk a két réteg szintjeit, elemeit, kapcsolatait.

2. A nyelvi réteg

Mégha az ontológiát nyelvfüggetlennek gondoljuk is, beszélni róla (a rendszer egészéről, a részeiről, az elemeiről) csak valamilyen nyelv segítségével tudunk. Azért van mindenképpen szükség nyelvi kötések létrehozására, hogy az ontológiát használó emberek is beszélni tudjanak róla. A MEO-projekt számára olyan modellt terveztünk, amelyben egyrészt egyértelműen el lehet különíteni a fogalmi és nyelvi rétegeket egymástól (persze mindezt úgy, hogy eközben a fogalmakhoz hozzákötjük az adott nyelven a fogalmakra mutató kifejezéseket, lexikai egységeket), másrészt tetszőleges (számú és típusú) nyelvet lehet a fogalmi réteghez kapcsolni. A nyelvi réteg entitásait mutatja az 1. ábra.

A nyelvi rétegben három – egymásra épülő, de egymától jól elkülöníthető – entitást definiáltunk. A ,,legalsó szinten" van a *szóalak* (ami lehet írás- vagy hangalak). A szóalak még nem igazán nyelvfüggő entitás, amire azért van szükség, mert a korpuszokból számolt gyakorisági értéket csak ehhez tudjuk hozzárendelni. A szóalakhoz valamilyen morfológiai jegyhalmazt rendelve jutunk a *morfológiai egység* fogalmához. Ez a nyelvi entitás már egyértelműen nyelvfüggő, és a hozzá kapcsolt morfológiai jegyek révén lehet megérteni a szóalakok nyelvi megnyilatkozáson belüli viselkedését. Meg kell jegyezzük azonban, hogy a nyelvi réteg első két entitása, a szóalak és a morfológiai egység még nem igazán használható szemantikai, ontológiai célokra.[7] A nyelvi réteg harmadik entitása, a *konstrukció* az, ami kapcsolatot teremt a nyelvi és fogalmi réteg (vagyis a morfológiai és

[7] A projekt ,,melléktermékeként" ugyan belekezdtünk egy nyelvi ontológia felépítésébe, de ez nem a MEO közvetlen céljainak eléréséhez volt szükséges.



1. ábra. A MEO modell nyelvi rétege

szintaktikai viselkedés, illetve a jelentés) között. Ha a jelentést az ontológiai réteg hordozza, akkor a konstrukció az, ami a nyelvi megnyilatkozásokat összeköti a szemantikával. A konstrukció gyakran lexikai egység (lexéma), de az ontológiában leírhatjuk más nyelvi egységek (pl. a toldalékok) jelentését is, és az ilyen típusú konstrukciókon keresztül kapcsolatot teremthetünk más típusú morfológiai egységekhez is. A konstrukció természetesen a morfológiai egység szintjéről örökli annak nyelvi kapcsolatát, tehát nyelvfüggő elem.

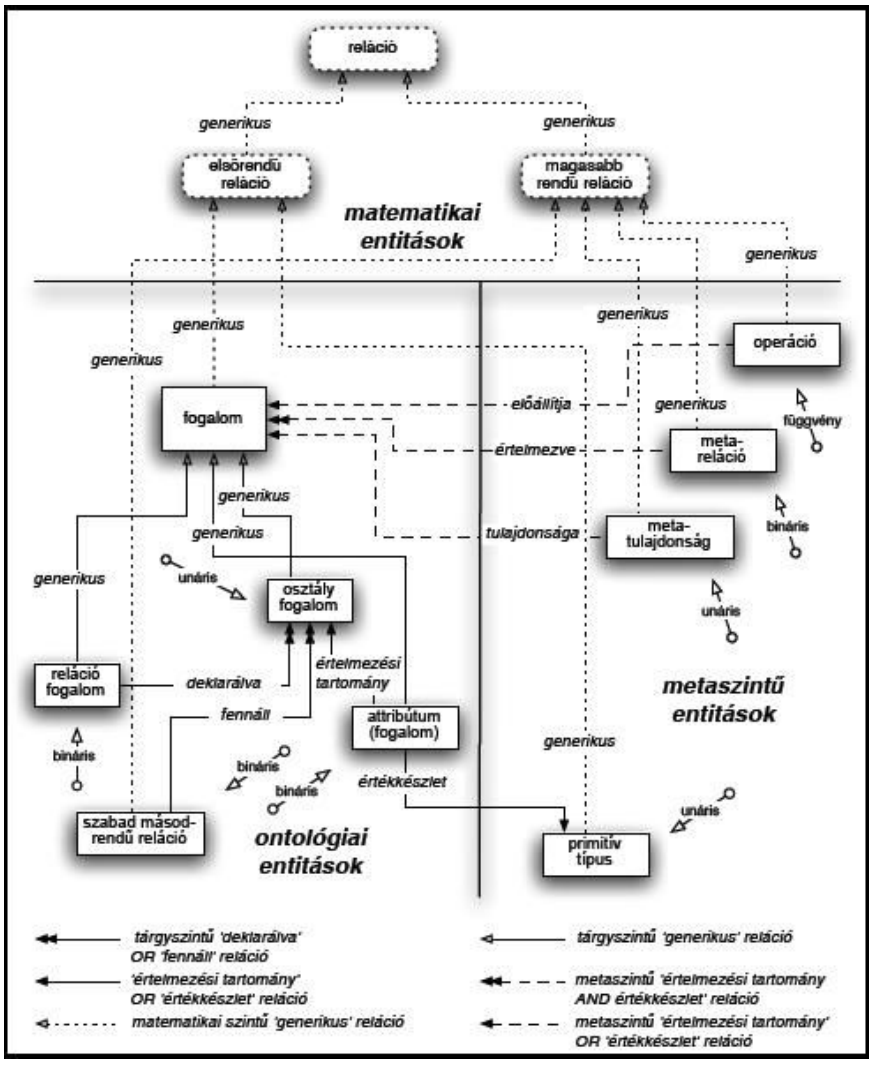
Mivel a konstrukció a nyelvi réteg eleme, ami a kapcsolt fogalmon keresztül kap jelentést, tetszőleges számú nyelven kifejezett konstrukciót lehet ugyanahhoz a fogalomhoz rendelni, ami biztosítja azt, hogy az ontológia elemeihez akárhány nyelven lehessen nyelvi megnyilatkozásokat fűzni. Az ontológia- és szótárépítési munkát ezáltal egymástól elválasztva, de egymáshoz mégis illeszkedve lehet végezni. Ennek során csak arra a szabályra kell figyelni, hogy minden új fogalomhoz biztosítani kell egy alapértelmezett nyelvi konstrukciót, így, ha valamely nyelven nincs egy ontológiai egységnek megfelelő nyelvi konstrukció, akkor még megjeleníttethető az ontológiai rétegben rögzített információ. Ilyen esetekben a ,,nyelvi lyukakat" az ,,idegen nyelven" megjelenő konstrukciók felbukkanása jelzi.

Bár széles körben hangoztatják, hogy az ontológiákba foglalt tudás nyelvfüggetlen, nem nagyon tudunk olyan ontológiaépítő projektről, illetve olyan ontológiaszerkesztő alkalmazásról, amely hatékonyan támogatná a nyelvfüggetlen építkezést. Ezért döntöttünk a projekt félidejében úgy, hogy saját szerkesztő fejlesztésébe kezdünk

(MEOdit), mert biztosítani szerettük volna, hogy a MEO csúcsontológia elemeihez valóban több nyelven lehessen nyelvi szótárt illeszteni (jelenleg négy nyelven, magyarul, angolul, latinul és lengyelül áll rendelkezésre a MEO csúcsfogalmak durván 2700 elemű készletének nyelvi kötése).

3. A fogalmi réteg

A MEO ontológiamodell másik fontos része a fogalmi réteg, mely az ontológiák építéséhez használható, metaszinten rögzített, a tárgyszinten nem változtatható metafogalmakat és a tárgyszinten az ontológiaépítők által szabadon építhető fogalmakat tartalmazza. A fogalmi réteg vázlatos ábrája a következő:



2. ábra. A MEO modell fogalmi rétege

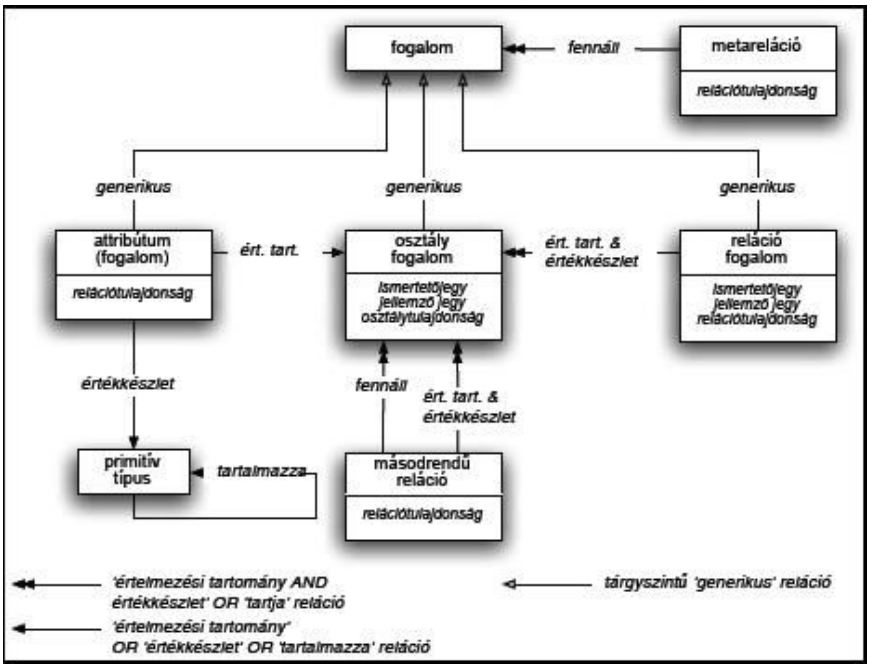
A fogalmi réteg felépítésekor arra törekedtünk, hogy egyértelműen el tudjuk választani a tárgy- és a metaszintű (illetve meta-metaszintű) entitásokat egymástól. A tárgyszint (és persze az egész ontológia) központi entitása a *fogalom*, amelynek három típusát, az *osztály-*, a *reláció-* és az *attribútumfogalmakat* különítettük el (szemben azzal a szokásos megoldással, ami csak az osztályfogalmakat veszi figyelembe). A relációfogalmon elsőrendű relációkat értünk, amelyeket az ontológia építői az osztályfogalmakon deklarálhatnak. Emellett mód van még másodrendű relációk felvételére is. Az attribútumfogalom is relációnak minősíthető, a különbség csak annyi közöttük, hogy az attribútum értékkészlete a metaszinten rögzített primitív adattípusokból állhat. A fogalom típusait (tehát az osztályt, a relációt és az attribútumot) generikus reláció köti a fogalomhoz. A *generikus alárendeltje* reláció (amire használják az is-a, a szubszumpció, az inklúzió, a részosztály és még sok minden más terminust) már a metaszinthez tartozik, és másodrendű reláció, hiszen fogalmakon (vagy másként: predikátumokon) érvényesített univerzális kvantor segítségével lehet definiálni.

A metaszinthez kötöttük az új fogalmak képzésére alkalmazható *műveleteket* (az egyszerű relációszorzással állíthatjuk elő az apja fogalmából a nagyapja fogalmát), illetve ehhez a szinthez rendeltük a fogalmak *metatulajdonságait* (mint például az OntoClean módszertan rigid vagy esszenciális tulajdonságait vagy a relációk algebrai tulajdonságait, mint a szimmetricitás vagy tranzitivitás). A metaszint elemeinek nagyon fontos vonása az, hogy azokat egy adott ontológia építői nem vehetik fel (és nem is törtölhetik) szabadon, azok a ,,rendszerbe vannak égetve". A MEO építkezése során ez a megoldás minősíthető a legerősebb ontológiai elkötelezettségnek. A tárgyszintre ebből az ontológiai elkötelezettségből csak a fogalom alá sorolt legfőbb típusok létezésének elismerése került be. Ettől eltekintve szabadon építhető a tárgyszint fogalmi rendszere.

A metaszinten rögzített metafogalmak meghatározását sem végezhetjük el másként, mint hogy valamilyen relációs definíciót adunk rájuk (a művelet homogén, balról egyértelmű, jobbról totális, kétargumentumú reláció, míg a *tulajdonsága* metafogalom olyan másodrendű reláció, amely egy tetszőleges fogalmat és egy tulajdonságot kapcsol össze a tárgyszinten). Ehhez természetesen újra szükség van a reláció fogalmára, vagyis már meta-meta szinten kell a definíciókat megadni. Erre szolgál a modell harmadik szintje, ahol már ,,matematikai értelemben" használjuk a relációfogalmakat.

A fogalmi réteg a tárgyszint fogalmaira fókuszál, ami az ontológiák föntről lefelé történő (top-down) építését teszi lehetővé. A fogalmi rendszerek kibontásakor azonban nemcsak a fogalmakat szokták megadni, definiálni. Olykor megpróbálják meghatározni a fogalmak (pontosabban az általuk ,,leírt", ,,hivatkozott" létező entitások) tulajdonságait, jegyeit is. Elképzelhető tehát olyan ontológiaépítkezés, amely az alulról fölfele történő (bottom up) építkezési módszertanra támaszkodik. A MEO-ban kezdettől fogva arra törekedtünk, hogy a modellünkben meg tudjuk ragadni ezt a kettőssséget. Ezért egyrészt a projekt egyik munkacsoportjában a tulajdonságok egymásra épülő rendszerének egy lehetséges modelljét próbáltuk meg kidolgozni, másrészt a top-down modellen belül is megteremtettük annak le-

hetőségét, hogy a fogalmakhoz rendelhető ismertetőjegyek rendszerét ki lehessen formálni. Ezt mutatja be a 3. ábra.



3. ábra. A fogalmakhoz rendelhető jegyek rendszere

Az ábrán látható, hogy a fogalmakat leíró tulajdonságok, jegyek változnak a fogalom típusától függően. A MEO-modell alapján megvalósuló ontológiaépítkezési munkák során a jegyeket bármiféle kényszer nélkül lehet csak a fogalmakhoz kapcsolni, tehát – egyelőre – nem lehet a fogalmak közötti relációk és a fogalmakat jellemző jegyek közti összefüggéseket nyomon követni (és kihasználni). A vízió szintjén természetesen megfogalmaztuk azt az Arisztotelészre visszavezethető elképzelést, miszerint csak úgy lehet megadni egy új fogalmat, ha meghatározzuk a fogalmat egyértelműen jellemző, ,,azonosító" tulajdonságot (differentia specifikát), de a tényleges gyakorlatban mindezt – egyelőre – nem tudtuk implementálni.

Az ontológiaépítő munka egyik fontos összetevője az életciklus-menedzsment. Az ontológiák esetében mindez azt jelenti, hogy kezelni kell tudnunk az ontológiák részeit, részrendszereit (vagy másként, az ontológiát gráfként értelmezve, a részgráfokat), az ontológia összetevőire (elemeire, részeire, részrendszereire) vonatkozó egyéni és csoportos jogosítványokat. Az ontológia szerkesztéséhez, a fogalmak definiálásához persze egyfajta konszenzusmenedzsmentre is szükség van, mivel arra kell felkészülni, hogy az ontológiák létrehozását, fenntartását több ember kooperációjától remélhetjük csak, így a közösen épített rendsze-

rek konzisztenciájához biztosítani kell valahogy a vitás kérdések rendezésének lehetőségét. Túl sok választási lehetőségünk persze nincs. Ezt a helyzetet a közösségi döntések egyik típusának minősítve csak az lehet a nyitott kérdés, hogy a közösségi döntéshozatali eljáráshoz milyen konkrét megoldásokat, algoritmusokat választunk magunknak a társadalmi választások elmélete által kínált repertoárból.

Említettük, hogy a tárgyszintre kevés ontológiai elkötelezettséget „vittünk be". Az már a projekt indulása előtt is nyilvánvaló volt számunkra, amit aztán a projekt tényleges beindulásával a saját tapasztalataink is megerősítettek, hogy a tárgyszinten, főleg a csúcsfogalmak esetében alternatív fogalomértelmezésekre kell felkészülni, vagyis nincs esély és remény a teljeskörű konszenzus kialakítására. Minél közelebb (vagyis a hierarchiában minél lejjebb) kerülünk azonban a hétköznapi szinten gyakran használt fogalmakhoz, annál valószínűbb, hogy konszenzust lehet találni ezek meghatározásában, illetve a fogalmak ontológiai struktrúrában elfoglalt helyeinek megállapításában. Ezt a konszenzusmenedzsment egyik különös feladatának minősíthetjük. Természetesen a MEO csúcsontológiában határozott ontológiai elkötelezettség érhető tetten, hiszen nekünk is rögzíteni kellett valahogyan a csúcsfogalmainkat, de azt reméljük, hogy az ontológiaszerkesztés elveinek, fogalomkészletének tisztázásával és publikálásával lehetővé tesszük bárki számára, hogy saját ontológiát tudjon előállítani céljai eléréséhez.

Hivatkozások

1. Corcho, O., Fernández-López, M., Gómez-Pérez, A., Methodologies, tools and languages for building ontologies. Where is their meeting point? In: *Data & Knowledge Engineering*, Vol. 46, 2003, pp.41-64.
2. Green, R., Bean, C.A., Myaeng, S.H., *The Semantics of Relationships: An Interdisciplinary Perspective*, Dordrecht: Kluwer, 2001.
3. Guarino, N., Welty, C., A Formal Ontology of Properties. In: *Proceedings of 12th Int. Conf. on Knowledge Engineering and Knowledge Management Lecture Notes of Computer Science*, Springer Verlag, 2000.
4. Guarino, N., Welty, C., Supporting ontological analysis of taxonomic relationships. In: *Data & Knowledge Engineering* Vol. 39, 2001, pp 51-74.
5. Staab, S., Studer, R. (eds.), *Handbook of Ontologies*, Springer Verlag, 2004.
6. Szakadát I., Szőts, M., Gyepesi, Gy., MEO - Ontology Infrastructure. In: Gabor Magyar, Gabor Knapp, Wita Wojtkowski, Gregory Wojtkowski, Joze Zupancic, Stanislaw Wrycza (eds.) *Advances in Information Systems Development: New Methods and Practice for the Networked Society, Proceedings Information Systems Development*, Springer (megjelenés alatt).
7. http://ontologia.hu/meo

Ontológiaalapú szöveganntotáció a Sintagma projektben

Szekeres András Márk[1], Varga László Zsolt[1], Krauth Péter[2]

[1] MTA SZTAKI,
Budapest, 1111 Lágymányosi u. 10.
{szekeres, laszlo.varga}sztaki.hu
[2] IQSYS Informatikai ZRt.,
Budapest, 1134 Hun u. 2.
krauth.peter@kfki.com

Kivonat: A természetes nyelvű szövegek számítógépes feldolgozása során általános igény, hogy tovább lehessen lépni egyszerű karaktersorozatok felismerésének szintjén a szövegben. A Sintagma-projektben fejlesztett ontológiaalapú szöveganntotáció egyfelől a szöveget struktúráltan ragadja meg, vagyis a rendelkezésre álló karaktersorozatból szintaktikai elemzés segítségével levezetési fákat ill. más hasonló nyelvtani szerkezetet állít elő. Másfelől nemcsak a mondat szerkezetének, hanem a benne szereplő szavaknak is struktúráltabb reprezentálását használja, és ennek eredményeképp szavak mellett ontológiabeli fogalmakkal is képes *indexelni* a szöveget. A fejlesztés újszerű megközelítései közé tartozik a *többértelműségek* feloldására alkalmazott algoritmus és az egyedi példányokra, objektumokra utaló olyan *speciális kifejezések* (pl. rendszám, telefonszám, számlaszám, termékkód) felismerésének és értelmezésének módja, amelyek gyakran fordulnak elő szakszövegekben nehezítve ezzel a szövegek feldolgozását. Ez utóbbit a szöveganntotáció számára a Sintagma *szemantikus információintegráló rendszer* biztosítja, amely fogalmi és példányszintű háttérinformációkkal képes ellátni a szövegfeldolgozást, és növelni annak eredményességét.

1 Bevezetés

Az „ontológia” kifejezést különféle szűkebb-tágabb értelemben szokták használni. A továbbiakban alapvető tulajdonságának azt tekintjük, hogy öröklődést biztosító hierarchikus struktúrával rendelkezik.

A természetes nyelvű szövegekkel foglalkozó projektek különféle célokat tűznek ki: a keresés hatékonyabbá tételétől kezdve a szöveg tartalmának logikai rekonstrukciójáig terjedően meglehetősen széles a paletta. A több éve zajló kutatás azzal a problémával foglalkozik, hogy hogyan kapcsolható össze a szöveg az értelmezéséhez használt ontológiával, vagyis hogy az ontológia fogalmait hogyan lehet felismerni a szövegben. Az utóbbi két évben a kutatásnak a Sintagma-projekt keretein belüli fejlesztések adtak további perspektívát.

A jelen cikk a ragozott szavakból a szótövek előállítását megoldott problémának tekinti, az ezt követő lépésekre fókuszál. A kutatás tárgya ezért a többértelműségek és referenciák feloldása volt, vagyis az, hogy olyan esetekben is beazonosítható legyen egy adott fogalom, ha egy másik szóval hivatkoznak rá.

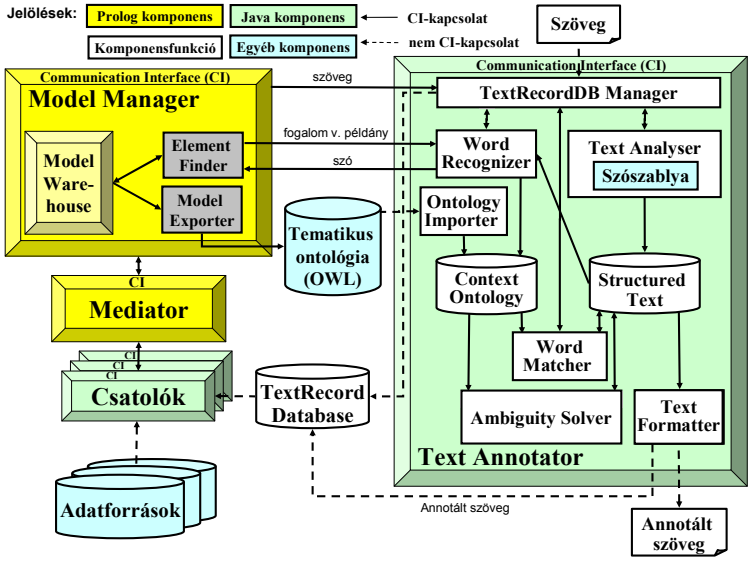


Fig. 1. A Text Annotator felépítése és kapcsolata a szemantikus integrációval.

Egy másik ontológiákhoz kapcsolódó fejlesztésre nagy méretű példány-adatbázisok kapcsán merült fel az igény. Például közéleti szereplők különféle információkkal ellátott nyilvántartásánál nem praktikus egy OWL-ben leírt ontológiába több ezer, esetleg tízezer példányt felvenni: a felismerő algoritmust jelentősen lelassítaná. Ezért az ontológiába a példányok és metainformációik „on-demand” (szükség szerint) dinamikusan töltődnek be a háttérként szolgáló, a Sintagma által integrált adatforrásokból.

Fontos felismerés volt az is, hogy a háttérinformációk adatelemzésével (adatprofilozás) olyan metainformációk (pl. adatmaszkok) származtathatók, amelyek segítik egyes speciális formátumú kifejezések felismerését mint pl. egy rendszám (ABC-123) alapján annak a kikövetkeztetését, hogy valójában valamilyen járműről van szó. Ha van hozzáférés a vonatkozó adatbázishoz, akkor a szóbanforgó jármű (ti. az „ABC-123” rendszámú) egyéb tulajdonságait (pl. tulajdonos, típus) is származtatni lehet.

2 Többértelműségek feloldása

Többértelműségek feloldása alatt sokkal tágabb problémakör értendő, mint ami általában szokás. Nem csupán a többértelmű szavakkal és nyelvtani utalószavak okozta

többértelműségekkel kell foglalkozni, hanem általában azzal a jelenséggel, amikor egy szó egy másik helyett szerepelhet, és így el kell dönteni, hogy melyiket jelenti.

2.1 Többértelműségi problémák

Egy tipikus esete ennek a jelenségnek a következő „Feltettem a rizst főni. Azután leültem a TV elé, és nem vettem észre, hogy odaégett az étel". Itt az „étel" szó a „rizs"-re utal, az emberi olvasó természetesen ezt rögtön megérti. A számítógépes feldolgozás során viszont minden ma létező megoldás, amikor megtalálja az „étel" szót a szövegben, majd az ontológiában/szótárban, akkor itt megáll, és fel sem merül, hogy kiderítse, vajon esetleg hivatkozik-e egy már korábban a szövegben megjelenő dologra?

Ez a jelenség igen gyakori a szövegekben, különösen újságcikkek esetén, ahol általában a szóismétlést súlyos stilisztikai hibának tekintik. Az újságcikkek esetében a mondatok többségében szerepel legalább egy ilyen fajta referencia. De formális szövegekben is előfordul, például orvosi zárójelentésekben: pl. egy fülvizsgálat után az orvos egyszerűen „nyálkahártyát" ír a „dobüregi nyálkahártya" helyett.

2.2 A megoldás fő elve

A probléma feloldása két lépésből áll, először generálni kell a jelentésjelölteket, másodszor pedig dönteni kell közöttük.

A jelöltgenerálást úgy történik, hogy az ontológiában megtalált fogalom leszármazottait tekinti jelölteknek. Ugyanis leggyakrabban a szóismétlés elkerülése érdekében használt szavak az eredeti szó *általánosabb kategóriái* (mint például a „rizs" helyett használt „étel" kifejezés). Informális szövegekben előfordulnak *tulajdonságokon alapuló referenciák* is: például egy jelenet szereplői közül úgy hivatkoznak az egyikre, mint „a kövérebbikre". Jelenleg ezzel az esettel az algoritmus nem foglalkozik, de éppen a már említett háttérinformációk kezelése teremti meg a lehetőséget arra, hogy a példányokat tulajdonságokkal is azonosítani lehessen.

A leszármazottak generálása mellett „hagyományosabb" módon is vesz fel jelölteket, szinonímalisták alapján. Így a többértelmű szavak problémáját is ugyanaz az algoritmus oldja meg: például a „körte" szó feldolgozása során a körte mint „gyümölcs" és mint „villanykörte" is a jelöltek közé kerül.

A jelöltek közül a kontextus alapján az ontológia segítségével történik a választás. A szövegkörnyezetben már beazonosított fogalmak és az adott jelölt között az ontológiában a relációk mentén mért *távolságon* alapul a választás. Ez egyrészt lehet saját maga korábbi előfordulása (mint például a „rizs-étel" példánál), de akár akkor is működik, mikor maga a fogalom nem is szerepelt. Például a „szögek beverésére alkalmas szerszám" mondatrész esetében, ha az ontológiába felvettünk olyan relációt, hogy „bekalapál: eszköze kalapács, tárgya: szög", akkor a „szerszám" fogalom leszármazottai közül a „kalapács" van legközelebb a kontextushoz (mivel a „bekalapál" reláció összeköti a „szög"-gel), és az algoritmus azt az eredményt hozza ki, hogy a „szerszám" szó itt speficikusan a „kalapács" fogalomra utal.

Jelentésreprezentáció ontológiában[87]

Szőts Miklós[1]

[1] Alkalmazott Logikai Laboratórium,
Hankóczy u. 7, Budapest 1202
szots@all.hu

Az a kutatás, amelyről beszámolunk, nem nyelvészeti ihletből született, hanem az ontológiák tanulmányozásából. Ennek megfelelően a nyelvészeti vonatkozások még kidolgozatlanok, de a „vizió" működését már demo szintű rendszerrel is bemutatjuk (l. Grőbler Tamás benyújtott előadását).

A kutatás célja, hogy egy szöveg jelentését formálisan reprezentáljuk. Ennek több praktikus felhasználási területe lehet, a jelenlegi projekt orvosi szabad szövegeket formális kórlapstruktúrába interpretál, de az igazi célkitűzés szemantikus kereső (text mining eszköz) fejlesztése.

Egyelőre a nyelvnek csak a leíró funkcióját kezeljük, tehát kijelentő mondatokból álló szövegeket elemzünk. Egy szöveg (szövegegység) **jelentése a szöveg által leírt szituáció reprezentálása egy világmodelben**[88]. A módszer a következő modulokból álló tudásbázison alapul:

a világról szóló tudás, azaz a **világmodell**,
a **nyelvi tudás**,
a kettő közti összefüggést reprezentáló **leképzés**.

Mind a világmodellt, mind a nyelvi tudást ontológiával reprezentáljuk.

A **világmodell**t reprezentáló ontológia relációkkal bőven felfegyverzett. Célunkhoz mérve[89] az ontológia szerkezetének legfontosabb elemei az eseményszerűségek, és a tulajdonságok. Az eseményszerűségek jellemzésére a szereprelációkat használjuk – lásd a 2005-ös előadást. A tulajdonságok reprezentálásánál ezek időfüggését, ill. más paramétertől való függést láthatóvá és kezelhetővé teszünk az ontológiában. Különösen jelentős az idő és a hely kezelése.

A **nyelvi tudás**t tároló ontológia, – nevezzük ezt lexikonnak, – „fogalmai" az önálló jelentéssel bíró nyelvi egységek, – lehetnek ragok, szavak, vagy több szóból álló frázisok, – nevezzük ezeket lexémáknak. A lexikon szerkezetét egy, a módszerhez jól illő, – feltehetően valamilyen lexikális felépítésű, – nyelvtan adná.

A módszer a világmodell, és nyelvi ontológia közti **leképzés**en alapul. Ennek a leképzésnek kell meghatároznia a nyelvi elemek által **referált** világmodellbeli egyedeket. Természetes módon a lexémákhoz fogalmakat, vagy fogalmak közti relációkat rendel – ezek előfordulásaira referálnak a lexémák. Ugyanakkor azokhoz a szintakti-

[87] a kutatás részint az NKFP-2/042/04, részint a GVOP-3.1.1-2004-05-0363/3.0 projekt keretében folyt.

[88] természetesen világmodell alatt egy jelenségkör egy adott szempontból felépített modelljét értjük.

[89] és meggyőződésünk szerint általában

kai viszonyokhoz, amelyek lexémákból mondatot fűznek össze, hozzárendel ontológiabeli relációkat.

Ez utóbbi adja annak lehetőségét, hogy a mondatokból az általuk referált szituáció leírását kapjuk. Legnyilvánvalóbb az igék vonzatkerete. Az ige[90] eseményszerűségre referál, a vonzataik pedig megfelelnek a szereprelációnak. Mivel a vonzatok jelentése nem univerzális, minden igénél külön a vonzatoknak megfelelő leképzés jelöli ki az ige referátumának megfelelő szereprelációt.

Mivel ez sarkalatos kérdés, hadd hozzunk már itt egy példát: az „ad" ige az ADÁS fogalomra referál, nominatívusz vonzata pedig az ADÁS-ból kiinduló *aktora* relációra. A „kap" ige ugyancsak az ADÁS fogalomra referál, viszont nominatívusz vonzata az ADÁS-ból kiinduló *recipiense* relációra.

A módszer nem kíván teljes lexikont: akkor is képes jelentésreprezentációt építeni, ha a szövegnek nem minden lexémája szerepel a lexikonban, – természetesen ekkor a jelentésreprezentáció is hiányos lesz.

Az elemző algoritmus jelenleg készen kapja a mondat alaktani és szintaktikus elemzését. Az alaktani elemzéssel szemben igényesek vagyunk, de a szintaktikai elemzés hiányosságait a reprezentációt építő algoritmus át tudja hidalni, akár mozaik elemzésből is felépíti a helyes reprezentációt. A fontos az, hogy a szintaktikai elemzés helyes lexémacsoportokat fogjon össze, és jól határozza meg a fejet. Ugyanis a jelenlegi algoritmus a fejből indul ki: az ez által referált fogalomhoz a vonzatoknak megfelelő relációkkal fűzi a fej által uralt lexémák referáltjait. A szintaktikai többértelműséget kezelni tudjuk: a vonzat - szerepreláció megfeleltetés kiszűri a szintaktikusan helyes, de értelmetlen elemzéseket.

A jelenlegi állapotban az alaktani, szintaktikai és szemantikai még el van választva, de olyan továbbfejlesztésről álmodunk, amikor ezek egymását segítve összedolgoznak.

Annak ellenére, hogy nem nyelvészeti ihletésre indult kutatásunk, nyilvánvalóak nyelvészeti vonatkozásai:

egy DR-hez hasonló struktúrát építünk, – a módszer jellegzetessége az, hogy ezt egy ontológiába építi be. Ezáltal a szemantika önálló életre lép, és kontrolálhatja, vezérelheti az elemzést.

a szereprelációkat a tematikus szerepek ihlették, még ha nem is pontosan azok;

módszerünk közel áll a nem transzformációs nyelvtanokhoz, pl. a konstrukciós, vagy a fejvezérelt fázisstruktúra nyelvtanhoz, az előadásban erre részletesebben kitérünk.

[90] a létige kivétel – azt nem vesszük „igének".

IX. Laptopos bemutatók

ALL-SPIDSY – Beszélőazonosító rendszer

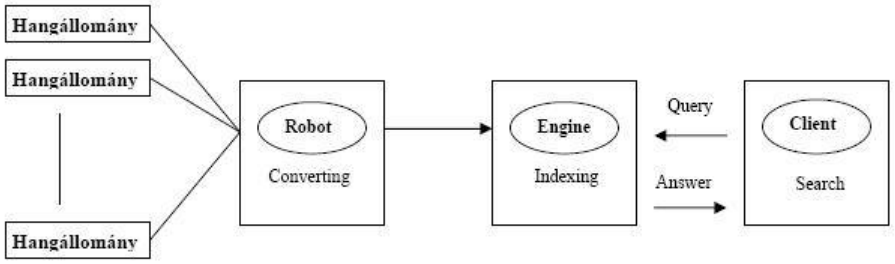
Karsai Győző

Alkalmazott Logikai Laboratórium
1022. Budapest Hankóczy J. u. 7., Hungary
gykarsai@all.hu

1. Bevezetés

A beszélőazonosító rendszerben (továbbiakban modul) egy vagy több kliens program (felhasználói felülettel ellátott alkalmazás) fut és felhasználja egy hang állomány indexelő program szolgáltatásait. A modul feladata hogy hang állományokban előre meghatározott beszélőket gyorsan és hatékonyan felismerjen ill. az összes előfordulásukat lokalizálja. A modult az Alkalmazott Logikai Laboratórium fejlesztette ki, ill. jelenleg is fejleszti tovább. Az itt bemutatott modul egy nagy elképzelés csomag része, mely csomag egy része már megvalósult. A beszélőazonosító modulban az egyes rendszer elemek külön önálló életet élnek (aszinkron működés), és csak a meghatározott módon kommunikálnak egymással, tehát a lehető legjobban megőrzik az egyes rendszer elemek az integritásukat és függetlenségüket. Ennek az az előnye is megvan, hogy pl. az indexelő programot le lehet cserélni egy olyanra ami nem beszélőket keres hangállományokban hanem valamilyen szignált, vagy szöveget vagy éppen egy dallamot. Ezért a kialakított rendszer elemei – Robot, Engine, Interface, Client - teljesen kompaktak abban az értelemben, hogy bármelyik kicserélhető egy újabb verzióra anélkül, hogy a többi elem ezt „észrevenné".

A beszélőazonosító modul felépítése



Egy kereső motor, amely a modul legfontosabb egysége. Több is lehet belőle. Azaz általában egy Engine-t használunk, de több, összehangoltan működő Engine is lehet a rendszerben főleg hatékonysági megfontolásokból. Az Engine a szolgáltatását több fázisban hajtja végre. Tipikusan először egy indexelést végez egy adott hang állományon, ezzel teszik lehetővé annak kereshetőségét. A már sikeresen indexelt hangállományok kereshetők, azaz megjelenhetnek, mint egy keresés eredménye.

Indexing

Az eljárás, amivel az Engine a számára biztosított hang állományból előállít egy adatstruktúrát, amelyen a konkrét beszélő keresése, ill. amivel a keresési közegben a kurrens keresés már gyorsan végrehajtható,

Client

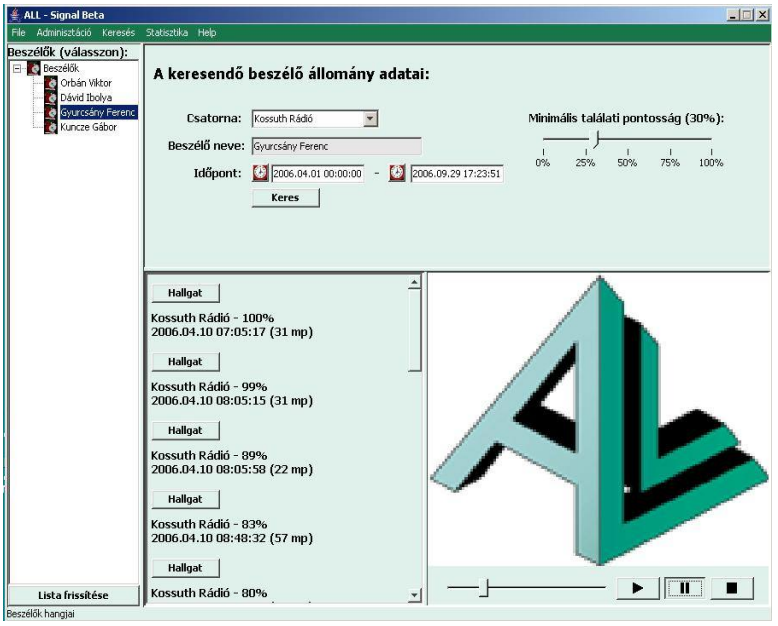
Az Engine szolgáltatásait használó rendszer komponens. Ő kezdeményez minden akciót az Engine felé, ill létrehozza az eredmény listát és a statisztikákat a keresés eredményéből.

Robot

Egy konvertáló motor. A Robot egy hang állományon végrehajt egy Converting-et (ha szükséges), majd továbbítja azt az Engine felé feldolgozásra.

Converting

Mivel az Engine csak egy formátumot ismer, ezért hang állomány legelterjedtebb formátumaiból elő kell állítani ezt a formátumot megfelelő konvertálással. Ezt természetesen csak akkor kell elvégezni, ha szükséges.



1. ábra Egy találati eredményeket tartalmazó képernyő:

Működés

A beszélőazonosító modul úgy működik, hogy egyrészt van egy Robot, ami felkutatja az elérhető, és feldolgozható „hang" állományokat. Amikor megtalál egy ilyen „hang" állományt - ami éppenséggel lehet egy Mpeg állomány is – átkonvertálja azt olyan „igazi" hang állománnyá, amit az indexelő program, az Engine fel tud dolgozni. A Robot a konvertált hang állományt átadja az Engine-k. Az indexelő programnak egyéb módon is át lehet adni hang állományokat indexelésre (batch, felhasználói felület). Tehát a feldolgozást egy Engine végzi, ami indexeli a hangállomány tartalmát, s egy előkeresést is végrehajt rajta. Az előkeresés célja az Engine által már ismert beszélők felkutatása az éppen indexelt állományon. Ez az indexelés-

előkeresés folyamat akkor is lezajlik, ha egy beszélőt leíró hang állománnyal ismertetjük meg az Engine-t. A beszélő felismerő modul harmadik fontos eleme a Client, ami az előkeresés eredményeire támaszkodva azonnal elő tudja állítani a kurrens keresés találati listáját a felhasználói felületen. A találati lista megjelenítésén túl a Client aktuális statisztikákat is gyárt a kurrens keresésben szereplő adatokból. Az indexelés sebességét illusztrálja, hogy egy 24 órás hangállomány 25-30 perc alatt indexelődik (kb. 2%-a a valós időnek). További példa az előkeresés sebességére vonatkozik: a sebesség egy 24 órás hangállományra kevesebb, mint 1 mp beszélőnként. A keresendő beszélők számának ill. a keresési közeg méretének csak a rendelkezésre álló hardware szab korlátot, mivel a modulok működéséhez minimális erőforrás szükséges. A modulok bármelyike több példányban is működhet összehangoltan, s ennek csak a rendelkezésre álló futtató környezet korlátai szabnak határt.

Referent Systems and Argument Structure

Kracht Marcus

Magyar Tudományos Akadémia, Nyelvtudományi Intézet,
Benczur u. 33, 1068 Budapest, e-mail: kracht@nytud.hu
and Department of Linguistics, UCLA, 3125 Campbell Hall, PO Box 951543,
Los Angeles, CA 90095-1543, email: kracht@humnet.ucla.edu

Abstract. The assignment of variables plays a pivotal role in the construction of semantics of complex expressions. In this paper we discuss the theory and implementation of an algorithm to identify variable names. It is based on referent systems, introduced in [5]. The theory is exposed in [3] and discusses in-depth the properties of referent systems. [1]

Keywords: argument structure, referent systems

1 Introduction

A crucial problem in the composition of meanings is the problem of *variable names*. Montague originally devised a semantics that assigned closed expressions to each word, thus relegating the problem to the λ-calculus. However, [1] has pointed out that this is problematic in view of transsentential binding. He proposed an alternative that led to the development of discourse representation theory (DRT), which uses—at least in its original form—no λ-binding mechanism at all. All unquantified variables where free. What needed to be solved, then, was the assignment of variable names.

Kamp and Reyle describe in [2] an algorithm to derive the semantics of a sentence. This algorithm needs a parser that not only produces a structure but also distributes indices to the constituents. Thus, the input to the semantical translation of (1.a) is (1.b) rather than (1.c).

(1) (*a*) `Egy fekete macska látja az egeret.`
(*b*) [[$\texttt{egy}_1$ [$\texttt{fekete}_1$ $\texttt{macska}_1$]] [$\texttt{látja}_{1,2}$ [$\texttt{az}_2$ $\texttt{egeret}_2$]]]
(*c*) [[`egy` [`fekete` `macska`]] [`látja` [`az` `egeret`]]]

The reason for this is that in DRT every variable is global. The variable x points to the same object independently of the DRS in which it occurs. To see this,

[1] The author wishes to thank Váradi Tamás and Kenesei István for their generous support.

look at the way two DRSs are merged. The phrase `egy fekete macska` consists of three DRRs, each of which uses at least one variable.

(2)

/egy/	/fekete/	/macska/
x	∅	∅
∅	$\mathsf{black}'(x)$	$\mathsf{cat}(x)$

Merge is associative and consists in taking the set union of the upper and the low box, respectively. Merging the first two, for example, results in accidental capture of the variables of the second DRS by the quantifier of the first, like this:

(3)

/egy macska/
x
$\mathsf{black}'(x).$

Merging the upper three we get

(4)

/egy fekete macska/
x
$\mathsf{black}'(x); \mathsf{cat}'(x).$

The problem with this approach is that if, for example, the middle DRS uses y in place of x we get an incorrect result:

(5)

/egy fekete macska/
x
$\mathsf{black}'(y); \mathsf{cat}'(x).$

This is because the merge operation cannot know whether two variables in different DRSs are meant to be 'the same' or not. To solve this problem, [2] simply relegated the problem to the parser; it was the responsibility of the parser to distribute the correct indexation to each lexical entry. The indices, in addition to being useful for syntax, provided the essential information to insert the correct names for the variables. The index i is simply translated by the variable $\mathtt{x}_i$. (Notice, by the way, that entries with several free variables need several indices, and the order matters.)

2 Referent Systems

[5] has shown that from a logical point of view there is no need to do indexing if variables are instead considered local. Instead of considering two variables of two DRSs identical if they carry the same string, we assume by default that the variables of distinct DRSs are different *unless stated otherwise*. To say that two variable are to be identified, we associate a so-called 'name' with a variable. (Names are optional; if a variable has no name, it simply cannot be identified.) In OCaML, such names have effectively the same mechanics (and are used for

similar purposes) as *labels*. Names can be everything, but the idea is that in natural languages names are morpho-sytactic properties, like cases and grammatical roles. When merging two structures two occurrences of the same variable (or of two different variables) are made the same in the output DRS if (and only if) they carry the same name. This type of variable is called a **referent**. With the help of referent systems the argument structure can be enriched in such a way that the indexation proceeds automatically. If an entry, say a verb, needs several arguments, we want to allow it to take each argument in turn. Most syntactic theories postulate a canonical deep order in which the arguments are consumed in the same way as programming languages insist on the arguments being fed to a

function in the order specified. In OCaML, for example, we may declare a function in the following way:

(6) `let f x y = 2 * x + y;;`

In this case the variables `x` and `y` are plain variables, and bound inside the function declaration. Order matters. Evaluating `f 3 2` gives 8, evaluating `f 2 3` given 7.

Freedom from this order regime comes in the form of *labels*. Consider this slightly different definitions using the tilde convention:

(7) `let f ~x:a ~y:b = 2 * a + b;;`

Here, `~x` and `~y` are labels; `a` and `b` are the associated variables. The advantage is that order is now irrelevant: `f ~x:2 ~y:3` and `f ~y:3 ~x:2` both yield 7, and `f ~x:3 ~y:2` and `f ~y:2 ~x:3` both yield 8. Basically, it is the *order independence* of this mechanism that we exploit.

3 Argument Structure

We aspire for a surface oriented approach, that is, we want to interpret every constituent where it actually occurs. Moreover, we require arguments to be *adjacent* to each other. Given these requirements we must accommodate free word order not by distinguishing two different syntactic representations, but by allowing arguments to be identified by other means than their surface position. This leads to the idea of using inherent properties of the arguments as a way to identify them with variables of the head. These properties are, in Hungarian, foremost case, but also person, number, and definiteness. A verb decides not only the basis of position but on the basis of case which constituent is its subject and which one is object etc. The properties thus constitute the *name* of the argument.

The theory by Vermeulen is insufficient in certain respects. In its original form it distinguishes a left incoming name from a right outgoing name; however, the left-right distinction is relegated here to the morphology and does not figure at all in the semantics. However, the notion of incoming and outgoing names is

important. Consider merging two constituents A and B. Then one of them, say A, will assume the role of the functor, taking the other, B, as argument. The variable x of A and y of B are identified if the outgoing name of y in B matches the incoming name of x in A. In tandem with names, each variable is associated with a diacritic that states whether or not the variable actually has an incoming and/or outgoing name. For names can be dropped, in which case a variable loses its ability to be identified with other variables in further computation. By default, incoming names and outgoing names are the same; but they need not be. For syntactic purposes we need to distinguish arguments from adjuncts; also, we need to distinguish an ordinary variable from a parameter. All these characteristics are unified into a so-called argument identification statement (AIS). For each syntactic argument such an AIS must be issued. An **argument structure** (AS) is a sequence of AISs. (To stress: it is not necessary to have an AIS for every variable; an AIS is only needed if the variable needs to be manipulated.) It contains the name of the variable, it contains a so-called *diacritic*, specifying whether the variable is imported ($\triangledown$) or exported ($\triangle$), or both ($\Diamond$). Furthermore, if a variable is imported, a name is given under which it is imported; if it is exported, a name is given under which it is exported. Names have the form of an attribute value structures, with usual notion of unification (thus allowing for certain types of abstractness).

A constituent A can be merged with a constituent B only if the semantic merge identifies at least one variable. (Two variables is also possible, for example in control structures.) We note here that the merge of a single variable has consequences on a different set of referents, called *parameters*. Unlike ordinary variables, parameters are identified through their role (eg *reference time*, *event time*, *worlds* and so on). An AIS associates with a variable two sets of parameter statements. These have the form [*role* : *ref*], with *role* a role and *ref* a referent. The first set describes the incoming parameters, the second the outgoing parameters. When A takes B as argument, and x is identified with y, then an *incoming* parameter for a role ρ is identified with the *outgoing* parameter for the role ρ of B. Parameters not mentioned in the lists are simply passed unchanged. It is possible to reevaluate parameters but also to make them change roles. For example, in Russian the reference time in the subordinate clause of an indirect speech act equals the event time of the main clause, while in English it equals the reference time of the main clause. Thus this mechanism can be used for sequencing context parameters, such as time, person, world and location (sequence-of-world, sequence-of-time, sequence-of-person and so on, as described in [4]).

Thus, a complete entry has three components:

1. an exponent, for example a string (but more complex exponents are implemented, see below);
2. an argument structure (AS). This is a sequence of argument identification statements.
3. a semantics, for example a DRS.

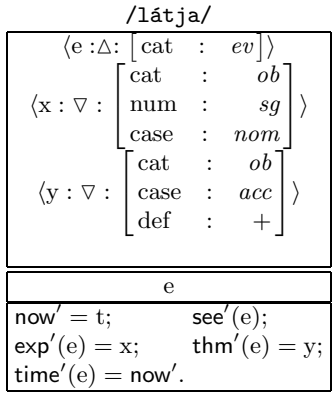
The structure (1) shows all three components, the string on top, the argument structure in the middle, and a DRS at the bottom.

The AS replaces not only the indexation but in fact a lot of the structure building itself. The syntactic categories are encoded mainly in the outgoing names. The complexity of merge is very low. Unification proceeds in $O(m + n)$ time. Thus to compute the semantics of a sentence is very fast ($O(n^3)$ for context free grammars). This is *not* true of our implementation, since the implementation also returns *all* parse terms and evaluates them into meanings. Since structural ambiguities can in worst cases be exponential in the length of the string the implementation runs in exponential time as worst case. (This is mainly due to the fact that the implementation is meant to reflect the theory as accurately as possible.)

4 Agreement

Fig. 1 gives an example of an entry. The referent x belongs to an argument ($\triangledown$) which has to be a thing (cat : *ob*), in the nominative (case : *nom*) and singular (num : *sg*). The e referent however has a $\triangle$, which means that it does not belong to an argument. Thus it is indicated that the denotation of the word `látja` is an event (cat : *ev*). The surface orientation of our approach

Fig. 1. The argument structure and semantics of the word 'látja'



has the following consequence. Names are needed to identify referents across structures; they must therefore be computable properties of the argument itself. The appearance of incoming names for the arguments is therefore correlated with surface differences in the arguments themselves. We see definiteness figure in the identification statement for the object for the reason that there is a different set of endings for definite objects. The verb flags for its object to be definite. It also flags for it to have accusative case. Here is an example from German.

(8) `Dir schärfsten Kritiker hat die Präsidentin in ihrer Heimat.`
The harshest critics has the president in her home [country].

Both arguments can be both nominative and accusative. However, when the verb shows singular agreement, excluding the first from being subject, since it unequivocally plural.

5 The Implementation: Description

The implementation is written in OCaML, a functional programming language. It has both a command line interface and a Tk-interface to allow for interactive sessions. The software is designed to support Unicode and multiple languages. At present, it can be both installed and run in English and German. Documentation is also available in both languages.

The output is sent to a .tex-file, which is translated using LaTeX, and is then shown to the user, but can also be stored independently for later use.

The algorithm proceeds via so-called *entries*, which are records consisting of four fields corresponding to

- the morphology: this is a set of *morphs*;
- the argument structure, which is an array of *argument identification statements*;
- the semantics, which is a DRS;
- the set of parse terms, which show the analysis terms of the entry.

Morphs consist of

1. an exponent, which is an array of string;
2. an array of subcategorisation statements. These consist in turn in a specification for each argument that the entry takes of
   (a) the required morphological class of the argument (possibly also its form)
   (b) the class (and shape) of the element produced when the argument is consumed,
   (c) the way the functor and argument morphology need to be combined (concatenation, reduplication and so on).

Argument identification statements consist of

1. a variable name;
2. a diacritic displaying the way in which the variable must be handled during merge;
3. a syntactic class for the argument to be consumed;
4. a syntactic class of the element constructed if the argument is consumed;
5. a parameter statement, showing the way in which parameters are consumed and passed up.

Notice that morphs need not consist of a single string, they can consist of several strings (thus we can accommodate circumfixes, but also the fact that in Hungarian the verbal prefix can be split from its verbal root). There are however no functions that allow to change any letter, and there may not be any empty morphs.

There is no inbuilt distinction between words and morphemes. The blank is considered a symbol of its own, like punctuation. Given a string as input, the system will match the morphs of the dictionary against any combination of substrings. If morphs consist of at most k units (typically, $k = 2$ is sufficient), then this gives $O(n^{2k})$ many occurrences, where n is the length of the string. The algorithm is a chart, which is implemented as a hash-table over pairs (ℓ, k), where ℓ is the overall length of the occurrence (the sum over all lengths of the parts), and k is the index of the leftmost occurrence. The chart is constructed by induction over the length and is finished when that length equals the string length. The output is then returned.

The chart parser operates on *occurrences*. These are quadruples, consisting of an argument structure, an occurrence of a morph (an array of pairs of positions), some morphological components (stating how the element may be further combined) and a term. The semantics is absent. It is inserted only when the successful terms are finally evaluated. This allows to keep the burden on the parser small. Currently, it is not very efficient, but it can be made much faster if need be.

The most flexible session is the standalone-session. After compilation, dictionaries can be loaded and unloaded dynamically. A useful tool is the command `diagnose`. Given two entries it documents the calculations in a step-by-step fashion. All outputs can be saved in a file and used for different purposes.

6 Documentation and Source

The software was originally designed to allow for the evaluation of the theory of argument structure. The software and the theory are now under simultaneous development. At the time of writing, some developments of the software are not yet reflected in the documentation. The current version of the software is 5.0. Installation currently is possible for Unix platforms only and has been tested on several of them, including MacOS X. Both the software and the manuscript can be freely obtained from

`http://kracht.humnet.ucla.edu/marcus/referent`

subject only to usual open license conditions.

References

1. Kamp, Hans: A theory of truth and semantic representation, in: Groenendijk, Jeroen (ed.): Formal methods in the study of language, Mathematisch Centrum, (1981).
2. Kamp, Hans, Reyle, Uwe: From Discourse to Logic, Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory, Kluwer, Dordrecht,
3. Kracht, Marcus: Agreement Mophology, Argument Structure and Syntax, Manuscript, UCLA, 2006.

4. Schlenker, Philippe: A Plea for Monsters, Linguistics and Philosophy 26, 29–120, (2003).
5. Vermeulen, Kees F. M., Merging without Mystery or: Variables in Dynamic Semantics, Journal of Philosophical Logic (24), 405–450, (1995).

Automatikus verselemzés tanuló algoritmusok alkalmazásával

Lesi Zoltán[1]

[1] Nokia Hungary Kft. 1092 Budapest Köztelek utca 6.
zoli@nix.hu

Kivonat: Cikkünkben a számítógépes verstani elemzés tárgykörét, valamint ezen új területen elért eredményeinket mutatjuk be. Egy automatikus vers elemző jelentősen megkönnyítheti a nyelvészek, irodalmárok munkáját, különösen nagy korpusz vizsgálata esetén. Segítheti az irodalmi és nyelvi állítások bizonyítását. A fonetikai vizsgálatok képezik a verstani analizálás alapját. Részletesebben a metrika, alliteráció és rím gépi meghatározásával foglalkozunk, melyhez tanuló algoritmusokat alkalmazunk. Eredményeinket TEI P4 konzorciumnak megfelelő XML-lel írjuk le, mely tartalmazza a versek fonetikai, szótagszerkezeti és metrikai leírását, végrímeket, alliterációkat és a szófaji elemzést is. Az algoritmusunkat Weöres Sándor szonettjein és szonett fordításain teszteltük, igazolva elképzeléseink helyességét.

1. Bevezetés

A szépirodalmi szövegeket többféle szempont szerint csoportosíthatjuk. Horváth Iván *A vers* [3] című könyvében három megközelítést mutat be. Ha a transzcendens versolvasó szemüveget tesszük fel, a verset nem tudjuk másnak látni, mint költeménynek, ilyenkor a verselmélet egybeolvad a költészet elméletével. Tekinthetjük a verset nyelvi egyetemességnek: olyan megjelölt beszédmódnak, amely valamiképpen minden természetes nyelvben létezik. A harmadik módszer szerint a vers az, amit egy bizonyos irodalmi hagyomány részesei annak tartanak.

A számítógépes nyelvészetben a verselemzés új kutatási terület, hiszen magyar nyelvű szövegekre kidolgozott (vagy magyar szerző munkájaként ismert) automatikus verselemző programról nincs tudomásunk. A probléma megoldása talán azért is váratott magára, mert nem könnyű elhatárolni a megoldható és egyelőre megoldhatatlannak tűnő részeket.

Fónagy Iván nemzetközi hírű nyelvész, pszichológus a hatvanas években megtervezetett [2] egy programot, amely prózai és verses szövegekkel foglalkozik, majd kiegészítette két másik fejezettel: „Program köznyelvi szövegekre", „Program költői szövegekre". A tervek világos, pontokba szedett szempontokat tartalmaznak, amelyek statisztikai jellegű információkra mutatnak. Ez a szempontrendszer adta a nyelvészeti és verstani alapot programunkhoz.

1.1 Weöres Sándor és a korpusz

Weöres Sándor (1913-1989) költő, műfordító, író. Bori Imre tanulmányában [1] írja, hogy jellemző Weöres költészetére egy fontos zenei mozzanat a kettősség: ha Weöres verseinek nagyobbik hányadát a zene fogalmával helyettesíthetjük, úgy van egy verscsoportja, amely a „nem-zenét" jelenti. Ez a zenei elv összefogja Weöres költőiségének alapvető törekvését, amely a harmónia utáni vágyban s a költői megvalósulásában nyilatkozik meg. Weöres szerint: „A szonett első nyolc sora a nyolcoldalú kristály, az oktaéder: a végső hat sor az előbbiek ismétlése, más összeállításban, más összefüggésekkel."[10]

Nagy L. János és Alexin Zoltán 1999 és 2002 között létrehozták a virtuális kritikai kiadás 'editio princeps'-ét, hogy minél teljesebb Weöres-korpuszon dolgozhassanak. [6] Az 1986-ban megjelent háromkötetes „Egybegyűjtött írások" című gyűjteményt vették alapul.

2. Automatikus fonetikai elemzés

A szövegek fonetikus konverziójához Kassai Ilona fonetikus átíró szabályokat tartalmazó táblázatát alkalmaztuk. [4] Megvizsgálva a karakter környezetét, szintaktikai elemzés alapján döntjük el, hogy az adott helyen digramma van-e. Összetett szavak határán előfordulhat olyan eset, mikor a látszólag összetartozó karaktereket különböző hangokká kell alakítani (pl. százszor).

Amennyiben egy magyar szóról van szó, akkor a fonologikus szabályokat, idegen szó esetén az „idegen szavak átírási szótárát" vesszük figyelembe. A fonetikai átírás másik problémája, hogy az idegen nevek, szavak más átírást követelnek, mint a magyarok. Tovább nehezíti a helyzetet, hogy ezek többnyire magányosan szerepelnek. A problémát nyelvazonosító rendszer [5] alkalmazásával sikerült megoldani. Különféleképpen jelöljük az allofónokat, hangváltozatokat.

3. Automatikus verstani elemző

A morfológiai elemzést (melyet a versek HuMor [8] elemzéséből nyerünk) és a hangtani vizsgálatokat felhasználva meghatározhatóak a számunka fontos alkalmazások a metrika, az alliterációk és a rímek. A megalapozott végeredményeket egy XML fájlban összegezzük, és lekérdezéseket hajtunk végre.

A sorvégi egybecsengések illetve alliterációk vizsgálatához szükség van a hangok összehasonlítására. A fónagyi tervezet tartalmaz egy leírást a fonémák eltérési fokáról, mely egyszerűen algoritmizálható.

A metrika, az alliteráció és a sorvégi egybecsengések alkalmazásait heurisztikus és tanuló algoritmusokkal is meghatároztuk, majd a részeredményeket szavazással egyesítettük.

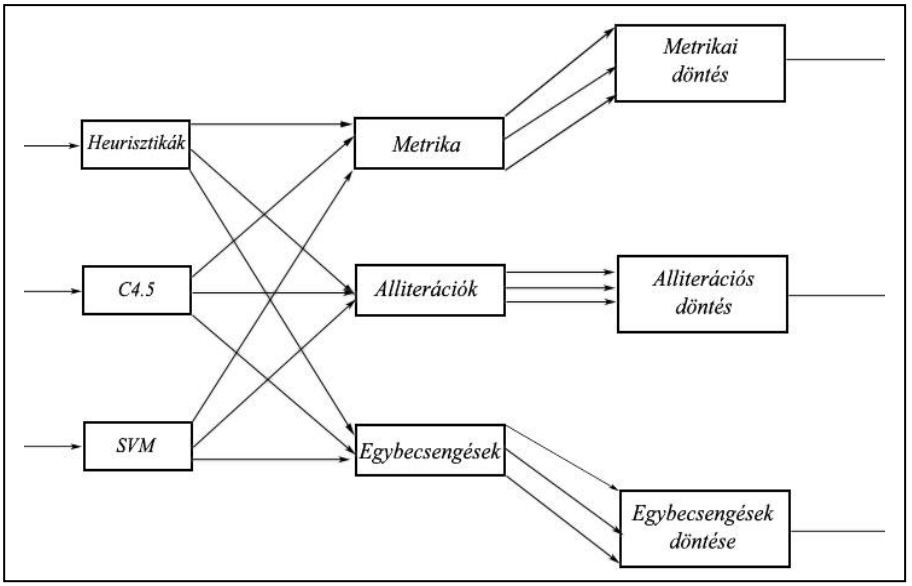


Fig. 2. A heurisztikák és a tanuló algoritmusok együttműködése

A következő táblázat a 152 szonettből álló korpuszra lefuttatott verselemző program eredményeit mutatja be:

	Heurisztikák	C4.5	SVM	Döntés után
Metrika	100%	100%	99.87%	100%
Alliterációk	97.36%	97.30%	90.92%	99.59%
Sorvégi egybecsengések	53.75%	81.54%	78.94%	81.73%

4. Az elkészült verselemző rendszer

A 2004. októberében elkezdett tudományos munka legfőbb eredménye, hogy a Fónagy-tervezet fejezetei alapján megterveztük és implementáltuk az első magyar automatikus verselemzőt. Az első időszak feladatai: a Fónagy-tervezet értelmezése; megvalósíthatósági tanulmány, DTD tervezet készítése, valamint a verstani, hangtani ismeretanyag összegyűjtése. A fónagyi tervezetben a statisztikai szempontokat visszavezettük alapadatokra. Nagy L. János kurzusán teszteltük a szempontrendszert, amely később a program alapjává vált. A program ellenpontjaként a diákok elemezték a verseket.

Összegyűjtöttük a virtuális kritikai kiadásból Weöres Sándor összes (101) szonettjét, valamint 91 szonett fordítását. Annak ellenére ragaszkodtunk a szonett formához, hogy a korpuszunk így aránylag kis méretű lett. Mindenképpen homogén versformájú anyagot akartunk: szonettekből találtunk a legtöbbet. Az egyes versekre

adott eredmény könnyebben összevethető a korpuszéval, és a szonett formát is jellemzi.

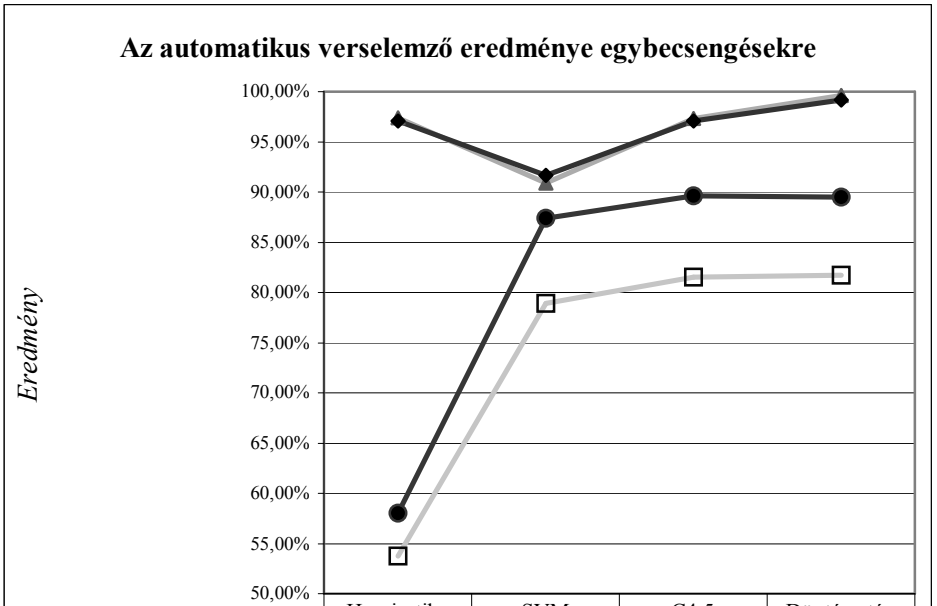
A fonetikai és a fonológiai elemző alapja a fonetikai átíró szabályrendszer. A további vizsgálataink miatt pontos fonetikai eredmények szükségeltettek. A digrammák szétvágását (pl. százszor) morfológiai elemzés alapján végeztük, tesztjeink igazolják a módszer helyességét. A versekben előforduló idegen szavak elszigetelten fordulnak elő, felismerésükhöz nyelvazonosító rendszert használtunk. Bizonyos hangoknak több ejtésváltozata (allofónja) van, így néhány újabb szabállyal kellett bővíteni a rendszert, valamint a „méh" típusú *h* miatt a szóvégmutató szótárt [7] alkalmaztuk. A hangok egymásra hatásakor megváltozik a fonetikus leírás, ezért fonologikus szabályokat alkalmazunk.

A magánhangzók eltérési fokának meghatározása hiányzott a Fónagy-tervezetből, ezt „A mássalhangzók eltérési foka" című fejezet alapján pótoltuk. A metrika, az alliterációk és a rímek heurisztikus meghatározása a Fónagy-tervezet alapján történt.

Az eredmények pontosításához C4.5 és SVM tanuló algoritmusokat használtuk. Létrehoztuk a három alkalmazás(metrika, alliteráció és sorvégi egybecsengések) modelljét, és dekomponáltuk a feladatokat. A metrikánál a szótag hosszúságát, az alliterációnál a kezdőbetűk egybecsengését elemeztük. A morfológiai elemzés segítségével a kötőszavakat és a névelőket kiszűrtük, mert csak a szűkebb értelemben vett alliterációkat kerestük. A rímeket szétbontottuk egy szótagos egybecsengésekre, ahol elsősorban a szótag tulajdonságait, a tanulóhalmazban pedig az egymással rímelő sorpárok gyakoriságát vizsgáltuk.

A két tanuló algoritmus és a heurisztika eredményeit szavazással egyesítettük, a végeredmény a sorvégi egybecsengések vizsgálatakor pontosabb, a másik két alkalmazásban pedig mind a tanuló modellek, mind a heurisztikus algoritmusok kiváló eredményt adtak, ezért érdemes volt tanuló algoritmusokat használni. Meglepő eredmény, hogy a C4.5 pontosabban osztályoz, mint az SVM, ezt a két szintű osztályozás és a diszkrét értékek okozzák. Biztosak lehetünk abban, hogy más feladatokra, más adatokon az SVM lenne a megfelelőbb.

A következő diagram a 152 szonettből álló tesztkorpuszra, és az *Átváltozások* ciklusra lefuttatott sorvégi egybecsengéseket, alliterációkat vizsgáló alkalmazások eredményeit mutatja be.



	Heurisztika	SVM	C4.5	Döntés után
Sorvégi egybecsengések 152 szonettre	53,78%	78,94%	81,54%	81,73%
Sorvégi egybecsengések az "Átváltozások" 40 szonettjére	58,01%	87,39%	89,61%	89,47%
Alliterációk 152 szonettre	97,36%	90,92%	97,30%	99,59%
Alliterációk az "Átváltozások" 40 szonettjére	97,05%	91,64%	97,05%	99,20%

Fig. 3. Az automatikus verselemző eredménye a két vizsgált korpuszra

A végeredményeket egy nemzetközileg elismert TEI kompatibilis XML fájlban összegeztük [9]. ehhez felhasználtuk a már kész DTD terveket. A TEI ajánlást betartva, minimális módosítással, oldottuk meg a konverziót. A DTD a TEI ajánláshoz képest a következő új elemeket tartalmazza: <w>, <ana>, <cons>, <c>, <link>, <linkGrp>, valamint bővíteni kellett a <body>, <lg>, <l> elemeket is.

Az elemzés elvégzése után – a Fónagy-tervezet alapján – lekérdezéseket hajtottunk végre. Néhány fonetika, metrika, alliteráció és sorvégi egybecsengések lekérdezéseit megvalósítottunk.

Magyarországon eddig nem készült automatikus verselemző szoftver. Automatizálási programunk célja, hogy a magyar nyelvű gépi verselemzés kutatását elindítsuk, és a nemzetközi tudományos áramlatba bekerüljünk.

5. A számítógépes verselemzés távlatai

A fonetikai rendszerben, komoly gondot okozott az idegen szavak átírása. Hiba már a nyelvazonosításnál felléphet, ha ismeretlen nyelvvel találkozik, mert az átírását nem biztos, hogy ismerjük. Hibás fonetikus átírás esetén, a fonemikus jegyekre, szótagszerkezetekre, digrammákra, szóhosszúságokra, rímekre, alliterációkra és metrikára pontatlan eredményeket kapunk.

A szintaktikai elemzés rohamosan fejlődik, de pontatlansága problémát okoz a szófajok, alliterációk vizsgálatakor. A Fónagy tervezetben szereplő szempontok automatizálhatóak, azonban többértelműségek miatt pontatlanságra számíthatunk. Például: A központozás hiánya bizonytalanná teheti a tagmondatokra bontást, a felsorolások, közbeékelt mondatok felismerését, a mondatok hosszát és a modalitását.

A versekben nem jelölt metrikai kétértelműségek elemzése a gép számára megoldhatatlan. Pl. „A mint B" (W.S: Önéletrajz) Az „A" hosszan ejtendő, de ezt nem jelöli semmi a gép számára. A versrendszer (időmértékes, ütemhangsúlyos vagy szimultán) felismerése is probléma lehet a gépi elemzés számára – mivel megfelel a szabályoknak – az *Átváltozások* ciklusban szereplő *A nyüzsgés* című szonett jambikus, de valójában nem is időmértékes.

A rímképletek meghatározása a rímelő sorok ismeretében előfordulhat, hogy a két sor között létezik egybecsengés, de ez a képletben nem jelenik meg.

A formai jegyek szerepét, illetve a rímek és az alliterációk jelentőségét – a mű értelmezésének tükrében – nem tudjuk számítógéppel meghatározni.

Bibliográfia

1. Bori Imre: A szintézisteremtő. In: Bori Imre huszonöt tanulmánya a XX. Századi magyar irodalomról, Forum Kiado, Újvidék, 1984.
2. Fónagy Iván: Program prózai és verses szövegek elemzéséhez. Kézirat. 1966. Javított változat: Antony, 1997.
3. Horváth Iván: A Vers. Gondolat Kiadó, Budapest, 1990.
4. Kassai Ilona: Fonetika. Nemzeti Tankönyvkiadó, Budapest, 1998.
5. Kiss Géza, Németh Géza: Skálázható szöveg-alapú nyelvazonosító módszer beszédszintézis céljára, In: III. Magyar Számítógépes Nyelvészeti Konferencia 2005. SZTE TTK Informatikai Tanszékcsoport, Szeged. Szerk.: Alexin Zoltán, Csendes Dóra.
6. Nagy L. János, Alexin Zoltán: Weöres költői nyelvének számítógépes feldolgozása. In: II. Magyar Számítógépes Nyelvészeti Konferencia 2004. SZTE TTK Informatikai Tanszékcsoport, Szeged. Szerk.: Alexin Zoltán, Csendes Dóra.
7. Papp Ferenc: Szóvégmutató szótár. Akadémiai Kiadó, Budapest. 1966.
8. Gábor Prószéky, and Balázs Kis, 1999. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 261-268. College Park, Maryland, USA
9. The Text Encoding Initiative Consortium. (http://www.tei-c.org)
10. Weöres Sándor: Oktaéder-kristály. In: Új Írás, 1977. 4

Névmutató