Általános Nyelvészeti
Tanulmányok
XXIV.

Általános Nyelvészeti Tanulmányok XXIV.

Nyelvtechnológiai kutatások

Főszerkesztő:

Kenesei István

Szerkesztette:

Prószéky Gábor és Váradi Tamás

Akadémiai Kiadó, Budapest

A kiadvány a Magyar Tudományos Akadémia támogatásával készült

ISBN 978 963 05 9308 3

Kiadja az Akadémiai Kiadó,
az 1795-ben alapított Magyar Könyvkiadók
és Könyvterjesztők Egyesülésének tagja
1117 Budapest, Prielle Kornélia u. 21–35.
www.akademiaikiado.hu

Első magyar nyelvű kiadás: 2012

A kiadásért felelős az Akadémiai Kiadó Zrt. igazgatója
Felelős szerkesztő: Vajda Lőrinc
Termékmenedzser: Egri Róbert
A számítógépes szerkesztés G. Kiss Zoltán munkája LaTeX 2_ε rendszerrel
A nyomdai munkálatokat a Prime Rate Kft. végezte
Felelős vezető: Tomcsányi Péter
Budapest, 2012
Kiadványszám: TK120063
Megjelent 31,46 (A/5) ív terjedelemben

HU ISSN 0569-1338

Szépe György (1931–2012)

Kötetünket az Általános Nyelvészeti Tanulmányok*at*
a kezdetektől gondozó Szépe György emlékének ajánljuk

Tartalomjegyzék

Szerkesztői bevezetés

Az olvasó az *Általános Nyelvészeti Tanulmányok* egy újabb tematikus kötetét tartja kezében, amelynek alcíme: *Nyelvtechnológiai kutatások*. A *nyelvtechnológia* szó talán többek számára magyarázatra szorul. Ez a fogalom a 20. század második felében inkább *számítógépes nyelvészet* néven volt ismert. Korábban még az ezzel rokon *matematikai nyelvészet* kifejezés is használatos volt: 1962-ben a Magyar Tudományos Akadémia Nyelvtudományi Intézete munkaértekezletet szervezett *A matematikai nyelvészet és a gépi fordítás kérdései* címmel. Ennek a kereken fél évszázaddal ezelőtti eseménynek az előadásai láttak napvilágot 1964-ben az *Általános Nyelvészeti Tanulmányok* II. kötetében. Sorozatunkban azóta a számítógép és a nyelv kapcsolatáról nem jelent meg írás. Ez alatt az ötven év alatt viszont a számítógép oda jutott, hogy szinte minden rajta futó alkalmazás találkozik az emberi nyelvek beszélt vagy írásos formájának valamelyikével: egymásnak szánt szövegeink, leveleink, híreink, feljegyzéseink, dolgozataink, folyóirataink, könyveink, tudományos publikációink – és még sorolhatnánk – valamilyen emberi nyelven íródnak, és a gépek ezeket a szövegeket segítenek létrehozni, kijavítani, lefordítani, vagy éppen keresni bennük. Ez angol nyelvterületen persze angolul történik, spanyol nyelvterületen spanyolul, Magyarországon pedig magyarul. Az ezeket a tevékenységeket leíró számítógépes nyelvészeti irodalomban sokat használt kifejezés a szakterület megnevezésére a *természetesnyelv-feldolgozás (natural language processing),* bár angolul napjainkban egyre inkább a *human language technologies* elnevezést használják. Magyarul ez a fogalom vonult be *nyelvtechnológia* néven a szakmai köztudatba.

Mivel is foglalkozik a nyelvtechnológia? A nyelvtechnológia a nyelvhasználatból indul ki, azaz konkrét szöveggel, konkrét beszéddel foglalkozik: bátran felvállalja tehát – az elméleti nyelvészet által leírni szándékozott kompetenciával szemben – a performancia vizsgálatát. Jellemzően a kidolgozott eljárások, technológiák valamilyen alkalmazás céljából (pl. gépi fordítás, beszédfelismerés) születnek, ezért a nyelvnek olyan szempontú vizsgálata is megjelenik az írások között, amely az elméleti nyelvészetben ritka vagy ismeretlen, hiszen az elméleti nyelvészek számára egyszerűen nem vetődnek fel ezek a kérdések. Gondolunk itt például a szófaji egyértelműsítés problémáira, amely a hagyományos nyelvészet-

ben nem is létezik. Azért nem, mert az ember óhatatlanul használja teljes nyelvi tudását és világismeretét a szöveg értelmezésében, és nemcsak az adott mondat betűire hagyatkozik a többértelmű kifejezések kezelésében. Összességében is igaz: a nyelvtechnológia számos olyan jelenséggel foglalkozik az emberek nyelvi kompetenciáját megközelítő pontossággal és hatékonysággal, amelyet elméleti nyelvészek triviálisnak tartanak, vagy ami egyáltalán nem jelenik meg számukra problémaként.

A nyelvtechnológia központi kihívása az, hogy a számítógépek számára tegye érthetővé és értelmezhetővé az emberi nyelvet, azaz – ha úgy tetszik – a legszigorúbb módon valósítsa meg a generatív nyelvészetben Chomsky által meghirdetett programot: egyfajta explicit nyelvleírást szorgalmaz, amely nem támaszkodik az emberi intuícióra a jelenségek értelmezésében. Az explicit és nem explicit nyelvi leírás különbségének illusztrálására említhetjük a szótárak példáját. Manapság már a legtöbb szótár digitális technológiával készül, és szinte mindegyik elérhető elektronikus adathordozón vagy a világhálón. Ettől azonban a tartalma, azaz az adatok megjelenítése változatlanul „emberi fogyasztásra" szolgál, azaz igen nagy mértékben támaszkodik a szótár olvasóinak nyelvi intelligenciájára (hogy mindazon prezentációs fogások dekódolási készségéről ne is beszéljünk, amelyeket részletesen sorolnak a szótárak előszavai). Ezek a szótárak azonban közvetlenül nem alkalmasak arra, hogy számítógépes nyelvfeldolgozó rendszerek szótári komponensei legyenek: az ilyen szótáraknak a nyelvtechnológia számára történő átalakítása jelentős erőfeszítést és megfelelő nyelvtechnológiai előképzettséget kíván.

Izgalmas probléma, hogy a nyelvtechnológia mennyire alkalmas „ellenőrző" eszköze az elméleti nyelvészet nyelvleírásának. Mint említettük, a nyelvtechnológia abból a szempontból az elméleti nyelvészet számára is kihívást jelent, hogy a lehető legexplicitebb leírásra kényszeríti a nyelvészt. Ha a gép, azaz nem az ember a készülő grammatika felhasználója, akkor például a „stb."-vel végződő felsorolásoknak nem lehet helye a nyelvleírásban. Ugyanakkor nem állíthatjuk, hogy a nyelvtechnológiai alkalmazás eredményessége egyben a nyelvelméletek közvetlen validálásának mércéje lehetne; már csak azért sem, mert nyelvi kompetencia tekintetében a jelenlegi legösszetettebb szuperszámítógép teljesítménye is messze elmarad az emberi agyétól. Bizonyosra vehető, hogy ez alapvetően nem a kapacitás, hanem az eltérő felépítés miatt van. A nyelvtechnológia közvetlenül tehát nem tűzi ki a nyelvi kompetencia modellezését, azaz nem akarja feltétlenül a beszéd- és nyelvhasználat mentális folyamatait leképezni a nyelvtechnológiai algoritmusokban.

Bár a nyelvtechnológiának sokszor a gyakorlati kényszer szülte alapelvei kezdetben ellentmondtak az uralkodó elméleti nyelvészeti felfogásnak (például

erős empirikus irányultsága, a nyelvhasználat vizsgálata, vagy a jelenségek gyakoriságára épülő statisztikai módszerek alkalmazása miatt), ma már ezek lépésről lépésre tért hódítanak az elméleti nyelvészet keretein belül is. Az utóbbi időben egyre inkább terjedő gépi tanulásos módszerek népszerűsége ellenére korántsem akarnánk azt állítani, hogy a nyelvész intuíciójának semmi szerepét nem látjuk a nyelvtechnológiában. Éppen ellenkezőleg: a nyelvtechnológiát az különbözteti meg az általában vett számítógépes adatkezeléstől, hogy a nyelvészet elvi felismeréseit építi be a technológiákba. Az például, hogy egy szociolingvisztikai kérdőív adatait Excel-táblákban vagy valamilyen adatbázis-kezelő program segítségével tároljuk, nem több mint számítógéppel segített szociolingvisztikai kutatás. Senki nem tekinti számítógépes nyelvészetnek azt, ha kedvenc példamondatainkat számítógépes fájlokban tároljuk, és onnan másoljuk be a szövegszerkesztővel készített tanulmányunkba. Az viszont már nyelvtechnológia (még ha megint csak egy látszólag triviális problémát old is meg), ha egy szöveg és annak idegen nyelvi fordítása között meg akarjuk találni a mondatok szintjén a fordítási megfeleléseket. Lehet, hogy ehhez kezdetben egy olyan egyszerű algoritmust használtunk, hogy a rövid mondatok fordítása is várhatóan rövid lesz és fordítva. Ugyanakkor ez az egyszerű elv is meglepően hatékonynak bizonyult a párhuzamos korpuszok mondatszintű illesztésében, ami viszont a ma már tömegesen használt, statisztikai gépi fordító rendszerek kifejlesztésében alapvető szerepet játszik. Ez utóbbi példa rávilágít a nyelvtechnológia társadalmi hasznosságára és küldetésére. A szövegszerkesztőkben használt helyesírás-ellenőrzők, a gépi fordítás, vagy akár a felolvasó- és beszédfelismerő programok mind bevonultak mindennapjaink számítógépes eszköztárába. Ily módon a nyelvtechnológia kiválóan alkalmas arra, hogy a társadalom széles körében érthetővé és hasznossá tegye a nyelvészetet, amelyet sokan egyébként elég elvont diszciplínának tartanak.

A kötet tanulmányainak válogatásában az egyik rendező elv az volt, hogy reprezentatív áttekintést adjunk a magyar nyelvtechnológia jelenleg használt módszereiről és eredményeiről. Fontosnak tartottuk azonban azt is, hogy olyan kutatásoknak is adjunk teret a kötetben, ahol maga a szerzői gárda vagy az olvasó könnyen eljuthat olyan konklúziókra, amelyek már túlmutatnak a pusztán adat alapú, gyakorlatorientált projekteken.

Prószéky Gábor kötetindító tanulmánya *A magyarországi számítógépes nyelvészet történeti áttekintése* címmel azt a folyamatot vázolja, amely Magyarországon már a hatvanas évek elején az akkori gépi fordítási munkálatokba való bekapcsolódással elindult, és többszöri megszakítással, hol számítógépes nyelvészet, hol természetesnyelv-feldolgozás néven élte túl a 20. század utolsó évtizedeinek hazai kutatás-fejlesztési nehézségeit. A kitartó kutatók munkája végül is

azokhoz a magyar nyelvtechnológiai eredményekhez vezetett, amelyeket ma már nemzetközileg is számon tartanak.

Az említett történeti folyamat utolsó időszakának, a 21. század első évtizedének eredményeit mutatja be tehát a kötet, mégpedig többé-kevésbé a gépi nyelvfeldolgozási szintek szerinti elrendezésben. Elsőként így egy szóalaktani problémákkal foglalkozó írás szerepel benne: Rebrus Péter, Kornai András és Varga Dániel *Egy általános célú morfológiai annotáció* című dolgozata a nyelvtechnológiában kulcsfontosságú szóalaktani annotációs sémák problémáival foglalkozik, majd a magyar főnévi, igei és egyéb inflexiós paradigmák ezek segítségével való kódolását tárgyalja részletesen. A leírás alapelvei nemcsak teljesen általánosak és nyelvfüggetlenek, hanem a gyakorlatban, a szabadon elérhető *hunmorph* programban is megtalálhatók.

Ezt követően egy, a szintagmaszint problematikájával foglalkozó írás következik: Recski Gábor és Varga Dániel *Magyar főnévi csoportok azonosítása* címmel az ún. NP-darabolóról (angolul *NP chunker*) ír, amely magyar nyelvű főnévi csoportok azonosítását teszi lehetővé itt éppen egy felügyelt gépi tanulási módszer segítségével. A módszer a gyakorlatban *hunchunk* néven érhető el.

Ezt követően a főnévi csoportokéhoz hasonló problematikájú névkifejezések kezelése következik: Vincze Veronika és Farkas Richárd *Tulajdonnevek a számítógépes nyelvészetben* című írása az angolul *named entity recognition* néven ismert problémakörrel, a névelem-felismeréssel foglalkozik. A tulajdonnevek és más szövegbeli entitások, például email-címek, weblapok, rendszámok, telefonszámok, dátumok, vagy orvosi-biológiai szövegekben fehérjenevek, génnevek, kémiai szövegekben a vegyületek neveinek és képletének felismerése is ide tartozik. Sokszor a felismerésen túl további – a szöveg tartalmától is függő – belső osztályozást is illik adni a megtalált elemeknek, hiszen például a jogi szövegekben előforduló személynevek igen különböző szerepeket testesíthetnek meg a bírótól a vádlottig.

A névelem-felismerésben használt gépi tanulási eljárások gyakran induktív módszereken alapulnak. Kálmán László tanulmánya, az *Analógiás tanulás asszociatív memóriamodellel* ezzel szemben egy abduktív eljárást mutat be, azaz egy következtetés konklúzióját reprezentáló formulahalmazhoz keres minél nagyobb konzisztens premisszahalmazt egy olyan adatbázisban, amelyben különböző valószínűséggel igaznak tekinthető formulák vannak tárolva. Ez az adatbázis nem más, mint a korábbi tapasztalatokat tároló memória. Kálmán kutatásának legfőbb eredménye, hogy laboratóriumi méretekben sikerült egy olyan memóriaalapú modell alapjait lefektetni, amely hosszabb távon képes lehet megragadni a nyelvi viselkedés legáltalánosabb mechanizmusaira jellemző folyamatokat, és

így alapjául szolgálhat a nyelvi produkció és a nyelvi megértés minden eddiginél hatékonyabb szimulációinak.

A mondatszintű leírással kötetünkben Alberti Gábor, Károly Márton és Kleiber Judit *A mondatoktól a hatóköri relációkig – és vissza* című munkája foglalkozik. Ők a magyar kijelentő mondatok információszerkezetét tárják fel gépi módszerekkel, az – esetükben totálisan lexikalista irányultságú – generatív grammatika alapelvei mentén. A bemenő betűsorhoz minden lehetséges intonációs mintázatot hozzárendelnek, így igyekeznek az írott bemenetnek a hangzó beszéddel való kapcsolatát is kezelni. Mivel kutatásuk távlati célja a gépi fordítás, az ellenkező iránnyal, az információszerkezetből intonációs jelekkel ellátott mondatot előállító algoritmussal is foglalkoznak, amire a dolgozat címe is utal.

A mondatszintaxis tárgyalása a modern nyelvészet központi kérdéskörét jelenti, ám a nyelvtechnológiai kutatásokban, bár ez a nyelvi szint is fontos, nem feltétlenül játszik központi szerepet. A következőkben tárgyalandó koreferenciaviszonyok kilépnek a mondatszintről. Ráadásul itt már megjelenik a tudásalapú közelítés is, vagyis a szemantika és a világismeret bevonása a gépi elemzésbe. Miháltz Márton *Tudásalapú koreferencia- és birtokviszony-feloldás magyar szövegekben* címmel arról ír, hogy milyen gépi algoritmusokkal lehetséges a szövegbeli entitások közötti kapcsolatok – koreferenciaviszonyok, birtokviszonyok – automatikus felismerése. Ennek a problémának a megoldása gyakorlatilag a nyelvtechnológia minden területén (a gépi fordításban, az információ-kivonatolásban, a szöveg-összefoglalásban, vagy a véleményanalízisben) egyaránt fontos. Főnévi csoportok koreferenciáinak feloldásán az egy dokumentumban megjelenő különböző, de a világban azonos entitásra referáló főnévi csoportok közötti viszonyok azonosítását értjük. A birtokviszony-feloldás az egymástól a mondatban különvált birtokos szerkezet birtokosának és birtokának felismerését és párosítását jelenti. Ezekre a feladatokra ad algoritmikus megoldást a dolgozat.

A jelentéssel kapcsolatos ismereteket a nyelvtechnológiai eszközök az ezeket is leíró gépi lexikonok világából szerzik be. Héja Enikő és Gábor Kata *Igék lexikai reprezentációja és a nyelvtechnológia* címmel arról ír, hogy milyen elvárásoknak kell megfelelnie a nyelvtechnológiai alkalmazások igei lexikonjának. Elvárható, hogy egy ilyen adatbázisban az ugyanolyan típusú dolgok ugyanúgy legyenek reprezentálva, azaz a lexikonnak koherensnek kell lennie. Másfelől a lexikai adatbázisnak explicitnek is kell lennie, vagyis nem támaszkodhat a felhasználó intuíciójára. A szerzők körüljárják, hogy hogyan határozható meg a produktív igei bővítmények köre. Ezáltal a többek között általuk korábban kidolgozott igei vonzatkeret-adatbázist olyan információkkal bővítik ki, amelyek segítségével hasznos általánosítások tehetők az igék bővítménykeretére vonatkozóan, így növelve az adatbázis koherenciáját és explicitségét.

Az ezt követő négy dolgozat a modern nyelvtechnológia legújabb alkalmazási területeit villantja fel: a szociolingvisztikát (ahol élőbeszéd-átiratok segítségével valós beszédhelyzetek számítógépes elemzése történik); a metaforikus nyelvhasználat gépi kezelését; az ember és a gép közötti kommunikáció különféle aspektusait nyelvtechnológiai szempontból vizsgáló kutatást; végül pedig a nyelvtechnológiának a nyelvtörténeti kutatásban való felhasználását.

Váradi Tamás, Oravecz Csaba és Peredy Márta *A Budapesti Szociolingvisztikai Interjú lexikai és szintaktikai jellemzői* című tanulmányának célja a magyar nyelvű társalgási szövegek lexikai és szintaktikai elemzése nyelvtechnológiai módszerekkel és ennek segítségével a szóbeli és írásbeli nyelvhasználat közötti különbségek kvantitatív megfogalmazása. Az elemzőprogram a számítógépes elemzéssel annotált szövegkorpuszt elsősorban statisztikai eljárásokkal vizsgálja. A BUSZI társalgási nyelvhasználatát a szerzők a Magyar Nemzeti Szövegtárból vett minta segítségével az írott nyelvhasználat jellemzőivel vetik össze. Az ismertetett vizsgálatok a magyar nyelvre még nagyrészt feltáratlan lehetőségeket mutatják be, azaz elsősorban a kezdetet jelentik ezen a gépi eszközökkel korábban nem kutatott területen.

Babarczy Anna és Simon Eszter *A fogalmi metaforák és a szövegstatisztika szerepe a metaforák felismerésében* című munkája a metaforikus kifejezések automatikus számítógépes felismerését vizsgálja. Az emberi metaforaértelmezés két elméleti modelljét, a fogalmimetafora-elméletet és a statisztikai megközelítést vetik össze. A két elmélet alapján pszicholingvisztikai és korpusznyelvészeti módszerek felhasználásával a metaforikus használatra utaló nyelvi jelek listáit hozták létre, majd ezek valós metaforajósló erejét számítógépes modellel tesztelték. Az eredmények alapján a statisztikai módszer tűnik a legsikeresebbnek, bár ennek a teljesítménye is elmarad a várakozásoktól, nagy valószínűséggel a metafora jelenségének megfoghatatlansága, illetve magának a fogalommeghatározásnak az elméleti pontatlanságai miatt.

Simon Eszter és Sass Bálint tanulmánya *Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből* címmel szerepel a kötetben. A nyelvi kulturális örökség széles körű elérhetővé tételében manapság világszerte kulcsszerep jut a nyelvtechnológiának. A gépi módszerekkel a kutatók eddig nem látott, egységes, következetes, rengeteg kiegészítő nyelvi információval ellátott adatbázisokhoz juthatnak. A dolgozatban bemutatásra kerül a nyelvtörténészek és a nyelvtechnológusok első hazai közös kutatási területe, a történeti szövegkorpuszok építése. Ezek segítségével a kutatók egységes, akár egy egész korra jellemző, átfogó keresési eredményekhez is juthatnak, amelyekkel elméleti feltevéseik könnyebben igazolhatóvá válnak. A minderre kiváló terepet szolgáltató ómagyar

nyelvtörténeti szövegadatbázis létrehozásának és a hozzá tartozó gépi lekérdező eszközök alkalmazásának problematikájáról esik szó az írásban.

Az ember–gép kommunikáció elméleti-technológiai modellje és nyelvtechnológiai vonatkozásai címmel Hunyadi László, Földesi András, Szekrényes István, Staudt Alexandra, Kiss Hermina, Abuczki Ágnes és Bódog Alexa számol be a HuComTech korpusz létrehozásáról, amelynek a motivációja az volt, hogy létrejöjjön az ember–gép kommunikáció olyan technológiai modellje, amely alapvetően épít az ember–ember kommunikáció lényeges és e feladat szempontjából releváns jellemzőire. A modell fontos tulajdonsága, hogy kétirányú, azaz egyaránt szolgálja a szintézist (egy kommunikatív esemény technológiai megvalósítását) és az analízist (ezen esemény interpretációját, „megértését"). Ráadásul lehetővé teszi e két, ellentétes irányú folyamat egyidejű kezelését is, miáltal alkalmassá válik az ember–gép kommunikáció kétirányú folyamatának egységes kezelésére. A tanulmány az ehhez szükséges multimodális (video-, akusztikai, tekintet-, gesztikuláció-, szintaktikai és pragmatikai) annotálási folyamatot mutatja be, valamint az adatbázis lekérdezése alapján már elérhető egyes eredményeket.

A kötetet a beszédtechnológiai terület két kutatásának összefoglalója zárja: az egyik a beszéd gépi felismerésében elengedhetetlen akusztikus modellek, a másik a beszélt nyelvi adatbázisok létrehozásának problémakörét járja körül.

Tóth László *Kísérletek beszédfelismerők akusztikus modelljének nyelvek közötti átvitelére* című dolgozata a szokásos beszédhang-alapú beszédfelismerőkkel szemben a fonológiai megkülönböztető jegyekre épülő módszereket járja körül. Mivel a megkülönböztető jegyek jóval univerzálisabbak és kevesebben vannak, mint a beszédhangok, így a hipotézis az, hogy ezekre építve jóval könnyebb és hatékonyabb nyelvfüggetlen akusztikus modellt készíteni. A szerző angol nyelvre betanított rendszerekből készített két magyar nyelvű akusztikus modellt, ahol az eredeti, angol felismerő az egyik esetben beszédhangok, a másik esetben megkülönböztető jegyek felismerésére volt betanítva. Eredményei meglepőek, ugyanis egyik angol nyelvről átültetett modell sem éri el a tisztán magyar tanítású modell teljesítményét. Így nem teljesül tehát az a remény, hogy a nagy mennyiségű adaton tanított angol modellekből kiindulva elkerülhető, hogy a magyarra is hasonló hatalmas korpuszokat kelljen összegyűjtenünk. A szerző érdekes általános konklúzióra jut a gépi tanulási módszerekkel kapcsolatban, ha ezek – mint a bemutatott kutatásban is – az intuíciónak ellentmondó eredményeket adnak: ilyenkor sokszor nem az alapkoncepcióval van a baj, hanem a tanulóalgoritmus paramétereivel, modellválasztásával, optimumkeresési módszerével, vagy egyéb technológiai tényezővel.

Gósy Mária dolgozata egy *Multifunkcionális beszélt nyelvi adatbázis*, a számos tekintetben nemzetközileg is jelentős BEA munkálatait foglalja össze. Ez

az első sok beszélővel rögzített, nagy mennyiségű hangzó anyagot és különböző szintű átiratukat, illetve annotálásukat tartalmazó adatbázis, amelynek a felvételi körülményei állandóak. A jól megtervezett és kivitelezett, annotált és lekérdezhető adatbázis kiváltja az időigényes felvételek készítésének munkáját, hatalmas adathalmazt biztosít sokféle kutatáshoz, és a nyelv valós használatát tükrözi. A BEA adatbázis révén magyar nyelven először vált lehetővé az összes magánhangzó akusztikai-fonetikai szerkezetének leírása, a koartikulációs mezők jellemzése, a beszédhangok semlegesedésének, a gyakori szavak ejtési sajátosságainak, a zöngeminőség kommunikációs funkcióinak az elemzése, avagy a prozódia szerepének vizsgálata a spontán beszéd tagolásában. A szoros értelemben vett fonetikai kutatások mellett a szerző számos, a BEA segítségével lehetővé váló új kutatási irányra is ráirányítja az olvasó figyelmét.

* * *

A szerkesztők köszönetüket fejezik ki mindazoknak, akik hozzájárultak az *Általános Nyelvészeti Tanulmányok* nyelvtechnológiával foglalkozó XXIV. kötetének létrejöttéhez. A lektorok figyelmes munkája és a szerzők türelmes együttműködése következtében ez a kötet, még ha nem is az eredetileg elképzelt sebességgel, de végül is az eredeti elveknek megfelelően készülhetett el. Külön köszönet illeti Kenesei István sorozatszerkesztőt és Siptár Péter technikai szerkesztőt, valamint Pintér Tibort, aki lelkiismeretes szervezőmunkájával járult hozzá a kötet létrejöttéhez.

Prószéky Gábor, Váradi Tamás

A magyarországi számítógépes nyelvészet történeti áttekintése

Prószéky Gábor

MTA–PPKE Nyelvtechnológiai Kutatócsoport, PPKE ITK & MorphoLogic, Budapest
proszeky@morphologic.hu

Összefoglaló néven nyelv- és beszédtechnológiának hívják manapság azt a komplex tudományterületet, amely a számítógép és az emberi nyelv, illetve az emberi beszéd kapcsolódási pontján alakult ki. Korábban ezt számítógépes nyelvészetnek nevezték. Tanulmányunk felépítése kutatási témánként igyekszik – amennyire lehetséges, azon belül időrendben – követni a hazai nyelvtechnológiai tevékenységeket. A bevezető rész után áttekintjük a számítógépes morfológia és a gépi szintaxis hazai kutatási eredményeit, ezután végigvesszük a korpusznyelvészeti kutatásokat, majd a számítógépes lexikográfiával, végül a gépi fordítással kapcsolatos kutatások hazai helyzetét. A korai időszak áttekintésében Prószéky (1989)-re, a későbbiekében Prószéky–Olaszy–Váradi (2006) tanulmányára támaszkodtunk.

Kulcsszavak: számítógépes nyelvészet, történeti áttekintés, beszéd- és nyelvtechnológia, magyar nyelvtechnológiai alkalmazások, a nyelvi rendszerek

Annak ellenére, hogy az angol az utóbbi évtizedekben egyeduralkodó világnyelvvé lett, a nemzeti nyelvek és kultúrák szerepe egyértelműen felértékelődött az informatikában. A magyar nyelvtechnológiai kutatások eredményeképpen létrejött nyelvi szoftvereszközöket ma már többszázezren használják naponta, és hatásuk a magyar nyelvhasználókra – ennek következtében a magyar nyelv jövőjére – lényegesen nagyobb, mint gondolnánk.

1. A kezdetekről

A számítógépes nyelvészeti kutatások Magyarországon gyakorlatilag már a számítógép hazai megjelenésekor elindultak. 1958 őszétől Fodor István, Papp Ferenc, Tarján Rezső és Szalai Sándor többször is tartottak előadást a gépi fordításról a Nyelvtudományi Társaságban és az MTA Nyelvtudományi Intézetében. 1960-ban a gépi fordítás előkészítése az MTA távlati terveibe is bekerült. Ennek az évnek a végén lezajlott az első magyarországi interdiszciplináris értekezlet is

a nyelvészek, logikusok és – az akkor megjelenő névvel kibernetikusoknak nevezett – számítógépesek részvételével. Az első gyakorlati eredmény e téren: Hell György és Sipőczy Győző a BME-n a Vezetékes Híradástechnika Tanszék jelfogós gépén magyarra fordított egy orosz mondatot. 1962 elején Hell György – Dömölki Bálint segítségével – megkezdte az első orosz–magyar gépi fordító algoritmus alapjainak kidolgozását az MTA Számítóközpont M-3 számítógépén. A kísérletek ellenére a magyarról vagy magyarra való fordítás átfogó leírásáról ez időben nem jelent még meg komoly publikáció. 1962-ben két fontos tanácskozás is volt hazánkban: Budapesten az MTA munkaértekezlete Kalmár László vezetésével *A matematikai nyelvészet és a gépi fordítás kérdései* címmel, valamint Tihanyban *A matematika alapjai, matematikai gépek és alkalmazásaik* konferencia *Matematikai nyelvészet és gépi fordítás* szekcióülése. Ebben az időben Budapesten, az MTA Számítóközpontjában, a BME Gépészkari Idegennyelvi Lektorátusán és Debrecenben, a KLTE Szláv Filológiai Intézetében folytak számítógépes nyelvészeti kísérletek. Ezekről elsősorban az *Általános Nyelvészeti Tanulmányok* II. számában és főként különböző könyvtári feldolgozásokkal kapcsolatos kiadványokban, egy-két alkalommal a *Magyar Nyelvőr* hasábjain, valamint 1963-tól kezdve az MTA Számítóközpont által többé-kevésbé évente megjelentetett angol nyelvű kiadványban, a *Computational Linguistics*-ben olvashattak az érdeklődők. Az MTA Számítóközpontjában 1966-ban – az Egyesült Államokban folyó gépi fordítási kutatások nagy részének leállítását kezdeményező ALPAC-tanulmány megjelenésével egyidejűleg – a gépi nyelvészeti munkacsoport átalakult, és Dokumentációs Nyelvészeti Csoport néven, megváltozott összetételben már csak részben folytatta a jogelődje által megkezdett munkát. 1967-ben a csoport és az OMKDK közös rendezésében *MASPEREVOD-67* néven sor került a szocialista országok első gépi fordítási találkozójára is. 1968 őszén Balatonszabadiban volt egy matematikai nyelvészeti konferencia, amelyen a csoportban folyó szintaktikai kutatásokról szintén hangzott el előadás.

A magyar számítógépes nyelvészeti törekvések megbecsülését is jelentette, hogy 1971-ben a kétévenként megrendezésre kerülő Nemzetközi Számítógépes Nyelvészeti Kongresszus (a későbbi COLING világkonferenciák elődje) színhelye a téma kutatásában élen álló Grenoble és Stockholm után Debrecen lett. 1966-tól kezdve ugyanis a város egyeteme egyre inkább a számítógépes nyelvészet egyik – elsősorban a filológiai munkákhoz elengedhetetlen – ágának, a Papp Ferenc nevével fémjelzett számítógépes lexikológiának a bölcsőjévé vált. A Dokumentációs Nyelvészeti Csoport felszámolásával az MTA Számítóközpontjában minden számítógépes nyelvészettel kapcsolatos szervezett munka megszűnt a hetvenes évekre. Az MTA Számítóközpont jogutódjaként működő MTA SZTAKI kiadványaként időnként megjelent az átalakított *Computational Linguistics and*

Computer Languages kiadvány, de az is inkább a formális nyelvekkel kapcsolatos kérdésekre helyezte a hangsúlyt. A hetvenes években Debrecenben a Papp Ferenc vezetésével működő kutatócsoport már elsősorban nem a szövegfeldolgozás szempontjából jelentős, hanem az irodalmár-filológus kutatók igényeinek jobban megfelelő kvantitatív nyelvészeti, illetve kimondottan lexikológiai feldolgozásokra összpontosított. A magyar számítógépes nyelvészeti kutatások nyelvészeti szempontból legjelentősebb kiadványa, a Papp Ferenc által írt *A magyar főnév paradigmatikus rendszere* (Papp 1975) épp ebben az időszakban jelenik meg, bár az alapjául szolgáló számítógépes munka az előző korszak eredményeit idézi. A debreceni csoport figyelme a hetvenes évek végétől inkább a nyelvoktatás számára használható számítógépes programok irányába fordul. Ez idő tájt az MTA Nyelvtudományi Intézetében is szinte kizárólag kvantitatív jellegű számítógépes munkálatok *(A magyar köznyelv és irodalmi nyelv gyakorisági szótára)* folynak egészen a hetvenes évek legvégéig, amikorra számítástechnikai eszközeink hardver és szoftver tekintetében egyaránt elérték azt a szintet, hogy az Európa Magyarországon kívüli részében mindenhol elterjedt nyelvfeldolgozó rendszerek (természetes nyelvű adatbázis-lekérdezés, szövegkivonatolás, dialógusrendszerek) hazai megvalósításának legalább a lehetősége felmerüljön. Így kerülhetett sor – az SZKI, és ezen belül is a korábbi gépi fordító csoport valahai tagjának, Dömölki Bálintnak a támogatásával – a mesterséges intelligencia céljaira fejlesztett és a hazai számítástechnikai életben nagy sikerrel bevezetett programozási nyelv, a Prolog kezdeti alkalmazásai közt számítógépes nyelvészeti kutatásokra is. A nyolcvanas években sikerült néhány korábbi számítógépes nyelvészeti anyagot „újraéleszteni" Kornai Andrásnak és Prószéky Gábornak (Kornai 1986; Papp 2000), akik ez idő tájt még inkább elméleti munkásságot folytattak. Az ő neveikhez fűződik egyébként az első átfogó hazai számítógépes nyelvészeti könyv (Prószéky 1989), illetve a nemzetközi matematikai nyelvészeti kutatások elmúlt évtizedeinek összefoglalása is (Kornai 2007).

Budapest-központú országunkban – mint jeleztük – a számítógépes nyelvészeti kutatások területén korábban Debrecen játszotta a legfontosabb nemfővárosi kutatóhely szerepét. Ma is folynak ott ilyen irányú kutatások (Hunyadi 2011), de a kilencvenes években – elsősorban Csirik János kutatócsoportja munkájának következtében – a Szegedi Tudományegyetem vált a legismertebb nem budapesti nyelvtechnológiai központtá. Egy másik, nagy múltú egyetemi városunkban, Pécsett Alberti Gábor munkatársaival szintén a kilencvenes években alakította ki a magyar gépi nyelvészet egy újabb központját. 1991-ben összeállt a hazai gépi nyelvészet első magánvállalkozása, a MorphoLogic, mely a kilencvenes évektől meghatározó szerepet játszott a hazai kutatásokban. A 2000-es évektől a kiemelt kutatás–fejlesztési témák közé bekerült a nyelvtechnológia is.

Eleinte elsősorban a fent említett magánvállalkozás, az MTA Nyelvtudományi Intézete, valamint az SZTE[1] Informatikai Tanszékcsoportja kutatóinak együttműködésében valósult meg több alapvető szövegnyelvészeti, illetve elsősorban a BME TMIT-en[2] néhány alapvető beszédtechnológiai kutatás. Az évek folyamán további szereplők jelentkeztek: az elsősorban a mesterségesintelligencia-alkalmazásaival híressé lett AITIA és ALL, vagy a fordítástámogató szoftvereszközök fejlesztésére koncentráló Kilgray magánvállalkozások, illetve az egyetemi kutatóhelyek közül pedig elsősorban a BME MOKK,[3] majd a PPKE ITK.[4] A BME-n, az SZTE-n és a PPKE ITK-n egyébként a 2000-es évektől a nyelvtechnológiai tárgyak szerves részét alkotják a BSc-, MSc- és PhD-programoknak. 2007-ben megalakult a Nyelv- és Beszédtechnológiai Platform, amely indulásakor nyolc (később további tíz) ipari és kutatási partnert tömörítő érdekképviseleti társulás volt, annak érdekében, hogy előmozdítsa a hazai nyelv- és beszédtechnológia fejlesztését és a már meglévő eszközök használatát, illetve jövőképet mutasson a nyelv- és beszédtechnológia mint leendő iparág számára.

2. A magyar számítógépes morfológia eredményei

A magyar nyelv grammatikájának viszonylag legkönnyebben – de semmiképpen sem könnyen – számítógépesíthető része a morfológia. Mivel a kétszintes morfológiai modell (Koskenniemi 1983) megjelenéséig nem volt olyan eszköz, amely egymaga használható lett volna elemzésre és generálásra egyaránt, a magyar morfológiai programok is két családra oszlanak, a szintetizálókra és az elemzőkre. A morfológiai szintetizáló rendszerek a magyar szóalakok esetében a két nagy szófajosztály, a névszók és az igék automatikus toldalékolását végző programok gyűjtőneve. A kétféle rendszer nem pusztán a toldalékok különbözősége miatt válik el egymástól – különösen mivel a toldalékok egy része (a birtokos személyragok és az igei személyragok) nem is különböznek –, hanem az igazi különbség a névszók toldalékolásának meglehetősen tisztán agglutináló és az igei toldalékok összemosódott, nehezen kielemezhető voltában van.

Az első komolyabb gépi morfológiai kísérlet hazánkban Vargha Dénes nevéhez fűződik: az ő szótővezérelt, a Dömölki-szűrőre (Dömölki 1964) épülő szukcesszív behatárolás módszerével működő morfológiai elemzése magyar szóala-

[1] Szegedi Tudományegyetem

[2] Budapesti Műszaki Egyetem, Távközlési és Médiainformatikai Tanszék

[3] Budapesti Műszaki Egyetem, Média Oktató és Kutató Központ

[4] Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar

kokat gyakorlatilag nem is elemzett, csak orosz nominális formákat (Vargha 1963). Sorra vette a vizsgálandó objektum, például egy szótári tő morfológiai tulajdonságait, majd megállapította, hogy a kívánt toldalékok által meghatározott grammatikai kategóriák közül melyek egyeztethetők össze velük. Így lépésenként, szukcesszíve szűkül le a vizsgált objektumokra vonatkoztatható kategóriák halmaza, míg elő nem áll a legszűkebb olyan halmaz, amelybe a vizsgált objektum még beletartozik. Kónyi (1965) a magyar főnevek gépi elemzéséről szóló írásában felsorolta a magyar főnévtípusok teljes paradigmáit. Nagyon fontos megjegyezni, hogy nem a tövek alakja vagy változása, hanem a paradigmák különbözősége szolgáltatta ezeket a típusokat. Melcsuk (1967) magyar főnevek szintézisét végző modellje egy tőhöz 842 paradigmatikus alakot volt képes előállítani. Klauszer (1965) a magyar főnevek szintézisét a Papp-féle Szóvégmutató Szótár elkészülte előtt nem alapozhatta a teljes nyelvi anyagra, hanem csak egy korábbi – a Nemes-féle gyakorisági szótár (Nemes 1941) segítségével kiválogatott –, kb. 700 elemet tartalmazó szójegyzékre. Így az ebből elvonatkoztatott „törvények" nem voltak maradéktalanul helyesek, de arra mindenképpen jók voltak, hogy későbbi szintéziskísérletek alapjául szolgáljanak. A rendszer a főnevek egyes szám tárgyesetbeli, többes szám alanyesetbeli és a birtokos ragozás egyes szám 3. személyű toldalékai tőhöz való kapcsolódásának megfigyelésein alapul. Stein főnévszintetizáló rendszere szintén a debreceni számítógépes nyelvészeti munkacsoport munkájának eredménye volt (Stein 1966). Jánoska igeszintézise a *Szóvégmutató Szótár* (Papp 1969) igető-alaptípusaira épül, bár annak könyv formában való megjelentetése előtt készült el szintén a debreceni számítógépes nyelvészeti munkacsoport kutatásaként (Jánoska 1967). A csoport vezetője, Papp Ferenc 1966-os főnévszintetizálási elképzelésével – minden hiányossága ellenére – jó kiindulópontot szolgáltatott ahhoz, hogy 1975-re a kutatás beérjen, és a szerző a kor legtökéletesebb algoritmusának leírásaként közölje. Ez a megoldás az Értelmező szótár teljes anyagára épített gondos elemző munka eredményeként (Papp 1975) már mentes volt a korábbi gépi morfológiai modellek hibáitól. A Papp-féle modell tulajdonképpen nem is tövekre, hanem a lehetséges szótövek alapjául szolgáló három bázisra épül: ezek segítségével egy ragozási típusba azok a szótövek tartoznak, amelyek ugyanazon sorszámú báziseleméhez a megfelelő toldalékmorfémák azonos allomorfjai kapcsolódnak. Egyes toldalékok csak egy-egy konkrét bázishoz járulhatnak, de vannak különböző tőtípusok esetén különböző bázishoz kapcsolódó toldalékok is. A tőtípusra jellemző, hogy melyik bázisa milyen jellegű toldalékok felvételére alkalmas. Vásárhelyi (1975) igeszintézise a Vargha Dénes által kidolgozott szukcesszív behatárolás módszerén alapult (Vargha 1963). Lugosiné 1975 igeszintetizáló modellje a személyragos alakokon és igeneveken kívül még a ható, műveltető és szenvedő alakok, ill. ezek továbbtoldalékolt formáinak előállítására

is alkalmas volt. Az Elekfi-féle alaki rendszer az Értelmező szótár igéit – egymástól tulajdonképpen sokszor csak minimálisan eltérő – ragozási típusokba sorolja, így a rendszer segítségével kapott toldaléktömbök közvetlenül a tőhöz járulnak, mindössze a hasonulás, a hangkiesés, illetve -beszúrás jelent apróbb nehézséget. Pajzs (1983) morfológiai szintetizáló programjának szótári információi szintén a tővariánsok ragozási típusba való sorolás nélküli előállítását szolgálják, csak itt a tőtípusba tartozás a „valódi" toldalék-előhangzókkal kiegészített tövet jelenti.

A magyar nyelvű szóalakok morfológiai elemzésére készült **GáZoLaj** modell (Prószéky et al. 1982) jobbról balra halad a szó belsejében. Az algoritmus nemcsak a további balra levő toldalékokra és a lehetséges tövekre tesz hipotéziseket, hanem ezek morfofonológiai tulajdonságaira is. A szabályok tulajdonképpen logikai állítások, és a megfelelő állítássorozat bizonyíthatósága jelenti a helyes morfológiai elemzést. A rendszer ilyenfajta interpretációját a megvalósítás nyelve, a Prolog logikai programnyelv ösztönözte (Sántáné-Tóth–Szeredi 1982). A GáZoLaj rendszer igealakok elemzését is végezte, ám mivel az igei paradigma egyes elemei nominális toldalékokat is felvehetnek, az igei végződéseknek így kapott két csoportját elkülönítették egymástól. A nem nominális igei toldalékok rendszere finitum végződésekből, ragozott és puszta infinitívuszokból, valamint határozói igenevekből áll. A finitum végződéseket és az infinitívuszi ragozást az eljárás komplex toldalékokként kezeli. Ennek oka a magyar igeragozás már többször említett, flektálóba hajló, a névszói ragozásnál kevésbé agglutináló jellegzetességeiben keresendő.

A mai napig a gépi morfológia területén a legátfogóbb szóalaktani rendszer a magyar nyelvhez az 1991-ben elkészített **Humor** *(High-speed Unification Morphology)* morfológiai elemző program volt. A rendszerhez egy leíró formanyelv is tartozott, mely a MorphoLogic cég első tudományos eredményének tekinthető (Prószéky–Kis 1999; Prószéky–Merényi 2012). Ennek a számítógépes szóalaktani rendszernek a kidolgozásához a magyar szavaknak olyan jellegű és részletességű osztályozása volt szükséges, amely korábban nem volt még kidolgozva (Prószéky 2000). A program belső összetevős szerkezet nélküli lapos morfsorozatokként elemzi a szavakat. Ennek az az oka, hogy a program reguláris szónyelvtant tartalmaz, amely egyfajta determinisztikus véges állapotú automataként van implementálva. Ez egyrészt nagy sebességet biztosít, másrészt elkerüli a sok irreleváns szerkezeti többértelműség előállítását, amit a megfelelő környezetfüggő elemző generálna, például a többszörösen képzett összetett szavak esetében. Az elemző olyan morfokat keres a szótárában, amelyeknek a felszíni alakja illeszkedik a megadott szó még elemzetlen részére. A lexikon nemcsak morfokat, hanem összevont morfsorozatokat is tartalmaz, amelyeket az elemző így egy lépésben ismer fel. Elemzés közben a program kétféle ellenőrzést hajt végre (gyakorlatilag

ez a program nevében szereplő unifikációs része a formalizmusnak): egyrészt lokális kompatibilitás-ellenőrzést végez az egymás mellett álló morfok között, azaz ellenőrzi a morfofonológiai és a lokálisan ellenőrizhető morfotaktikai feltételek teljesülését; másrészt azt is ellenőrzi, hogy az elemzést alkotó morfémák a nyelv lehetséges szókonstrukciói egyikét testesítik-e meg, azaz megfelelnek-e az adott nyelv morfológiai konstrukcióit leíró szónyelvtannak. A magyarban például a tő + képzők + ragok alakú morfémasorozatok jól formáltak, ugyanilyen kategóriájú morfémák más sorrendben azonban nem jók. A szónyelvtan nem szomszédos összetevők közötti megszorítások ellenőrzését is lehetővé teszi: pl. a *leg-* felsőfokjelet egy tőle jobbra álló morfémának (leggyakrabban a *-bb* középfokjelnek) engedélyeznie kell, közöttük azonban számos más morféma is állhat. A későbbiekben a formalizmushoz egy magas szintű leíró nyelv és az ebből a tényleges Humor-adatbázist előállító eszközkészlet is csatlakozott (Novák 2003; Novák–M. Pintér 2006). Ebbe a rendszerbe nagyon könnyen lehet új szavakat felvenni, mert csak azokat a megjósolhatatlan tulajdonságaikat kell megadni, amelyek eltérnek a szó alakjából következő alapértelmezett viselkedéstől. A Humor rendszer szóadatbázisa lefedi az Értelmező kéziszótár teljes szóanyagát, sőt mintegy ötvenezer további alapszóval gazdagítja is. A produktív toldalékolási és összetételi szabályok miatt a programrendszer – becslések szerint – több milliárd helyes magyar szóalak elemzésére képes, ugyanakkora helyigénnyel és ugyanolyan sebességgel, mint a néhány százezres adatbázisú nyelvekhez készített elemzőprogramok. A Humor rendszerhez kifejlesztett formanyelv – a magyar szóalaktan relatív bonyolultsága miatt – más nyelvekre is könnyen és eredményesen alkalmazható volt: a MorphoLogic nyelvi programtermékei ezt az elemzőmodult használták a lengyel (Wołosz 2005), a cseh, a román, az angol, a német, a francia és a spanyol esetén (Prószéky–Kis 1999). Az idők folyamán több kutatási pályázatban is uráli nyelvészek vezetésével a MorphoLogic leíró formalizmusát használva több kicsiny rokon nyelv (komi, udmurt, manysi, tundrai nyenyec, nganaszan stb.) morfológiájának leírása is megvalósult (Prószéky–Novák 2005).

A BME MOKK-ban kidolgozott **hunmorph** (Trón et al. 2005) nyílt forráskódú, nyelvfüggetlen morfológiai elemző helyesírás-ellenőrzésre, szótövesítésre és morfológiai elemzésre egyaránt használható. A hunmorph keretrendszer három fő részből áll: egy nyelvfüggetlen végződéskezelőből, egy lexikai adatbázisból (valójában egy morfológiai nyelvtanból) és egy magas szintű leíró formalizmusból, illetve az ennek működtetéséhez szükséges előfordítóból.

A nyelvtechnológiában van egy, a morfológiához szorosan kapcsolódó, ám az elméleti nyelvészetben nem szereplő terület, ami elméleti nyelvészeti körökben magyarázatra szorul: a szófaji egyértelműsítés. Ez a kategória azért nem létezik a nyelvtudomány más területein, mert az ember számára egy többértelmű

szó értelmezésekor mindig létezik olyan nyelvi szint, ahol csak egyetlen szófaji értelmezése van az illető szónak. A morfológiai többértelműségek kezelésében mindig segít a szintaxis, a szemantika vagy a pragmatika, vagy valami külső körülmény segítségével el tudjuk különíteni az egyik szófajt a másiktól (Prószéky 2012). A számítógépes módszerek sokszor nem lépnek át a magasabb nyelvi szintekre, de az adott nyelvi szinten elvárható volna tőlük a szöveg egyértelmű kódolása. A morfológiai elemzés több lehetséges felbontásából ki kell tehát választani azt az egyet, amely az adott környezetben szerepel. A magyar nyelvvel kapcsolatos szófaji egyértelműsítő módszerek kutatása több mint tízéves múltra néz vissza: szabályok alapján dolgozott Megyesi (1999) Svédországban, különféle valószínűségszámítási–statisztikai módszereket alkalmazott Oravecz és Dienes (2002) a Nyelvtudományi Intézetben, Kuba et al. (2004) Szegeden, Halácsy et al. (2006) a BME-n, valamint legutóbb Orosz (2011) a PPKE-n.

3. A magyar számítógépes szintaxis eredményei

A morfológiai rendszerek világához képest kisszámú és meglehetősen szerény képességű szintaktikai elemző és generáló modell készült a magyar nyelvre. Mivel kezdetben az orosz–magyar gépi fordítás megvalósítása volt a cél, a magyar szövegek szintetizálása állt a kutatások előterében, s ezt mindössze néhány kísérleti jellegű próbálkozás követte. A magyar szövegek szintaktikai szintézise a hatvanas évek elejének gépi fordítási lázában fontos kutatási területnek számított, ám egy-egy részterület tanulmányozásán túl az időszak jelentős eredmények nélkül zárult. Szintén a hatvanas években, Vargha Dénes elképzelései alapján az MTA Számítástechnikai Központjában indult meg az első automatikus mondattani elemzést végző program kidolgozása. Az eljárás alapjául Dömölki (1964) tetszőleges jelsorozatok felismerésére kidolgozott algoritmusa szolgált. A Varghaféle felfogásban a „nyelvtan" nem a hagyományos értelemben vett nyelvtant jelenti, mivel nem célja a mondatok és a nem mondatok megkülönböztetése képességének leírása. Annak, hogy a mondat a nyelvhez tartozik-e vagy sem, annyi köze van a nyelvtanhoz, mint egy tény igaz vagy hamis voltának egy róla szóló logikai állításhoz. Vargha megállapítja, hogy a szabad szórendű nyelvek elemzője nem használhat transzformációkat, mert azok vagy nem állíthatók elő ismert transzformációk (pl. a törlés) inverzeként, vagy a szerkezet ismerete nélkül nem alkalmazhatók. Maga az eljárás morfológiai elemzéssel kezdődik, és a szöveg morfémái helyett a szintaktikai elemző már csak kategóriakódjaikkal találkozik.

Hell György a hetvenes évek elején a BME Idegennyelvi Intézetében foglalkozott magyar mondatok szintaktikai elemzésével is: elképzelése a függőségi le-

íráson alapult. Kísérleti elemző algoritmusa csak egyszerű, ellipszismentes mondatokat kezelt. Gyakorlati megfigyelésekre épülő elemzőprogramját az Egyetemi Számítóközpont RAZDAN-1 gépén implementálták, gépi kódban (Hell 1975). Prószéky és Tóth (1979) szintaktikai elemzője szintén csak a kísérleti stádiumig jutott: az ELTE-1304 gépén futó FORTRAN nyelvű program egyszerű bővített (vesszőt nem tartalmazó) magyar mondatok nyelvtani elemzését végezte.

A MorphoLogic első kísérleti mondattani elemzőjének, a **HumorESK** rendszernek (Prószéky 1996) a segítségével valósult meg egy rövidhírek elemzésére készített rendszer, a **NewsPro** (Prószéky 2003). A kutatás a MorphoLogic, az MTA Nyelvtudományi Intézete és a SZTE Informatikai Tanszékcsoportjának közös projektjében zajlott, és eredménye egy olyan kísérleti elemző volt, amely egymondatos hírekből volt képes információt kivonatolni. A program gazdasági híreket kategorizált: mintegy 360 ún. „hírkeretet" különböztetett meg. Az említett szintaktikai modul olyan mondatszintű elemzést igénylő kutatás alapjául is szolgált, mint a pszichológiai szövegek elemzésére irányuló projekt a Pécsi Tudományegyetem Pszichológiai Tanszéke és a MorphoLogic együttműködésében. Az elkészült **LinTag** rendszer magyar nyelvű pszichológiai narratívumok nyelvi előelemzését végzi (László–Ehmann 2004). A program részleges, felszíni mondatelemzés útján kísérli meg a pszichológiai kutatás szempontjából releváns nyelvi markerek felismerését. A későbbiekben a László János vezette pszichológiai kutatócsoport és az MTA Nyelvtudományi Intézet kutatói az alább részletesebben említett **NooJ** rendszert és a **MetaMorpho** (szintén lásd alább) nyelvi elemzéseit is összekapcsolva megindították a narratív pszichológiai elemzésben a szemantikus szerepek vizsgálatát (Ehmann et al. 2011).

Az ezredforduló első éveiben elkészült egy másik, és a jelenleg is legátfogóbbnak tekinthető, a gyakorlatban is működő mondatelemző rendszer is: a **Moose** (Prószéky et al. 2004). Ennek segítségével további olyan új alkalmazási területeken sikerült mondattani megoldásokat ajánlani, ahol nemcsak az elemzés, hanem az azonnali eredménygenerálás is fontos. Ilyen volt például maga a **MetaMorpho** gépi fordító rendszer is. Az elemző érdekessége a szabad frázisrendű magyar nyelv különféle szintaktikai funkciójú nominális szerkezeteinek „begyűjtését" végző algoritmus (Merényi 2005). Ennek segítségével a magyar és az angol nyelv jelentős felszíni különbségei ellenére egyazon működtető formalizmus segítségével sikerült a szintaktikai elemzést megoldani.

A BME MOKK által készített **hunpars** szintaktikai elemző (Babarczy et al. 2005) bemenetként egy szövegfájlt kap mondatokkal, kimenetként pedig megadja a mondatok szintaktikai fáját egyszerű zárójelezéses jelölésben (illetve egy közvetlen szerkezetmegjelenítésre szolgáló grafikus formában).

A Pécsi Tudományegyetem számítógépes nyelvészeti kutatócsoportjának (Alberti 2011) kutatási célja kettős: egyrészt elméleti, egy saját kidolgozású totálisan lexikalista grammatika létjogosultságának és egzaktságának bizonyítása volt; másrészt gyakorlati, azaz egy komoly szemantikai komponenssel rendelkező elemzőprogram megalkotása is ott lebegett a célok között. Kutatásaik során elkészítettek egy Prolog programnyelvű elemzőt is, amely az elméletet volt hivatott demonstrálni, ám amely csak igen kis számú adattal működött. A program a jól formált (angol vagy magyar nyelvű) mondatokhoz morfofonológiai, szintaktikai és szemantikai reprezentációt társít, és a két nyelv egyszerű szerkezetei között egyfajta gépi fordítást is megvalósít. Napjainkban a munkálatok a **ReALIS** projekt keretében folynak (Alberti 2011), amely már egy nagy mennyiségű adatokkal is működni képes adatbázis-szerkezetet ígér a szintaktikai és szemantikai elemzés megvalósítására.

A **NooJ** valójában olyan integrált nyelvelemző környezet, amely egyaránt használható korpuszlekérdező eszköznek, komplex grammatikaépítő eszköznek, sőt nyelvészetet oktató eszköznek is. A szoftvert Max Silberztein fejlesztette ki a francia nyelv feldolgozásához (Silberztein 1993), de azóta már sok más nyelvre is átdolgozták. A magyarra 2003 óta folynak ezzel kapcsolatos fejlesztések az MTA Nyelvtudományi Intézetében. A rendszer meghonosítását nemcsak a robusztus és gyors véges állapotú technológia indokolja, hanem a fejlesztőknek az a szándéka is, hogy viszonylag könnyen használható oktatási eszközt is adjanak a nem informatikus nyelvészek számára. Első megközelítésben a NooJ egy gyors korpuszkezelő eszköznek tűnik, amely amint betöltöttünk egy sima formázatlan szöveget, máris készen áll arra, hogy lekérdezhessük reguláris kifejezések segítségével. A reguláris kifejezések azonban nemcsak a szavak alakjára, hanem nyelvi (morfoszintaktikai vagy akár szemantikai) jegyeikre is utalhatnak. Ezek az információk a szótári komponensből származnak, amely a rendszer központi részét alkotja. A szótár egy-, illetve többtagú kifejezések tára, amelyekben szóalakok találhatók, a lemmával és tetszőleges társított nyelvi információval, mindez igen hatékony véges állapotú belső reprezentációban. A rendszer egyedi sajátossága, hogy a szótár, a szöveg, valamint a szövegre alkalmazott grammatika egyaránt véges állapotú technológiával van megvalósítva. Ami a rendszert széles körben is különösen használhatóvá teszi, az a grafikai felület, amelyen viszonylag egyszerűen szerkeszthetjük és kezelhetjük a lexikai elemek vagy szintaktikai szerkezetek leírására szolgáló véges állapotú grammatikákat. A NooJ rendszer szótári modulja azonnal előállítja a szöveg morfológiai elemzését is. Az egyszerűbb szóalaktanú nyelvek esetében ezt úgy oldották meg, hogy az egy-egy szótőhöz tartozó összes képzett és ragozott alakot tételesen felsorolták egy szótárban, ami a magyar morfológia gazdagsága és produktivitása miatt nem járható út. A magyar változat

előállításához tehát meg kellett oldani a NooJ-on belüli morfológiai elemzés kérdését is (Vajda et al. 2004).

4. A magyar korpusznyelvészet eredményei

Az egyik legjelentősebb új nyelvtechnológiai fejlemény, amelyet a számítógépek kapacitásának növekedése okozott, a korpusznyelvészet megszületése. Ennek a kutatási területnek a segítségével a nyelvhasználat rejtett dimenziói kerülnek felszínre, méghozzá pontosan adatolt formában. Mint említettük, a korpusznyelvészet magyarországi története az 1980-as évek elejére, az Akadémiai Nagyszótár munkálatainak újraindításához vezethető vissza. Eredetileg 10 millió szövegszó összeállítása szerepelt a tervekben, amelyeket századonként egy-egy, főleg filológusokból álló szakértőbizottság állított össze 16–20. századbeli szövegekből. Az úgynevezett **Történeti Korpusz** mintegy 23 millió szövegszót tartalmaz, és 1772 és 2000 között keletkezett szépirodalmi, tudományos ismeretterjesztő és publicisztikai művekből, műrészletekből áll (Pajzs et al. 2004). A korpusz minden szava morfológiailag elemzett alakban szerepel. A szövegek feldolgozásánál különleges problémát jelentett a régies helyesírás, illetve az archaikus alakváltozatok kezelése. A ma már nem élő helyesírási alakok kódolására egy – már az indulásnál erre a célra bevezetett – speciális kódkészlet (Prószéky 1985) kiterjesztett változatát használják, azaz az alapbetű mellé tett szám segítségével kódolják az illető alapkarakter diakritikus jelekkel ellátott változatait. A kihalt alakok kezelésére külön heurisztikus eljárást kellett kidolgozni (Kiss et al. 2001).

A Történeti Korpusz munkálatainak befejeztével felmerült az igény, hogy a diakrón korpusz mellett szükség lenne egy nagyméretű, az aktuális nyelvhasználatot tükröző szinkrón korpusz összeállítására is. E nagyméretű vállalkozás hívta létre 1997-ben az MTA Nyelvtudományi Intézetén belül a Korpusznyelvészeti osztályt, melynek központi feladata a ma már **Magyar Nemzeti Szövegtár** (MNSZ) néven ismert korpusz megalkotása lett (Váradi 1999). Az eredeti cél egy 100 millió szónyi korpusz összeállítását irányozta elő, amely a legújabb írásos nyelvhasználatot volt hivatott tükrözni, méghozzá öt markánsan elkülönülő nyelvhasználati terület – a sajtó, a szépirodalom, a tudományos nyelv, a hivatali nyelvhasználat és a személyes közlések – külön-külön is lekérdezhető részkorpuszainak segítségével. A szépirodalmi alkorpusz teljes egészében tartalmazza a Digitális Irodalmi Akadémia anyagát, azaz az élő magyar irodalom anyagai is vizsgálhatók a korpusznyelvészet módszereivel. Az MNSZ szövegei bibliográfiai adatokkal jelzik az eredeti forrásokat, valamint az átvett szöveganyag fő szerkezeti és tartalmi egységei is jelölve vannak benne. Ezen felül minden egyes szöveg-

szó morfológiailag elemzett és egyértelműsített alakban szerepel. A morfológiai elemzés a MorphoLogic Humor morfológiai elemzőjével készült, az egyértelműsítés pedig egy erre a célra kidolgozott statisztikai alapú eljárással (Oravecz–Dienes 2002). 2003-ban megkezdődött az anyag kiegészítése a határon túli nyelvváltozatok szövegeivel. A **Kárpát-medencei Magyar Korpusz** megalkotásában az MTA Kisebbségkutató Intézete, illetve az MTA Nyelvtudományi Intézete koordinálásával négy határon túli kutatóállomás vett részt: a dunaszerdahelyi Gramma Nyelvi Iroda, a szabadkai Magyarságkutató Társaság, a Kárpátaljai Nyelvi Iroda és a kolozsvári Szabó T. Attila Nyelvi Intézet.

A **Webkorpusz** 2003 telén született a **Szószablya** projekt keretében a BME MOKK-ban: több mint 1,48 milliárd szavával (szűretlenül, illetve 589 millió megszűrt szóval) ez jelenleg a legnagyobb magyar nyelvű korpusz. A gyűjtemény 18 millió magyar weboldalból áll. A többszörösen előforduló szövegállományokat, illetőleg a használható szöveget nem tartalmazó állományokat kiszűrték belőle. A szövegek teljes állományát alapul véve előállt egy gyakorisági szótár is, amely a különböző szűrési szintek mellett tartalmazza az egyes szóalakok gyakoriságát. A Webkorpusz kétféle formátumban tölthető le: a szövegeken alapuló gyakorisági szótárként és az eredeti szövegek összességeként.

Mivel a szóalaktani szint magában hordozza a többértelműséget, a szófaj egyértelmű megállapításához a szó környezetének tanulmányozására, illetve az ezt lehetővé tevő szövegkorpuszokra van szükség. Ezek megvalósításához a SZTE Informatikai Tanszékcsoport és a MorphoLogic együttműködésével alakult konzorcium 2000 és 2002 között **Szeged Korpusz** néven elkészített egy magyar természetes nyelvi szövegadatbázist, valamint egy, a szófaji egyértelműsítést támogató szoftverrendszert (Csendes et al. 2005). A korpusz a szövegeket strukturáltan tárolja (cikk, bekezdések, mondatok). A szöveg minden egyes szava mellett szerepel a Humor morfológiai elemző kimenete, amely a lehetséges szófaji kódokat és szótöveket tartalmazza, valamint a kézi egyértelműsítéssel kiválasztott, az adott szövegkörnyezetnek megfelelő helyes kódolás és szótő. A szavak szófaji kódolása az európai nyelvekre azzal az MSD-kódrendszerrel történt, amelyet az MTA Nyelvtudományi Intézete és a MorphoLogic alakított ki – egy akkor már létező nemzetközi sztenderd, a MULTEXT alapján – a **MULTEXT-EAST** nevű Copernicus-pályázatban. Az öt kisebb témakörből származó – szépirodalmi, publicisztikai, számítástechnikai, jogi szövegekből, valamint tizenévesek rövid írásaiból álló –, összességében egymillió szövegszót tartalmazó magyar korpusz a TEI nemzetközi szövegkódolási ajánlásnak megfelelő XML-formátumban készült. A korpusz 1.0 változatát egy 200 ezer szóból álló, üzleti szövegeket tartalmazó részkorpusszal egészítette ki a gazdasági szövegek elemzését végző **NewsPro** rendszert (Prószéky 2003) megvalósító – és az MTA Nyelvtudományi Inté-

zetéből, a Szegedi Tudományegyetemből és a MorphoLogicból álló – konzorcium. Ezzel létrejött a korpusz 1,2 millió szövegszavas és 225 ezer írásjel méretű 2.0 verziója. Az annotálást követően a konzorcium kutatói megvizsgálták a gépi tanulási algoritmusok alkalmazhatóságát a lapos szintaktikai elemzés problémájára. Az algoritmusok hatékony működtetéséhez főnévicsoport-felismerő szabályokat vontak ki a korpuszból, majd ezeket szakértők által definiált szabályokkal kombinálták. A **Szeged Treebank** a Szeged Korpusz mondatszerkezeti egységeinek bejelölését is tartalmazó változata (Csendes et al. 2005), a **Szeged Dependencia Treebank** pedig a Szeged Treebank függőségi mondatszerkezetekkel való reprezentációja (Vincze et al. 2009).

Az egynyelvű korpuszok mellett az utóbbi időben egyre több figyelem irányul az ún. párhuzamos korpuszok kutatására. Párhuzamos korpusznak olyan két-, esetleg többnyelvű korpuszt nevezünk, ahol az egyik nyelvű korpusz szövegei a másik szövegeinek fordításai. Az ilyen korpuszok kutatásának célja az, hogy kiaknázza és újrafelhasználja a fordításokban megtestesülő emberi tudást. Ez különféle számítógépes alkalmazások, jelesül a gépi fordítás vagy a számítógéppel támogatott fordítás számára rendkívül értékes, de a fordítástudomány is egyre inkább támaszkodik az ilyen korpuszokra. A már említett MULTEXT-EAST projektum keretében elkészült egy párhuzamos korpusz, amely George Orwell *1984* című regényének angol eredetijét és annak számos nyelvre, közöttük a magyarra való fordítását is tartalmazza (Dimitrova et al. 1998). A korpusz értékét növeli, hogy akárcsak az MNSZ és a Szeged Korpusz, ez is gondosan van nyelvileg annotálva: minden szövegszó morfológiailag elemezve és egyértelműsítve van.

A **Hunglish Korpusz** egy angol–magyar kétnyelvű mondatgyűjtemény, amely az MTA Nyelvtudományi Intézete és a BME Médiaoktatási és Kutató Központja közreműködésében született (Halácsy et al. 2005). A **huntoken** program magyar nyelvű szövegeket mondatokra, azon belül pedig ún. tokenekre (szavakra és központozási jelekre) bont. Lexikonépítéshez, információ-visszakereséshez, szövegbányászathoz és sok egyéb természetesnyelv-feldolgozó alkalmazáshoz is használható. A **hunalign** egy szabadon felhasználható automatikus mondatszinkronizáló program párhuzamos korpuszok építésére.

Az utóbbi években újabb párhuzamos korpuszok is jelentkeztek: ilyen a **SzegedParalell** kézzel párhuzamosított angol–magyar korpusz (Tóth et al. 2008), illetőleg a **HunOr** magyar–orosz párhuzamos korpusz (Szabó et al. 2011).

5. A magyar számítógépes lexikográfia eredményei

A magyar lexikográfia számítógépes munkálatai elsősorban az irodalmi nyelv vizsgálatára irányulnak, legtöbbször szerzők szerinti bontásban. Az alábbiakban a magyar számítógépes lexikológiai kutatások közül azokkal foglalkozunk, amelyek a magyar nyelv számítógépes rendszerekben való alkalmazásához készültek, vagy megfelelő átalakítással ahhoz felhasználhatóak.

A magyar nyelv értelmező szótára 58 323 címszóból álló anyagát 1963-tól a debreceni KLTE oktatói és hallgatói vitték lyukkártyára Papp Ferenc vezetésével. Magát a kódokkal kiegészített és lyukkártyán tárolt anyagot származási helye után Debreceni Thészaurusznak is nevezik (Papp 2000). Az Értelmező szótár anyagának elkészítették a szóalakok vége szerinti rendezését, az anyag nyelvtani (elsősorban morfológiai) szempontok szerinti kódolását, statisztikákat a nyelvtani kódok alapján, valamint a bent levő információkhoz hozzávettek további, új szempontok szerinti kódokat. Az anyag 1969-ben könyv alakban is napvilágot látott **A magyar nyelv szóvégmutató szótára** címmel (Papp 1969). Ezzel gyakorlatilag egyidejűleg, Wolfgang Veenker német nyelvész könyv alakban megjelentette a magyar toldalékok és toldalékkombinációk *a tergo* jegyzékét (Veenker 1968). Ebben nyelvünk ragjai, jelei, sőt képzői is megtalálhatók, méghozzá minden lehetséges, illetve a szerző által lehetségesnek tartott kombinációban. A nyolcvanas években mind a Szóvégmutató Szótár, mind a Veenker-féle toldalék-adatbázis eredeti lyukkártyás formájában megtalált és újrafelhasználhatóvá tett anyagából az MTA SZTAKI akkori igen korszerű IBM 3031 számítógépén lekérdezhető adatbázis készült (Kornai 1986), amely innen jutott el az akkor már éledező személyi számítógépek világába.

Az **Értelmező kéziszótár** új változatának kidolgozásakor az MTA Nyelvtudományi Intézetében felmerült az igény, hogy ez a szótár már korszerű, formanyelven kódolt elektronikus változatban szülessen meg. Az ehhez szükséges kutatás a teljes szótár lexikai adatbázissá alakítását tűzte ki célul. A feldolgozás során számos igen munkaigényes feladatot kellett elvégezni annak érdekében, hogy az emberi olvasásra és megértésre készült szócikkekből nyelvtechnológiai felhasználásra alkalmas lexikai adatbázis alakuljon ki.

A nagyközönség számára csak könyv formában volt elérhető **A magyar nyelv gyakorisági szótára**, amely egy 500 ezer, 20. századi szépirodalmi szövegekből való szövegszót tartalmazó anyagon nyugszik. Ebből az anyagból az idők folyamán többféle, gépi statisztikai módszerekkel kialakított gyakorisági lista készült, amely bekerült az MTA SZTAKI már említett adatbázisába is (Kornai 1986).

Az MTA 1984-ben határozatban döntött arról, hogy létre kell hozni **A magyar nyelv nagyszótárát**, amely eredetileg a legutóbbi öt évszázad, jelenleg azon-

ban az elmúlt 230 év magyar nyelvének szóanyagát tartalmazó nyelvtörténeti szótár. Mintegy 110 ezres címszóállományát egy összesen 13 millió szövegszót tartalmazó szövegkorpuszból számítógépes segédlettel állítják elő. Az anyaggyűjtés, azaz a szótári cédulák kézírásos készítése korábban mintegy hetven éven át folyt, aminek eredményeképpen az 1970-es évekre 4,5–5 milliósra becsült szótári cédulatömeg gyűlt össze. A Magyar Tudományos Akadémia 1984-ben határozatban döntött a nagyszótári munkálatok folytatásáról, és egyben azt is kimondta, hogy a szótár munkálatait számítógép segítségével, az írásbeliség kezdetétől napjainkig ívelő számítógépes szöveges adatbázis, azaz számítógépes korpusz alapján kell végezni (Pajzs 1990). Ez a döntés nemcsak a számítógépes lexikográfia intézményes megerősödéséhez vezetett, hanem egyben ezeken az alapokon indult el hazánkban a korpusznyelvészet is. Az Akadémiai Nagyszótár ma már korszerű XML-adatbázisként készülő anyagából könyv alakban eddig négy kötet jelent meg.

Az MTA Nyelvtudományi Intézetében időközben megvalósult a könyv alakban korábban megjelent **Magyar ragozási szótár** (Elekfi 1994) adatbázissá való átalakítása is, mely eredetileg az Értelmező kéziszótár számára készült ragozási útmutatóból lett egy önálló, a szótár teljes szócikkállományát feldolgozó szótár. Ahhoz, hogy a Magyar ragozási szótár gazdag tartalmát számítógép számára kezelhető alakra hozzák, a szótárban rejlő implicit információt explicit alakra kellett alakítani. Ennek első lépéseként minden egyes paradigmatáblát elő kellett állítani, azaz az öröklött jegyeket az adott paradigma egyéni jegyeivel együtt le kellett generálni a tőalakváltozatok pontos feltüntetésével. További feladat volt a toldalékok lehetséges kombinációinak előállítása is, valamint a szótár eredeti céljain túl még a képzőket is bevonták az alakváltozatok leírásába.

Az MTA Nyelvtudományi Intézetének Korpusznyelvészeti Osztályán a 2000-es években létrejött a **Vonzatszótár-adatbázis**. Ez minden olyan vonzat jellegű információt tartalmaz, amely a magyar nyelv számítógépes szintaktikai elemzéséhez szükséges lehet. Szóanyagát a Magyar Nemzeti Szövegtár leggyakoribb 20 ezer szava, központi részét pedig egy több mint háromezer elemű igei adatbázis alkotja. A vonzatokat felszíni esetvégződésük szerint (pl. nominatívusz, akkuzatívusz és még legfeljebb két vonzat), a tematikus szerep megjelölése nélkül tartják számon. Emellett megszorító szabályok is vannak, amelyek a mondat főbb összetevőinek (alany, tárgy) jegyeire hivatkoznak (pl. élő alany, absztrakt tárgy stb.). A vonzatkeret mellett feltüntették a főmondat és az ige komplemenseként szereplő tagmondat közötti koreferenciális viszonyokat is.

Napjaink egyik legfontosabb nyelvtechnológiai célja, hogy a szavakat és jelentésüket egy egységes, nyelvi és világismeretet tartalmazó fogalmi rendszerben helyezzük el. Az egyik legszélesebb körben használt ilyen fogalmi rendszer a Princeton Egyetemen készített **WordNet** adatbázis (Miller et al. 1990), amely

több mint százezer nyelvi egység között definiál fogalmi viszonyokat. A vállalkozás annyira sikeresnek bizonyult, hogy több európai nyelvre is adaptálták az EuroWordNet projektum keretében. A magyar nyelv WordNethez kapcsolásával foglalkozó első kísérletek a 2000-es évek elején indultak el, amikor a MorphoLogic kutatói módszereket kezdtek keresni, illetve kidolgozni arra, hogy az angol nyelvű WordNet adatbázist – először csak a főnévi részét – minél automatikusabb módon lehessen átültetni magyarra (Prószéky–Miháltz 2002). Az eljárás mögött az a hipotézis áll, hogy a WordNet-rendszerben kódolt relációk többékevésbé nyelvfüggetlenek, ezért tehát, ha a rendszer csomópontjain álló lexikai elemekhez találunk magyar megfelelőt, a köztük lévő fogalmi kapcsolat az angol WordNetből egyszerűen átörökíthető. A kísérleteket egy már több intézmény által koordinált kutatás követte: a **Magyar WordNet** (sokszor: HuWN) teljes létrehozására irányuló munka 2005 és 2007 között folyt a MorphoLogic, az MTA Nyelvtudományi Intézete és a SZTE Informatikai Tanszékcsoportja közreműködésével (Prószéky–Miháltz 2008; Miháltz et al. 2008). Időközben több hazai intézmény kutatói úgy ítélték meg, hogy a szemantikai jegyek kódolását a jövőben szerencsés volna egységes formában végezni. Ezért 2004 és 2006 között folyt egy ezt megcélzó projekt, a **Magyar Egységes Ontológia**, az NKFP támogatásával.

6. A magyar számítógépes nyelvészet eredményei a gépi fordítás területén

A számítógépes nyelvészeti kutatás klasszikus problémája a gépi fordítás, amelynek természetesen csak tudományos, szakmai, esetleg köznapi szövegek (hírek, hirdetések stb.) lefordításában vagy megértésében van szerepe. A kutatások nem tudnak és nem is szándékoznak kiterjedni a szépirodalmi szövegek számítógépes vizsgálatára és a műfordításra. Ebben az irányban, tehát a hagyományos, teljesen gépi úton végzett fordítórendszer fejlesztése irányában is megindultak munkálatok. Az MTA Nyelvtudományi Intézetében az EU 5. keretprogramja által finanszírozott **MATCHPAD** projektum keretében folyt egy nagyszabású kísérlet egy angol–magyar fordítórendszer kifejlesztésére (Senellart et al. 2001). A szoftvertechnológiát a francia Systran cég nyújtotta, amely egyike az első generációs fordítórendszereknek, és jelenleg szinte az egyetlen olyan általános célú gépi fordítórendszer, amely bizonyítottan jól működik. A magyar nyelv ehhez szükséges leírása az MTA Nyelvtudományi Intézete és a MorphoLogic együttműködésével készült. A rendkívül gazdag morfológia, az indoeurópai nyelvekétől nagymértékben különböző elvű mondatszerkesztés igazi kihívást jelentett a francia

szoftvercég számára is, és bebizonyosodott, hogy a nemzeti nyelvek technológiai megoldásait nem lehet automatikusan importálni más nyelvek bevált rutinmegoldásaiból.

A MorphoLogic által 2000-től kezdődően fejlesztett **MetaMorpho** (Prószéky–Tihanyi 2002) gépi fordítórendszerben az igazi újdonságot egyrészt a szabályok és a példák egységes kezelése jelenti, másrészt a rendszer a hagyományos fordítóprogramoktól eltérő elvet használ: gyakorlatilag a forrásnyelvi elemzés „melléktermékeként" jön létre a célnyelvi szöveg. A fejlesztők nyelvi mintának neveznek minden olyan szimbolikus leírást, amelyet a szövegtest valamely részére helyezve a benne szereplő szimbólumok illeszkednek a szöveg megfelelő elemeire, legyen ez az illeszkedés betű szerinti, szófaji vagy jelentés alapú, vagy a nyelvész által definiált egyéb megfeleltetés. Ha a minták rövidek és specifikusak, akkor más elméletekben szótári elemeknek hívják őket; ha hosszabbak, akkor kollokációknak vagy idiómáknak. Ha viszont kevésbé specifikusak, akkor ezek a minták nem lexikális, hanem strukturális szegmensek, azaz nyelvi szerkezetek, címkézett zárójelezések. A több mint kétszázezer szabálysémát tartalmazó MetaMorpho mindezeket a mintákat egységesen kezeli, illeszthetőségük sikeressége esetén lehetővé teszi a hozzájuk tartozó célnyelvi minták megjelenését. A rendszer a célnyelvi oldalon a minták egymásba építését egyfajta függvényalkalmazásként oldja meg. A teljes MetaMorpho-formalizmus és a működtető rendszer, valamint az angol–magyar nyelvi adatbázis a MorphoLogic kutatóinak saját fejlesztése, a magyar–angol nyelvi adatbázis építéséhez az MTA Nyelvtudományi Intézete és a Szegedi Tudományegyetem kutatói csatlakoztak egy erre szolgáló pályázat keretében. A program ingyenesen használható 2005 óta a www.webforditas.hu weboldalon, valamint az ennek a programnak az alapötletére épülő és (épp ezért) magyar kutatók vezette nemzetközi konzorcium által 2012 elején publikussá tett www.itranslate4.hu weboldalakon is. Ezeken a magyar nyelv már nemcsak az angollal, hanem – az angolon keresztül más kutatópartnerek angol–X nyelvű moduljainak a kiegészítésével – sok világnyelvre, és gyakorlatilag az összes európai nyelvre, illetve ezekről a nyelvekről magyarra is képes fordítani.

7. A magyar nyelvtechnológiai kutatások gyakorlati eredményei

A számítógépes nyelvészeti alkalmazások gyakorlati jelentőségét az adja, hogy időközben a számítógép alapvetően és elsősorban a kinyomtatandó vagy fel-

olvasandó – és egyre inkább elektronikus formában felhasznált – dokumentumok előállításának eszközévé vált. A Humor morfológiai leíráson alapuló helyesírás-ellenőrzőként bevezetett **Helyes-e?**, valamint a szintén a kilencvenes évek elején kidolgozott **NyelvÉsz** – később **Lektor** (Seregy 1991) – valójában még csak szóellenőrzők voltak. A szószintű helyesírás-ellenőrzőnek „csőlátása" van, hiszen mindig csak azt az egy szót látja, amit odaadott neki a hívó program; fogalma sincs az előző és a következő szavakról. A fentiekkel szemben, ha valaki mondatszinten ellenőriz, akkor több mindent lát, kombinálni tudja a mondat szavainak nyelvi tulajdonságait, és ezáltal bonyolultabb jelenségeket, egybeírást–különírást, vesszőhibákat is képes kezelni. Ezt a fejlesztést végezte el a MorphoLogic a **Helyesebb** rendszer kidolgozásakor (Naszódi 1997). A kifejlesztett módszer az ún. részleges szintaktikus leírással adja meg az egyes hibajelenségek formális szabályait. A mondatszintű helyesírás-ellenőrző jelenleg körülbelül négyezer szabályt tartalmaz, de újabb jelenségek leírásával a korábbi szabályok módosítása nélkül is bővíthető. A magyar elválasztást nem lehet az elválasztási szabályok puszta gépi kezelésével megoldani. A **Helyesel** elválasztó rendszer (Prószéky–Kis 1999), amely a megjelenése után hamarosan összeépült a Helyes-e? helyesírás-ellenőrzővel, a szótagolás tökéletes megoldásához a Humor morfológiai elemző programot használja. Ennek a feladata ebben az esetben az egyes szóalakokat felépítő morfémák határainak megtalálása. A kérdéses szóalak morfológiai elemzése segítségével megállapítható, hogy az elválasztás szempontjából összetett szó-e, és ha igen, melyek azok a morfémahatárok, amelyek felülbírálják az egyszerű szótagolással kapott elválasztási pozíciókat. A választékos fogalmazás támogatására a MorphoLogic kidolgozott egy toldalékoló szinonimaszótárt, a **Helyette** rendszert (Prószéky–Tihanyi 1993). Ez három, nyelvi szempontból fontos funkciót valósít meg: felismeri a forrás-szóalak szótári tövét, megkeresi a forrásszó jelentésköreit, és az azokhoz tartozó szinonimákat; majd visszaírja a szövegbe a kiválasztott szinonima megfelelő alakját. A bemutatott szó- és mondatszintű helyesírás-ellenőrzőből, elválasztóból és szinonimaszótárból álló **Helyesek** magyar nyelvhelyesség-ellenőrző programcsomag 1993 óta beépült az összes magyarországi irodai rendszerbe (Prószéky–Kis 1999), sőt ugyanez a magyar technológia a román nyelv leírására alkalmazva 1996-tól elérhető az összes romániai irodai termékében is. Fontos nyelvpolitikai eredmény volt, hogy a MorphoLogic teljes magyar nyelvhelyességi csomagja 2000-ben bekerült a legelterjedtebb irodai programrendszer szlovák nyelvű változatába is. A szövegekben való keresés szerepe az utóbbi időben az internet előretörése miatt jelentősen megnőtt. A mai keresőprogramok egyszerűen egy rövidebb betűsorozatot próbálnak megkeresni egy nagyon hosszúban, méghozzá minden intelligencia nélkül, a keresett szövegnek csak azokat az előfordulásait jelezve, melyek pontosan, betűhíven meg-

egyeznek a keresendő betűsorozattal. Ennek a problémának a kiküszöbölésére fejlesztette ki a MorphoLogic a magyarra és más nyelvekre a **HelyesLem** lemmatizáló rendszert (Prószéky 1996), amelyet többek közt a Microsoft által több nyelv keresőmoduljába beépített **MorphoStem** kereséstámogató rendszer is használ (Prószéky 2001).

A nyelvtechnológiában sokszor van szükség egy szöveg nyelvének az azonosítására. Ha megvan a nyelv, meghívhatók az adott nyelvet kezelni képes nyelvtechnológiai eszközök. A nyelvazonosítást statisztikai módszerrel vagy szólista segítségével szokás végezni, de mindkettőhöz nagy mennyiségű, adott nyelvű szöveget kell feldolgozni. A statisztikai alapúnál különböző méretű szórészek előfordulási valószínűségéből hozzák meg a döntést, a szólista alapú megközelítés szógyakoriságok összehasonlításán alapul. Ez utóbbi módszeren alapuló rendszert fejlesztettek ki (Németh et al. 2000) a BME Távközlési és Telematikai Tanszékén (2003-tól Távközlési és Médiainformatikai Tanszék, röviden: TMIT). Ez azt határozza meg, hogy az adott levél szövege magyar, német, illetve angol nyelvű-e. A kialakított rendszer 96%-ban helyesen állapítja meg a dokumentum nyelvét, amennyiben az több mondatból áll. A MorphoLogic által működtetett ingyenes fordítóportálon, a www.webforditas.hu weboldalon egy szintén statisztikai alapú nyelvfelismerő, a **LangWitch** került beépítésre a fordítandó szöveg nyelvének azonosítására.

A 2000-es évek elején folyt még egy érdekes kutatás: a felismerőprogramok folytonos bemenetét szegmentálni képes eszköz kezeli az időben (akár beszédhanghossz, akár karakterszélesség alapján) és minőségben alulspecifikált információt és a nyelvi modulok párhuzamos kezeléséről is gondoskodik. A **Recognition Assistant** rendszer (Prószéky et al. 2002) először egy kézírás-felismerő rendszer prototípusának kialakításakor került beépítésre (Karacs et al. 2009).

A számítógépes, illetve mobiltelefonos gyakorlatban – különböző okok miatt – gyakoriak az olyan magyar szövegek, amelyekben az egyébként ékezetes betűket az ékezet nélküli legközelebbi megfelelőjükkel írják (e-levelek, SMS-szövegek). Amennyiben ilyen „csonka" szövegeket kell felolvastatni egy beszédszintetizátorral, a felolvasás előtt helyre kell állítani az ékezeteket. Ezt nevezik automatikus ékezetesítésnek. A magyarban öt olyan ékezet nélküli betű van, melynek legalább egy ékezetes párja is létezik. Vannak viszont olyan szavaink is, amelyeknek mind az ékezetes, mind az ékezet nélküli formája értelmes, ezért nehéz eldönteni, hogy a szöveg adott pontján melyik a helyes (pl. *meg, még*). Minél hosszabb egyébként egy szó, annál többféle ékezetesített változatot lehet vonatkoztatni rá (természetesen ezekből csak néhányra lehet azt mondani, hogy nyelvileg helyes). A nyelvi szabályokon alapuló ékezetesítő megoldás csak a magyar köznyelvi szóállományra végez sikeres ékezetesítést, a személy-, illetve cégnevek-

re például nem használható eredményesen, könnyen téveszthet. Ilyen feladatnál külön kivételszótárakat kell a nevek értelmezésére készíteni. Magyar nyelvre 1999-ben készült egy automatikus, statisztikai alapú ékezetesítő algoritmus a BME Távközlési és Telematikai Tanszékén az első magyar elektronikus levélfelolvasóhoz. A statisztikai elemzések egy 25 millió szavas szövegállományon alapulnak, és a segítségével készült ékezetesítő 95%-os pontossággal működött (Németh et al. 2000). Egy elsősorban morfológiai meggondolásokon alapuló ékezetesítő algoritmus működött a MorphoLogicnak a (ma már írásban nem használt, de sok magyar nyelvjárásban meglevő) zárt *ë* hangok szövegbeli bejelölését végző programjában is (Novák–Endrédy 2005), amelyet később általános ékezetesítési problémák megoldására is használhatóvá tettek.

A **MoBiMouse** szótárrendszer (Clark 2000; Prószéky–Kis 2002) egy szövegfelismerő modul, egy nyelvi elemző és számítógépes szótárak kombinációja. A felhasználó az egérmutatóval rámutat a szöveg valamely részére, a program az egérmutató alatti szót és környezetét „elolvassa", és a szó tövét – adott esetben a környezetben szereplő szavakkal együtt – úgy továbbítja az éppen aktív szótáraknak, hogy azok a lehetőségekhez mérten, környezetfüggő módon a szó aktuális környezetének megfelelő jelentéseit adják csak vissza, egyfajta dinamikus szócikk-előállító modul működésének következtében. A MoBiMouse rendszer felületének, valamint a MetaMorpho fordítóprogramnak a kombinációja az internetes szolgáltatásként működő **MoBiCAT** megértéstámogató–fordító, amely egy a mondat fölött megjelenő buborékban az aktuálisan kijelölt szót tartalmazó teljes mondat azonnali fordítását nyújtja (Tihanyi 2003).

8. A magyar nyelvtechnológia eredményei a beszéd kezelésében

Az írott nyelvvel kapcsolatos nyelvtechnológiai eredmények azért olyan fontosak, mert – az emberrel szemben – a számítógépnek az írott és nem a beszélt nyelv az „elsődleges nyelve". Ugyanakkor az egyre emberközelibb, továbbá az egyre táguló információtechnológiai alkalmazások igénylik azt is, hogy bizonyos információkat a gép szóban mondjon el (beszédszintézis), illetve, hogy a számítógép megértse az emberi beszédet (beszédfelismerés). Ez a terület – a nyelvtechnológiai meghatározást követve – a beszédtechnológia. Itt is kiváló eredményeket mutathat fel a magyar kutatás-fejlesztés. A beszéd mesterséges előállításának kiinduló alapja a szöveg, amit a gép felolvas. A beszéd megértésekor az elhangzó akusztikai jelből kell a gépnek eljutni a nyelvi formához. A beszédtechnológia

alapjainak elsajátításához ajánljuk az érdeklődőknek a Németh Géza és Olaszy Gábor által szerkesztett könyvet (Németh–Olaszy 2010). Az akusztikai, fizikai, jelfeldolgozási folyamatok professzionális kezelésén túl is azonban az a folyamat, amelyben a szövegtől a gépi beszédig vagy a gépi beszédtől az írott szövegig eljutunk, számos olyan nyelvtechnológiai megoldást tartalmaz, amelyben a szűkebb értelemben vett nyelvészet is érintett. Az automatikus beszéd-előállítás egyik legnehezebb problémaköre a név- és címfelolvasás jó minőségű megoldása (például cégbírósági adatok lekérése telefonon, tőzsdei információk beszéddel való megadása, automatikus telefonos tudakozó a szám alapján stb.), ugyanis meg kell határozni a név (cég-, illetve személynév) hangzó, kiejtési formáját (ami sok esetben nem egyszerű), majd a kiejtés prozódiai paramétereit (hol legyen hangsúly, szünet, milyen dallamformával kell „elmondani" a kért adatot), végül ki kell alakítani az esetleges szótagolási, betűzési formákhoz a szabályokat. A megoldásra nagy mennyiségű valós név- és címadatot kell feldolgozni, statisztikailag osztályozni, csoportokba sorolni, elemezni és kialakítani a megfelelő kiejtési szabályokat, prozódiai formákat. Magyarországon az első komplex név- és címfelolvasó 2003-ban készült el a BME TMIT fejlesztésében (Németh et al. 2003) egy automatikus számszerinti tudakozó alkalmazáshoz (mintegy négymillió telefon-előfizető adatainak felolvasására). A fejlesztés során végzett tesztek azt mutatták, hogy a nevek, cégnevek gépi felolvasásánál még fokozottabban érvényes a jó érthetőség biztosítása (esetleges túlbiztosítása), mint a normál szöveges felolvasásnál, hiszen ennek hiánya hibás információadást eredményez. Erre fejlesztették ki az úgynevezett „részletező" felolvasási formát (Fék et al. 2004), amely az első magyar beszélő szótagoló automatának is tekinthető. A részletező felolvasást kérő felhasználó szótagolva hallja az adott nevet, továbbá kiegészítő, pontosító információkat is tud kérni a név írásával kapcsolatban, pl. családnevek esetében.

A gépi beszédkeltés egyik kulcskérdése a beszéd dallam-, hangsúlyozási, ritmikai és intenzitásszerkezetének (a prozódiának) a helyes megvalósítása. A prozódia legfontosabb elemei a szöveg alapján előre jósolhatók. Ilyenek a mondatdallam, a hangsúlyos/hangsúlytalan szavak, a gondolati egységet alkotó szövegrészek (szintagmák) határai, a beszédsebesség lassulása/gyorsulása, a szünetek helye és hossza, valamint az átlagintenzitás változása. Ez az egyik legbonyolultabb nyelvi technológiai témakör, amelyre hazánkban inkább statisztikai alapú megoldásokat használnak, amelyek a gyakorlatban különféle közlekedési tájékoztató rendszerekben, ügyfélszolgálatoknál ugyanúgy megtalálhatók, mint az interneten: a **Profivox** rendszer a weben időjárás-jelentéseket, vagy látássérültek számára akár teljes szépirodalmi műveket képes jó minőségben felolvasni (Olaszy et al. 2000).

A magyar esetében több száz szabály biztosítja a korrekt szöveg–hang konverziót (az angolra például több ezer ilyen szabályt kell meghatározni). Az átalakítási folyamat eredménye, hogy a szövegből kialakul a kiejtendő hangsor hangjainak sorozata. Ebből már összeállítható a ténylegesen megszólaltatható nyers beszédhangsor. A hangsor fizikai megvalósítása általában előre eltárolt (emberi beszédből kivágott) hullámforma-részletek összekapcsolásával történik. Ebben a fázisban is lényeges szerepe van a nyelvtechnológiának annak kiválasztásában, hogy mik legyenek a beszédhangsort felépítő optimális elemek: hangok, hangkapcsolatok, szótagok, szavak vagy esetleg más egységek (Olaszy 1999). A beszéd időszerkezetével kapcsolatos modellkutatások eredményeként nagyméretű magyar szóadatbázis (1,5 millió szó) készült (Olaszy 2002), amely az összeállított hangfolyam időszerkezetének a meghatározásához alapvetően szükséges volt. Természetesen a prozódia megvalósításához is megfelelő modellt kellett készíteni. A modellből kapott adatok fizikai megvalósításához fejlett jelfeldolgozási algoritmusok álltak rendelkezésre (Gordos–Takács 1983), amelyekkel például ráültethető a hangsorra a kívánt dallammenet.

A gépi beszédfelismerés még a beszédkeltésnél is nehezebb feladat, és célja általában az elhangzott hangsor gépi átírása a helyesírásnak megfelelő írott alakba, illetve egy előre meghatározott elemhalmazból történő kiválasztás az elhangzott hangsor alapján, ami parancsszavas vezérlés, vagy kulcsszó-felismerés esetén szükséges. A BME TMIT-n végzett kutatások kimutatták, hogy a (fonetikai értelemben) környezetfüggő beszédhangmodellek alkalmazásával a felismerési hiba a harmadára csökkenthető (Fegyó et al. 2003). Ez a kutatási eredmény tette a gyakorlatban is használhatóvá a személyfüggetlen nagyszótáras beszédfelismerést. A magyar nyelvre is készült már a kiejtési szabályok alapján működő automatikus fonetikus átíró program (Mihajlik et al. 2002). Magyarországon az első ilyen általánosan használható, beszélőfüggetlen, ezres nagyságrendű szótárra épülő rendszert a BME TMIT-n dolgozták ki az AITIA Zrt.-vel közös kutatásban hanggal vezérelhető telefonközpontok kialakítására (Fegyó et al. 2003). Szegeden a kétezres években szintén megindult egy folytonos, magyar nyelvű beszédfelismerő rendszer kialakíthatóságának kutatása is. A rejtett Markov-technológián alapuló (orvosi diktálás célját szolgáló) prototípusrendszer akusztikai része a beszélő hangjához hozzáigazodó, ezáltal a pontosságot nagymértékben növelni képes modult is tartalmaz. A folyamatos diktálás nyelvi szintű algoritmikus támogatása szó-*n*-gramokat, különböző simítási módszereket és környezetfüggetlen nyelvtani modellezést is magába foglal.

Magyarországon az utóbbi 15 évben komoly kutatási eredmények születtek a speciális beszédadatbázisok tervezése, fejlesztése és használata területén. Ilyen volt például a **Babel** nevű, olvasott szövegű beszédadatbázis, amelyben a magyar

hangkapcsolatok 97%-ára van minta (Vicsi–Vig 1998). A **Speechdat** vezetékes telefonbeszéd-adatbázis magyar változata (Vicsi 2001), valamint annak mobiltelefonos változata (Vicsi et al. 2002) kifejezetten izolált szavakat és szókapcsolatokat, valamint dialógusszövegek leglényegesebb elemeit tartalmazza. A BME és a Szegedi Tudományegyetem kutatói egy diktálórendszer készítéséhez fejlesztettek irodai környezetben rögzített beszédadatbázist (Vicsi et al. 2004), amelynek szövegkészlete a magyar nyelv hangzókapcsolatainak statisztikai feldolgozásán alapszik.

Az ígéretes kutatási irányok között feltétlen meg kell említeni, hogy a PPKE ITK-n egy kutatási program keretében a hangzó beszédet valós időben egy ezt a hangsort produkáló ideális száj mozgásává konvertálták (Takács et al. 2006), lehetővé téve ezzel a siketek mobiltelefon-használatának alapjait.

9. Összegzés

Tanulmányunkban igyekeztünk összefoglalni a hazai nyelv- és beszédtechnológia legfontosabb eredményeit. A különféle nyelvi szinteknek megfelelő gépi kutatások természetes következményei az általános nyelvészet által kijelölt nyelvi szinteken történő kutatásoknak. Van azonban a gépi módszereknek olyan ága is, amely a hagyományos nyelvészeti irodalomban nem létezik. Ilyen például a szófaji egyértelműsítés, illetve ilyen maga az egész gépi fordítás is. Nagyon nehéz a szóba jöhető jövőbeli kutatási irányokról bármit is mondani, hiszen a hazai kutatások javarészt követik a világ nagy nyelvtechnológiai kutatási trendjeit, ám sokszor az agglutináló, szabadabb szórendű nyelvekre jellemző nyelvi jelenségek gépi kezelésének megvalósításával kiegészítik, pontosítják is őket.

Mivel „a nyelvi technológiák kifejlesztése a magyar nyelv modernizációjának legalapvetőbb tényezője és feltétele" (Kiefer 1999), igyekeztünk áttekinteni a magyar nyelvvel kapcsolatos nyelvtechnológiai kutatások eddigi fontosabb eredményeit. Több részterületen is szép eredmények születtek, bár a magyar nyelv sajátosságai nem tették lehetővé a nagyobb nyugat-európai nyelvekre kidolgozott technológiai megoldások egyszerű adaptálását. Bár az akadémiai kutatók és az üzleti alapon működő nyelvtechnológiai kutatóhelyek (pl. a MorphoLogic) tevékenységét a 2000-es évek első felében több K+F-pályázati lehetőség támogatta, napjainkra ezt a területet lényegesen nagyobb mértékben kellene támogatnia egy központi szándéknak, hiszen amint az MTA korábbi elnöke, Glatz Ferenc (1999) írta: „a kis nyelvek korszerűsítési programja sohasem történhet üzleti alapon: nem kifizetődő befektetés".

Irodalom

Alberti, Gábor 2011. ReALIS: interpretálók a világban, világok az interpretálóban. Budapest: Akadémiai Kiadó.

Alexin Zoltán – Csendes Dóra (szerk.) 2003. Az I. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem.

Alexin Zoltán – Csendes Dóra (szerk.) 2004. A II. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem.

Alexin Zoltán – Csendes Dóra (szerk.) 2005. A III. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem.

Babarczy Anna – Gábor Bálint – Hamp Gábor – Kárpáti András – Rung András – Szakadát István 2005. Hunpars: mondattani elemző alkalmazás. In: Alexin – Csendes (2005, 20–28).

Clark, Bob 2000. MoBiMouse, the world's first "no-click" dictionary program. International Journal of Language and Documentation 3: 26–27.

Csendes Dóra – Alexin Zoltán – Csirik János – Kocsor András 2005. A Szeged Korpusz és Treebank verzióinak története. In: Alexin – Csendes (2005, 409–412).

Dimitrova, Ludmila – Tomaz Erjavec – Nancy Ide – Heiki-Jan Kaalep – Vladimir Petkevic – Dan Tufis 1998. Multext-East: Parallel and comparable corpora and lexicons for six Central and Eastern European languages. In: Christian Boitet – Pete Whitelock (szerk.): Proceedings of the COLING-ACL 98. Montreal: Morgan Kaufman. 315–319.

Dömölki, Bálint 1964. An algorithm for syntactic analysis. Computational Linguistics 3: 19–46.

Ehmann Bea – Lendvai Piroska – Fritz Adorján – Miháltz Márton – Tihanyi László 2011. Szemantikus szerepek vizsgálata magyar nyelvű szövegek narratív pszichológiai elemzésében. In: Tanács – Vincze (2011, 223–230).

Elekfi László 1994. Magyar ragozási szótár. Budapest: MTA Nyelvtudományi Intézet.

Fegyó, Tibor – Péter Mihajlik – Péter Tatai 2003. Comparative study on Hungarian acoustic model sets and training methods. In: Jean Cedric Chappelier (szerk.): Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva: ACL. 829–832.

Fék Márk – Németh Géza – Olaszy Gábor 2004. Megértést segítő részletező gépi névfelolvasás magyar nyelvre. In: Alexin – Csendes (2004, 301–306).

Glatz Ferenc (szerk.) 1999. A magyar nyelv az informatika korában. Budapest: MTA.

Gordos Géza – Takács György 1983. Digitális beszédfeldolgozás. Budapest: Műszaki Kiadó.

Halácsy Péter – Kornai András – Németh László – Sass Bálint – Varga Dániel – Váradi Tamás – Vonyó Attila 2005. A Hunglish korpusz és szótár. In: Alexin – Csendes (2005, 134–142).

Halácsy, Péter – András Kornai – Csaba Oravecz – Viktor Trón – Dániel Varga 2006. Using a morphological analyzer in high precision POS tagging of Hungarian. In: Nicoletta Calzolari – Khalid Choukri (szerk.): Proceedings of LREC-2006. 2245–2248. http://www.lrec-conf.org/proceedings/lrec2006

Hell, György 1975. Generation of nominal constructions in Hungarian. Computational Linguistics 9: 73–78.

Hunyadi László. 2011. Az ember–gép kommunikáció elméleti-technológiai modellje. Háttér és alapkérdések. In: Bódog Alexa (szerk.): Az ember–gép kommunikáció technológiájának elméleti alapjai. IKUT zárókötet. Debreceni: Debreceni Egyetemi Kiadó. 6–12.

Jánoska Sándor 1967. A magyar ige automatikus toldalékolásának egy modellje. Nyelvtudományi Értekezések 58: 464–468.

Karacs, Kristóf – Gábor Prószéky – Tamás Roska 2009. Cellular wave computer algorithms with spatial semantic embedding for handwritten text recognition. International Journal of Circuit Theory and Applications 37: 1019–1050.

Kiefer Ferenc 1999. Néhány gondolat a nyelvi technológiákról. In: Glatz (1999, 128–132).

Kiss, Gabriella – Margit Kiss – Júlia Pajzs 2001. Normalisation of Hungarian archaic texts. In: Paul Rayson (szerk.): Papers in computational lexicography (COMPLEX-01). Birmingham: University of Birmingham. 83–95.

Klauszer Judit 1965. A magyar főnevek szintézisének kérdéséhez. Általános Nyelvészeti Tanulmányok 3: 117–129.

Kónyi Sándor 1965. A magyar főnevek elemzése. Általános Nyelvészeti Tanulmányok 3: 131–143.

Kornai András 1986. Szótári adatbázis az akadémiai nagyszámítógépen. Műhelymunkák a nyelvészet és társtudományai köréből 2: 65–79.

Kornai, András 2007. Mathematical linguistics. Dordrecht: Springer.

Koskenniemi, Kimmo 1983. Two-level morphology: A general computational model for word-form recognition and production. Helsinki: University of Helsinki.

Kuba, András – András Hócza – János Csirik 2004. POS tagging of Hungarian with combined statistical and rule-based methods. In: Ivan Kopeček – Karel Pala (szerk.): Proceedings of the Seventh International Conference on Text, Speech and Dialogue (LNAI 3206). Dordrecht: Springer. 113–121.

László János – Ehmann Bea 2004. A narratív pszichológiai tartalomelemzés új eljárása: a LAS VERTICUM. Magyar Pszichológiai Szemle 59: 363–375.

Lugosiné Papp, Mária 1975. One model of the Hungarian verb synthesis. Computational Linguistics 9: 39–97.

Megyesi, Beáta 1999. Improving Brill's PoS tagger for an agglutinative language. In: Pacale Fung – Joe Zhou (szerk.): Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. New Brunswick NJ: Association for Computational Linguistics. 275–284.

Melcsuk Igor 1967. A magyar főnévragozás egy modellje. Nyelvtudományi Értekezések 58: 499–502.

Merényi Csaba 2005. A MetaMorpho magyar–angol gépi fordító rendszer igei vonzatkereteit működtető nyelvtan. In: Alexin – Csendes (2005, 108–115).

Mihajlik, Péter – Tibor Révész – Péter Tatai 2002. Phonetic transcription in automatic speech transcription. Acta Linguistica Hungarica 49: 407–425.

Miháltz, Márton – Csaba Hatvani – Judit Kuti – György Szarvas – János Csirik – Gábor Prószéky – Tamás Váradi 2008. Methods and results of the Hungarian WordNet project. In: Attila Tanács – Dóra Csendes – Veronika Vincze – Christiane Fellbaum – Piek Vossen (szerk.): Proceedings of the Fourth Global WordNet Conference. Szeged: University of Szeged. 311–321.

Miller, George A. – Richard Beckwith – Christiane Fellbaum – Derek Gross – Katherine J. Miller 1990. Introduction to WordNet: An on-line lexical database. International Journal of Lexicography 3: 235–244.

Naszódi Mátyás 1997. Nyelvhelyesség-ellenőrzés számítógéppel (parciális szintaxis). In: Polyák Ildikó (szerk.): Hetedik Országos Alkalmazott Nyelvészeti Konferencia. Budapest: Külkereskedelmi Főiskola. 256–260.

Nemes Zoltán 1941. Szóstatisztika egymillió szótagot felölelő újságszövegek alapján. In: Az Egységes Magyar Gyorsírás Könyvtára. Szeged: Gyorsírási Ügyek M. Kir. Kormánybiztossága. 190.

Németh Géza – Olaszy Gábor (szerk.) 2010. A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek. Budapest: Akadémiai Kiadó.

Németh, Géza – Csaba Zainkó – László Fekete – Gábor Olaszy – Gábor Endrédi – Péter Olaszi – Géza Kiss – Péter Kis 2000. The design, implementation and operation of a Hungarian e-mail reader. International Journal of Speech Technology 3: 217–236.

Németh, Géza – Csaba Zainkó – Géza Kiss – Gábor Olaszy – Géza Gordos 2003. Language processing for name and address reading in Hungarian. In: Proceedings of IEEE International Conference of Natural Language Processing and Knowledge Engineering. Beijing: IEEE. 238–243.

Novák Attila 2003. Milyen a jó Humor? In: Alexin – Csendes (2003, 138–145).

Novák Attila – Endrédy István 2005. Automatikus zárt ë-jelölő program. In: Alexin – Csendes (2005, 453–454).

Novák Attila – M. Pintér Tibor 2006. Milyen a még jobb Humor? In: Alexin Zoltán – Csendes Dóra (szerk.): A IV. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 60–69.

Olaszy Gábor 1999. Beszédadatbázisok készítése gépi beszéd-előállításhoz. In: Gósy Mária (szerk.): Beszédkutatás 1999. Budapest: MTA Nyelvtudományi Intézet. 68–89.

Olaszy, Gábor 2002. Predicting Hungarian sound durations for continuous speech. Acta Linguistica Hungarica 49: 321–345.

Olaszy, Gábor – Géza Németh – Péter Olaszi – Géza Kiss – Csaba Zainkó – Géza Gordos 2000. Profivox: A Hungarian text-to-speech system for telecommunication applications. International Journal of Speech Technology 3/4: 201–215.

Oravecz, Csaba – Péter Dienes 2002. Efficient stochastic part-of-speech tagging for Hungarian. In: M. Gonzalez Rodriguez – C. P. Suarez Araujo (szerk.): Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas, Spain. 710–717. http://www.lrec-conf.org/proceedings/lrec2002

Orosz, György 2011. Investigating Hungarian POS-tagging methods. In: Tamás Roska (szerk.): Proceedings of the Multidisciplinary Doctoral School 2010–2011 academic year. Budapest: Pázmány University ePress. 77–81.

Pajzs Júlia 1983. A magyar szavak morfológiai szintézise számítógéppel. Szakdolgozat, ELTE.

Pajzs Júlia 1990. Számítógép és lexikográfia. Budapest: MTA Nyelvtudományi Intézet.

Pajzs Júlia – Kiss Gabriella – Kiss Margit 2004. A Nagyszótár történeti korpuszának elemzéséről. Magyar Nyelv 100: 185–191.

Papp Ferenc 1969. A magyar nyelv szóvégmutató szótára. Budapest: Akadémiai Kiadó.

Papp Ferenc 1975. A magyar főnév paradigmatikus rendszere. Budapest: Akadémiai Kiadó.

Papp Ferenc 2000. A Debreceni Thészaurusz (Linguistica Series C, Relationes 11). Budapest: MTA Nyelvtudományi Intézet.

Prószéky Gábor 1985. Magyar szövegek számítógépes morfológiai elemzése (A Nagyszótár számára rögzített folyamatos szövegek szövegszavainak tő- és toldalékmorfémákra való bontását megvalósító automata terve). Kézirat. Budapest: MTA Nyelvtudományi Intézet.

Prószéky Gábor 1989. Számítógépes nyelvészet (Természetes nyelvek használata számítógépes rendszerekben). Budapest: SZÁMALK.

Prószéky, Gábor 1996. Syntax as meta-morphology. In: Jun-ichi Tsujii (szerk.): Proceedings of the 16th International Conference on Computational Linguistics. Vol. 2. Copenhagen: Center for Sprogteknologi. 1123–1126.

Prószéky Gábor 2000. A magyar morfológia számítógépes kezelése. In: Ferenc Kiefer (szerk.): Strukturális magyar nyelvtan 3. Morfológia. Budapest: Akadémiai Kiadó. 1024–1065.

Prószéky Gábor 2001. A nyelvtechnológia és a modern nyelvészet viszonyáról. In: Andor József – Szűcs Tibor – Terts István (szerk.): Színes eszmék nem alszanak… Szépe György 70. születésnapjára. Pécs: Lingua Franca Csoport. 991–998.

Prószéky Gábor 2003. Automatikus információszerzés gazdasági rövidhírekből. In: Patkós Anna (szerk.): Információs és kommunikációs technológiák. Budapest: Oktatási Minisztérium Kutatás-fejlesztési Helyettes Államtitkárság. 28–38.

Prószéky Gábor 2012. A nyelvtechnológia és a magyar nyelvtudomány. Magyar Nyelv 108: 1–18.

Prószéky Gábor – Kis Balázs 1999. Számítógéppel emberi nyelven. Természetes nyelvi feladatok megoldása számítógéppel. Bicske: SZAK.

Prószéky Gábor – Kis Balázs 2002. Context-sensitive dictionaries. In: Tseng et al. (2002, 1268–1272).

Prószéky, Gábor – Zoltán Kiss – Lajos Tóth 1982. Morphological and morphonological analysis of Hungarian word forms by computer. Computational Linguistics and Computer Languages 15: 195–228.

Prószéky, Gábor – Csaba Merényi 2012. Language technology methods inspired by an agglutinative, free phrase-order language. In: Walther von Hahn – Cristina Vertan (szerk.): Multilingual processing in Eastern and Southern EU languages: Low-resourced technologies and translation. Cambridge: Cambridge University Press. 182–206.

Prószéky, Gábor – Márton Miháltz 2002. Automatism and user interaction: Building a Hungarian WordNet. In: Antonio Zampolli (szerk.): Proceedings of the 3rd International Conference on Language Resources and Evaluation. Vol. II. Las Palmas: ELRA. 957–961.

Prószéky Gábor – Miháltz Márton 2008. Magyar WordNet: az első magyar lexikális szemantikai adatbázis. Magyar Terminológia 1: 43–57.

Prószéky, Gábor – Mátyás Naszódi – Balázs Kis 2002. Recognition assistance. In: Tseng et al. (2002, 1263–1267).

Prószéky, Gábor – Attila Novák 2005. Computational morphologies for small Uralic languages. In: Antti Arppe – Lauri Carlson – Krister Linden – Jussi Piitulainen – Mickael Suominen – Martti Vainio – Hanna Westerlund – Anssi Yli-Jyrä (szerk.): Inquiries into words, constraints and contexts (Festschrift for Kimmo Koskenniemi on his 60th birthday). Stanford: CSLI Publications. 116–125.

Prószéky Gábor – Olaszy Gábor – Váradi Tamás 2006. Nyelvtechnológia. In: Kiefer Ferenc (szerk.): Magyar nyelv. Budapest: Akadémiai Kiadó. 1038–1072.

Prószéky, Gábor – László Tihanyi 1993. Helyette: Inflectional thesaurus for agglutinative languages. In: Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics. Utrecht: ACL. 473.

Prószéky, Gábor – László Tihanyi 2002. MetaMorpho: A pattern-based machine translation system. In: Proceedings of the 24th ASLIB Conference. London: ASLIB. 19–24.

Prószéky, Gábor – László Tihanyi – Gábor Ugray 2004. Moose: A robust high-performance parser and generator. In: John Hutchins – Michael Rosner (szerk.): Proceedings of the 9th Workshop of the European Association for Machine Translation. La Valletta, Malta: Foundation for International Studies. 138–142.

Prószéky Gábor – Tóth Lajos 1979. Magyar nyelvű mondatok számítógépes szintaktikai elemzése. Budapest: ELTE.

Sántáné-Tóth, Edit – Péter Szeredi 1982. PROLOG applications in Hungary. In: K. L. Clark – S.-Å. Tärnlund (szerk.): Logic programming. New York: Academic Press. 19–32.

Senellart, Jean – Péter Dienes – Tamás Váradi 2001. New generation systran translation system. In: John Hutchins (szerk.): Proceedings of the Eighth Machine Translation Summit. Santiago de Compostela: EAMT. 311–316.

Seregy Lajos 1991. NyelvÉsz (Számítógépes helyesírás-ellenőrző és -javító program). Édes Anyanyelvünk 3: 6–7.

Silberztein, Max 1993. Dictionnaires électroniques et analyse automatique de textes: le système INTEX. Paris: Masson.

Stein, Mária 1966. Synthese des ungarischen Hauptwortes mit elektronischen Rechenmaschine. Computational Linguistics 5: 169–176.

Szabó Martina Katalin – Schmalcz András – Nagy T. István – Vincze Veronika 2011. A HunOr magyar–orosz párhuzamos korpusz. In: Tanács – Vincze (2011, 341–347).

Takács György – Tihanyi Attila – Bárdi Tamás – Feldhoffer Gergely – Srancsik Bálint 2006. Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére. Híradástechnika 66: 31–3.

Tanács Attila – Vincze Veronika (szerk.) 2011. A VIII. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem.

Tihanyi László 2003. A MetaMorpho projekt története. In: Alexin – Csendes (2003, 247–25).

Trón, Viktor – László Németh – Péter Halácsy – András Kornai – György Gyepesi – Dániel Varga 2005. Hunmorph: Open source word analysis. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor: ACL. 77–85.

Tseng, Shu-Chuan – Tsuei-Er Chen – Yi-Fen Liu (szerk.) 2002. Proceedings of the 19th International Conference on Computational Linguistic. Vol. II. Taipei: Academia Sinica.

Tóth, Krisztina – Richárd Farkas – András Kocsor 2008. Hybrid algorithm for sentence alignment of Hungarian–English parallel corpora. Acta Cybernetica 18: 463–478.

Vajda Péter – Nagy Viktor – Dancsecs Erzsébet 2004. A Ragozási szótártól a NooJ morfológiai moduljáig. In: Alexin – Csendes (2004, 183–190).

Váradi, Tamás 1999. On developing the Hungarian National Corpus. In: Špela Vintar (szerk.): Proceedings of the Workshop "Language Technologies – Multilingual Aspects". Ljubljana: Societas Linguistica Europea. 57–63.

Vargha Dénes 1963. Morfológiai elemzés a szukcesszív behatárolás módszerével. In: Gépi fordítás. Algoritmusok orosz nyelvű szövegek elemzésére. Budapest: Országos Műszaki Könyvtár és Dokumentációs Központ. 244–271.

Vásárhelyi István 1975. Magyar igealakok szintézise. Nyelvtudományi Közlemények 77: 67–92.

Veenker, Wolfgang 1968. Verzeichnis des ungarischen Suffixe und Suffixkombinationen. Mitteilungen der Sozietas Uralo-Altaica Heft 3. Hamburg: Sozietas Uralo-Altaica.

Vicsi Klára 2001. Beszédadatbázisok a gépi beszédfelismerés segítésére. Híradástechnika 2001/1: 5–13.

Vicsi Klára – Kocsor András – Teleki Csaba – Tóth László 2004. Beszédadatbázis irodai számítógép-felhasználói környezetben. In: Alexin – Csendes (2004, 307–311).

Vicsi Klára – Tóth László – Kocsor András – Gordos Géza – Csirik János 2002. MTBA – magyar nyelvű telefonbeszéd-adatbázis. Híradástechnika 57: 35–43.

Vicsi Klára – Vig Attila 1998. Az első magyar nyelvű beszédadatbázis. In: Gósy Mária (szerk.): Beszédkutatás 1998. Budapest: MTA Nyelvtudományi Intézet. 163–178.

Vincze Veronika – Szauter Dóra – Almási Attila – Móra György – Alexin Zoltán – Csirik János 2009. A Szeged Treebank függőségi fa formátumban. In: Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): A VI. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 127–138.

Wołosz, Robert 2005. Efektywna metoda analizy i syntezy morfologicznej w języku polskim. Warszawa: Akademicka Oficyna Wydawnicza EXIT.

A historical overview of computational linguistics in Hungary

Abstract: Language technology and speech technology are two large fields within a complex set of disciplines that used to be called computational linguistics and that covers natural language processing, the interface area between computer science and the study of human language/human speech. The present paper tries to summarize the development of language technologies in Hungary, proceeding topic by topic and, as far as possible, in a temporal sequence within each topic. After a general introduction, we survey research results in computational morphology and computational syntax, then we turn to corpus linguistics, computational lexicography, and machine translation. The present overview of the earlier periods is based on Prószéky (1989); that of more recent developments is partly based on Prószéky–Olaszy–Váradi (2006).

Keywords: computational linguistics, historical overview, speech and language technology, Hungarian LT applications, structures of language

Egy általános célú morfológiai annotáció

Rebrus Péter[1] – Kornai András[2] – Varga Dániel[3]

[1]*Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Budapest*
[2]*MTA SZTAKI, Budapest*
[3]*BME MOKK, Budapest*
rerbrus@nytud.hu; kornai@sztaki.hu; daniel@mokk.bme.hu

A morfológiai annotáció célja az adott szóalakban lévő morfoszintaktikai információk megjelenítése; a kizárólag a jelentésre vagy a hangalakra/írásképre vonatkozó információk nem tartoznak bele. A direkt morfalapú, tehát az allomorfiára tekintettel levő annotációs rendszerek beleütköznek a morfológiai szegmentálás bizonytalanságába, amit a fúziós és szuppletív alakok létezése még tovább bonyolít. Ezzel szemben a jelen cikkben a morfológiai annotáció egy bináris morfoszintaktikai jegyekből és ezek pozitív vagy negatív értékeiből álló fastruktúrára épülő formalizmusát javasoljuk, amelyben csak a pozitív értékkel rendelkező jegy–érték párok csomópontjai dominálnak más csomópontokat. Ez lehetővé teszi, hogy a bináris jegyes hierarchikus szerkezet unáris jegyessé alakítható legyen, s így közvetlenül tükrözze az adott alak morfológiai jelöltségének mértékét. A cikk befejező része az itt bemutatott morfológiai reprezentációt felhasználó, már megvalósult gyakorlati alkalmazásokból mutat be egy csokorra valót.

Kulcsszavak: magyar nyelv, morfológia, annotáció, inflexió

1. Bevezetés

Cikkünk a morfológiai annotáció általános kérdéseit tárgyalja a magyar nyelv példáján keresztül. Az első részben a morfémákra közvetlenül támaszkodó konkrét annotációs sémák és a rögzített kódhosszúságú rendszerek problémáit írjuk le. A második részben a magyar főnévi, igei és egyéb inflexiós paradigmák részletes kódolásáról írunk. Annotációs rendszerünk, a **hunmorph** az említett kódolásokkal szemben absztrakt, változó kódhosszt használó rendszer, amelynek alapelvei teljesen általánosak és nyelvfüggetlenek. A harmadik részben pedig röviden érintjük a deriváció és a szóösszetétel kezelését. Cikkünk záró részében az annotációs rendszert használó nyílt forráskódú számítógépes nyelvészeti eszközöket ismertetjük.

2. Az allomorfalapú annotáció problémái

Kiindulópontunk az, hogy a morfológiai annotáció elsődleges célja az adott szóalakban levő **morfoszintaktikai** információk megjelenítése. Morfoszintaktikainak tekintjük a szóalakban meglevő olyan információkat, amelyeknek közvetlen szintaktikai hatása van, azaz amelyek az adott szóalak mondatbeli formai viselkedését (disztribúcióját) befolyásolják – ilyen elsősorban az a szintaktikai pozíció, ahol a szóalak a grammatikus mondatban megjelenhet, illetve az egyeztetés, amikor egy szóalak morfológiai jegyei befolyásolják egy másik szóalak morfológiai jegyeit. Ennélfogva az alábbi módszertani elvet követtük: a kizárólag a jelentésre és a hangalakra (vagy az írásképre) vonatkozó információk nem részei a morfoszintaktikai reprezentációnak. Egyes esetekben a szemantikai és a szintaktikai információk éles elkülönítése nehézségekbe ütközik, ezért egy általános célú morfológiai annotáció tervezésekor mérlegelnünk kell, hogy a potenciális alkalmazások számára mely szemantikai információk lehetnek lényegesek. A fenti módszertani elvet azért is érdemes szem előtt tartani, mert az annotáció elveinek transzparenseknek kell lenniük: egy formai tulajdonságot bárkinek könnyű betanítani (és így emberi erőforrás segítségével előállítani egy nagy pontossággal címkézett korpuszt), míg a szemantikai tulajdonságok nagy részére ez nem áll. Az egyszerre szintaktikai és szemantikai tulajdonságokon alapuló ilyan jegyek körébe tartozik többek között a főnév–melléknév megkülönböztetés vagy igéknél a modális (ható ige) és a múlt idő, amelyeket annotációs rendszerünk is megkülönböztet (erről l. később).

2.1. Allomorfia

A hangalakra (fonológiai formára) vagy az írásképre vonatkozó információknak a morfológiai annotációban való megjelenítése azért sem lenne szerencsés, mert nagyon gyakran önkényes döntéseket kellene hozni arról, hogy milyen alakot adjunk meg allomorfia esetén (azaz akkor, ha az adott morféma több alakban jelenhet meg). Vegyünk néhány példát: a *fára* alak a következő információkat hordozza: (i) lemmája: FA, (ii) morfoszintaktikai jegyei: **SZÁM**: *EGYES*, **BIRTOKOS**: *NINCS*, **BIRTOK**: *NINCS*, **ESET**: *SUBLATIVUS*. Ha az annotációban ezeken a jegyeken túl azt is meg akarnánk jeleníteni, hogy a szóban forgó *fára* alakban a tőbeli magánhangzó hosszú (szemben más alakokkal, pl. ilyen a *fa* tőalak, a toldalékolt *faként* alak, vagy a *facipő* szóösszetétel), akkor az elemzésben a tövet esetleg a `fá` és nem a `fa` alakban adhatnánk meg. A szóban forgó *fára* alakban jelenlevő toldalék azonban elölképzett magánhangzójú változatban is megjelen-

het (pl. *kép-**re***), így dönthetnénk úgy is, hogy az esetragnak ezt a jellegzetességét az annotációnak tükröznie kell, azaz valamilyen alulspecifikált alakban adhatnánk meg a toldalékot (pl. `-rA`, ahol a nagy `A` szimbólum a középnyílt elülső *e* és hátulsó *a* magánhangzók helyett áll). Hasonló a helyzet máskor is, ahol a szóalakban szereplő morfémák allomorfiát szenvednek el. Például a *szelek* vagy a *sarki* alakokban szintén tőallomorfiát találunk: *szél – szelek*, *sarok – sarki*, sőt az első esetben a többes szám jelölője más szóalakokban más és más alakban jelenhet meg (pl. *kár-**ok***, *ház-**ak***, *sün-**ök***, *zokni-**k***), ezért ennek a morfémának a jelölése sem nyilvánvaló (lehetne az előzőhöz hasonlóan alulspecifikált magánhangzóval `-Vk` vagy magánhangzó nélkül csupán `-k`).

Látható tehát, hogy ha a morfológiai annotációt az allomorfokkal vagy az allomorfoknak valamilyen absztrakt alakjával adjuk meg, akkor az esetek jelentős részében legalább három megoldást követhetünk (természetesen lehetségesek kevert megoldások is): (a) a **konkrét** elemzésben az adott szóalakban megjelenő allomorfokat (tulajdonképpen a teljes sztringet eredeti formájában) szerepeltetjük (pl. `fá+ra`, `szel+ek` és `sark+i`); (b) az **allomorfiamentes** elemzésben az allomorfok közül a leggyakoribbat vagy az alapallomorfot választjuk ki (ilyen, amikor a `fa`, a `szél`, és a `sarok` töveket adjuk meg a *fára, szelek*, illetve *sarki* alakok elemzésénél); és (c) az **absztrakt** elemzés, ahol allomorfia esetén az összes allomorfot lehetőség szerint szerepeltetjük: ez a `-rA` és a `-Vk` toldalékok vagy a `fÁ`, `szÉl` és a `sarOk` tövek esete, ahol egy alulspecifikált (nagybetűs) szimbólum mutatja a váltakozás helyszíneit (ez lehet nyúlás, rövidülés, hangkivetés, magánhangzó-harmónia stb.). Ez az alulspecifikációs megoldás azonban nem mindig lehetséges: vannak az allomorfiának olyan esetei, amelyekben a váltakozó szekvencia nem adható meg alulspecifikált szimbólummal: ilyen ún. nemfonológiai allomorfiákat találunk az igei paradigmában, ha a toldalékváltozatok között nincs fonológiai kapcsolat (sőt gyakran a szekvenciák hossza sem azonos): ilyen az E.2 alakokban a tővégtől függő *sz~ol/el/öl* váltakozás (pl. *kap-**sz*** ~ *mos-**ol***) vagy az E.3 definit alakokban a tő hangrendjétől függő *ja~i* váltakozás (pl. *lop-**ja*** ~ *lep-**i***).

Egy további probléma a tő- és toldalékallomorfok azonos alakúságával függ össze: a *szelek* tőalakja a *szél*, viszont van egy másik, nem-rövidülő magánhangzót tartalmazó azonos alakú lexéma, vö. *szél – szélek*. Hasonló igaz a *sarki* alakra: ez lehet a SAROK lexémához tartozó (pl. *sarki bolt*), de lehet a SARK lexémához tartozó is (pl. *sarki expedíció*). Tehát ha ezek az alakok önmagukban utalnának az aktuális lexémára, az nem lenne elegendő (ez természetesen más, nem allomorfikus esetben is így van, ekkor a lexémákat a lexikográfiai gyakorlatban sorszámok használatával – pl. ÁR1, ÁR2, ÁR3 – különítik el egymástól). Hasonló homonímiajelenségek léphetnek fel a toldalékokban is: a *-k* toldalék nemcsak névszók többes

számára utalhat (pl. *ház-a**k***), hanem igéknél az E.1 (pl. én *kap-j-a**k***) és bizonyos esetekben a T.3 (pl. ők *kap-t-a**k***) szám/személyre is. Ugyanígy az *-i* toldalék nemcsak melléknévképző lehet (pl. *sarki*), hanem utalhat a birtok többes számára is (pl. *hajó-**i**, Pál-é-**i***). Tehát a morfoszintaktikai kódoláshoz a toldalékok alakja sem ad elégséges információt. Tanulságos összjátékot mutat a homonímia és a nemfonológiai allomorfia az olyan alakoknál, mint amilyen az indefinit E.1 *kap-ta-m*, ahol a „szokásos" indefinit E.1 *-k* toldalék helyett *-m* toldalékot találunk. Ekkor a „konkrét" elemzésben (`kap+t+am`) az *-m* szerepel, ami félrevezető lehet, hiszen az *-m* a szokásos **definit** E.1 toldalékkal azonos (pl. *kap-om, kap-j-am (azt)*). Az allomorfiamentes elemzésben ezzel szemben a „szokásos" *-k* **indefinit** E.1. végződés szerepelne (`kap+t+k`), ami szintén félrevezető, hiszen egy E.3 alak pontosan ennek megfelelő alakú (*ők kaptak*). Az absztrakt elemzésben viszont szerepeltetni kellene mindkét allomorfot, hiszen a *-k* és *-m* toldalékallomorfokat nem lehet értelmes módon alulspecifikálni: `kap+t+m/k`. Az alábbi táblázatban összefoglaljuk az említett alakok háromféle elemzését (félkövérrel jelölve a problémás eseteket; a kérdőjel a többféle lehetséges elemzést jelöli valódi morfoszintaktikai különbség nélkül).

(1) Főbb elemzési lehetőségek allomorfia esetén

	a. konkrét elemzés	b. allomorfiamentes elemzés	c. absztrakt elemzés
fára	`fá+ra`	`fa+ra`	`fÁ+rA`
szelek	`szel+ek`	**`szél+k`**	**`szEl+Vk`**
sarki	**`sark+i`**	`sarok+i/sark+i`	`sarOk+i/sark+i`
kaptam	**`kap+t+am`**	`kap+t+k/kap+t+m`	**`kap+t+m?k`**

Az imént bemutatott allomorfia-alapú elemzések tehát több szempontból problematikusak: (i) nincs módszertani eszközünk arra, hogy **eldöntsük**, hogy a három ideáltípus közül melyik elemzési módot kövessük (pl. `fá+ra` vagy `fa+ra` avagy `fÁ+rA`); (ii) az absztrakt (és részben az allomorfiamentes) annotáció használata mögött hallgatólagosan olyan vitatott elemzések és így nyelvészeti **elméletek** kaphatnak szerepet, amelyekről a nyelvtudománynak nincs egységes álláspontja (pl. a kötőhangzó része-e a toldaléknak vagy sem, vagy nemfonológiai allomorfia esetén mely allomorfo(ka)t szerepeltessük); (iii) egyes elemzések összemossák a tő- vagy a toldalékallomorfokban potenciálisan jelen levő **homonímiákat**, így önmagukban nem elegendőek a szóalakban levő morfológiai információk megadásához (l. pl. a *szélek*, a *sarki* és a *kaptam vmit* alakok fenti esetét).

2.2. Szegmentálás

A fentebb bemutatott annotációs megközelítéseknek egy további súlyos következménnyel is szembesülniük kell, ez pedig a morfológiai **szegmentálás** bizonytalansága. A morfalapú annotációnak tartalmaznia kell egy határjelölőt, amely elválasztja a morfokat egymástól (a fenti hipotetikus elemzésekben erre a célra a + szimbólumot választottuk). Ez az elválasztás azonban sok esetben önkényes és nem ritkán problémákba ütközik. Lássunk néhány példát! A problémás esetek első típusa az írásképpel kapcsolatos. A grafémikus alakban a kettőzött digráfok speciális írásmódja miatt nem lehetséges az eredeti szóalak karaktereit megfelelő módon elválasztani; ez történik pl. a *hússzor, ésszerű* stb. alakok elemzésénél: a konkrét elemzésben a kettőzött digráfot meg kell osztani a tő és a toldalék között, ami félrevezető (pl. `hús+szor`, `és+szerű`); az absztraktabb elemzésben viszont nem pontosan a szóalak karakterei találhatók (pl. `húsz+szor`, `ész+szerű`). Hasonló a helyzet akkor, ha a szóalak kettős mássalhangzóra végződik, és a toldalék ugyanezzel a mássalhangzóval kezdődik (pl. *szebből, halottal*). A következő problematikus típus a morfhatáron lezajló hasonulásokkal kapcsolatos. Így például a *-val/-vel* toldaléknak vagy a felszólító mód *-j* toldalékának egyes mássalhangzó utáni változatai esetén nem világos a szegmentálás (`hát+tal`, `hátt+al` vagy `hát+val`, illetve `fus+sa` vagy `fut+ja`). Ha a tő digráfra végződik, akkor a két említett probléma együtt jelentkezik: pl. `ác+csal`, `áccs+al`, `ács+csal` vagy `ács+val`, illetve `ed+dze`, `edz+dze` vagy talán `edz+je`. A következő táblázatban ezeket az elemzési lehetőségeket foglaltuk össze.

(2) Szegmentálási lehetőségek különböző elemzések esetén

	a. eredeti sztring	a.′ átelemzett sztring	b. allomorfiamentes
hússzor	**hús+szor**	húsz+szor	
szebből	**szeb+ből**	szebb+ből	
háttal	**hát+tal hátt+al**		hát+val
áccsal	**ác+csal áccs+al**	ács+csal	ács+val
fussa	fus+s+a		**fut+j+a**
eddz	**ed+dz**	**edz+dz**	edz+j

A fentiekhez hasonló technikai problémákkal minden morfológiai elemzőprogramnak meg kell birkóznia. Az, hogy egy elemző technikailag melyik módszert követi az aktuális szóalakok (sztringek) manipulációja során, az elemzőprogram (és az erőforrások) felépítésétől, lehetőségeitől függ, és nem morfoszintaktikai információ, amelyet a végső kódolásnak meg kell jelenítenie. Azaz a szegmentálás és az allomorfok kiválasztása az elemző belügye, és **nem** lehet **része** a morfoszintaktikai **annotációnak**.

2.3. Fúzió és szuppletivizmus

Ki kell térnünk egy további problémakörre, amely azt is megmutatja, hogy egyes esetekben a sztringalapú elemzést nem is lehetséges ésszerű módon megvalósítani. Az ún. **fúziós morfémák** esetén több funkció szételemezhetetlenül társul egy morfhoz; a legismertebb példa a magyarban a birtokos alakok és igék szám/személy jelölése. Az igazán problematikus esetek azonban azok, amikor a fúzió csak bizonyos esetekben áll fenn, máskor a morfémák agglutinatív módon jelennek meg. Ezt a jelenséget láthatjuk az igei definitjelölésnél: E.1 és E.2 egyértelműen fúziós (pl. *ad-om, ad-od*), E.3 és T.2 agglutinatív (*ad-ja, ad-já-tok*). Néha még az is előfordul, hogy az E.3 definitjelölés a módjelölővel fuzionál, pl. az *ad-ná* alakban a *-ná* toldalék együtt fejezi ki a feltételes módot és a definitséget, tehát ezek az allomorfok nem alkalmasak a morfológiai annotáció jelölésére. A fúzióhoz tartozó jelenség az is, amikor a toldaléktömb formailag szételemezhető, viszont nem egyértelmű, hogy mely funkcióhoz mely szekvenciák tartoznak; ilyen a többes számú birtokosjelölős alakok esete (pl. *kalapjaim*), ahol a szételemzett `kalap+ja+i+m` alakban a `ja` szekvencia nem bír morfoszintaktikai szereppel (ennek absztraktabb nyelvészeti elemzését l. Melcsuk 1965).

A sztringalapú elemzés lehetetlenségét az ún. **szuppletív** alakok mutatják leginkább, ahol ugyanazon lexémához tartozó alakok töve teljesen különbözik (pl. *van* vs. *lehet, jön* vs. *gyere, sok* vs. *több* stb.). Hasonló jelenség lép fel egyes kis zárt szóosztályoknál, így a személyes és birtokos névmásoknál: az *engem, téged* stb. accusativusi alakok nem állíthatók elő mint `én+t`, `te+t` stb.; teljes szuppletivizmusra példa a *benneteket, bennünket* alakok, amelyek „tövének” alakja inessivusi, ennek ellenére ezek egyszerű accusativusi alakok: `ti+t`, `mi+t`. Hasonlóan az *enyém, tied* stb. alakok morfoszintaktikailag nem birtokosjelölős, hanem birtokjelölős alakok, tehát morfoszintaktikailag `én+é`, `te+é` elemzést kellene kapniuk (a személyes és birtokos névmások esetéről később részletesen is írunk). A *gyere, gyertek* alakok ugyanígy a JÖN lexémához tartoznak és kötő-felszólító módúak, annak ellenére, hogy alakilag sem a tő, sem az idő/mód jelölő nem látszik.

(3) Fúziós és szuppletív alakok elemzési problémái

	„formai” elemzés	„morfoszintaktikai” elemzés
adná	?`ad+na+a`	`ad+NÁ+JA`
kalapjaim	?`kalap+ja+i+m`	`kalap+K+m`
engem	?`én+m`	`ÉN+T`
enyém	?`én+m`	`ÉN+É`
benneteket	?`benn+etek+et`	`TI+t`
gyere	?`gyer+e`	`JÖN+J`

A következő részben azt tekintjük át, hogy milyen alternatív annotációs megoldás lehetséges.

3. A kizárólag morfoszintaktikai kategóriákon alapuló annotáció

3.1. Általános annotációs elvek

Az előző részben láttuk, hogy a morfológiai annotáció problémájára a megoldás nem a szóalakok (fonológiai vagy grafémikus) formáján alapuló kódolás, hanem egy nyelvészetileg megalapozott morfoszintaktikai kategóriákra épülő formalizmus adhat választ. Egy ilyen elterjedt annotáció az ún. MSD-kódrendszer (*Morphosyntactic Description*, l. Erjavec–Monachini 1997), amelyben a morfoszintaktikai **kód rögzített hosszúságú**: egy jegyértékekből álló sztring, amelynek minden pozíciójához eleve rögzített módon vannak hozzárendelve a jegyek: azaz a pozíciók azt adják meg, hogy mely értéknek a jegyeit töltjük ki. Lássunk néhány példát: a *fiú* és a *fiaitokéinak* főnévi, illetve az *ad* és az *adtátok* igei alakok MSD-kódrendszer szerinti annotációi az alábbiak:

(4) Két főnévi és igealak MSD-annotációja

fiú	`Nc-sn-n---`	*ad*	`Vmip3s---n------`
fiaitokéinak	`Nc-pd-yp2p`	*adtátok*	`Vmis2p---y------`

Amint a példákból is kitűnik, ennek a kódolásnak a hátránya az, hogy egyrészt nehezen kezelhető (rosszul olvasható a sok üresen hagyott érték és az értékek nem vagy csak kevéssé transzparens kódjai miatt). Másrészt nem hierarchikus, azaz az annotáció közvetlenül nem tükrözi az egyes értékek közötti összefüggéseket – például ilyen összefüggés az, hogy csak birtokos alakoknál van szükség a birtokos számának és személyének megjelölésére, vagy az, hogy a magyarban van egy speciális *-lak/-lek* toldalék, amely 2. személyű tárgyra utal, viszont az alanynak E.1-nek kell lennie: pl. *(én) lát**lak** (téged/titeket)*; azaz ez a morfoszintaktikai érték függ az ige szám/személyétől. Harmadrészt nem képes a morfológiai jelöltséget tükrözni: azaz egy formailag és funkcionálisan komplex szóalak (pl. *fiaitokéinak* vagy *adhattátok*) és egy ilyen szempontból jelöletlen szóalak (pl. *fiú* vagy *ad*) annotációja ugyanolyan komplexitású. További problémája az, hogy egyelőre csak inflexiós kódrendszer, és nem nyilvánvaló, hogy a morfoszintaktikailag releváns képzések hogyan illeszthetők bele (különösen igaz ez a szófajváltó képzésekre).

3.2. Jegy–érték szerkezetek

A fenti problémák egy részére megoldást jelent a hierarchikus jegy–érték struktúrák (pl. az ún. AVS-ek, *Attribute-Value Structures*, l. Trón 2002) használata. Az AVS-ek előnye a nyelvészeti és formális megalapozottság: ezt a formalizmust több szintaktikai elmélet használja. A teljesen kitöltött AVS-eknek is problémája azonban az, hogy az annotáció nem tesz különbséget morfológiailag jelölt és jelöletlen szóalakok között. Lássunk egy példát: a fenti *fiaitokéinak* és a *fiú* alak a következő morfoszintaktikai információkat hordozza (itt és a későbbiekben a jegyeket és értékeiket kiskapitálissal jelöltük, ezen belül félkövérrel a jegyeket, és kurzívval az értékeket; a hierarchikus viszonyok jelölésére tabulálást alkalmaztunk).

(5) Két főnévi alak sematikus jegy–érték struktúrája

a. *fiaitokéinak*

LEMMA	*FIÚ*
KATEGÓRIA	*FŐNÉV*
SZÁM	*TÖBBES*
BIRTOKOS	*IGEN*
SZÁMA	*TÖBBES*
SZEMÉLYE	*2.*
BIRTOK	*IGEN*
SZÁMA	*TÖBBES*
ESET	*DATIVUS*

b. *fiú*

LEMMA	*FIÚ*
KATEGÓRIA	*FŐNÉV*
SZÁM	*EGYES*
BIRTOKOS	*NEM*
(SZÁMA	X)
(SZEMÉLYE	X)
BIRTOK	*NEM*
(SZÁMA	X)
ESET	*NOMINATIVUS*

Hasonlóan az említett *adhattátok* és *ad* igék szokásos specifikációja az alábbi.

(6) Két igei alak sematikus jegy–érték struktúrája

a. *adhattátok*

LEMMA	*AD*
KATEGÓRIA	*IGE*
MODÁLIS	*IGEN*
IDŐ	*MÚLT*
MÓD	*KIJELENTŐ*
SZÁM	*TÖBBES*
SZEMÉLY	*2*
DEFINITSÉG	*IGEN*

b. *ad*

LEMMA	*AD*
KATEGÓRIA	*IGE*
MODÁLIS	*NEM*
IDŐ	*JELEN*
MÓD	*KIJELENTŐ*
SZÁM	*EGYES*
SZEMÉLY	*3*
DEFINITSÉG	*NEM*

A fenti (5) és (6) szerkezetekből látható, hogy nincs jelentős különbség a jelölt és jelöletlen alakok jegy–érték struktúrájának „bonyolultsága" között: azok ugyanazokat a jegyeket tartalmazzák. Ez azonban nem intuitív és nem is praktikus, hiszen a morfológiailag jelöletlen alakok általában rövidebbek (több zérusmorfot vagy morfémát tartalmaznak) és jelentősen gyakoribbak (funkciójuk általánosabb, használatuk kiterjedtebb).

3.3. Bináris és unáris jegyek: főnevek

Ha azonban az AVS-ekben a jegyeket úgy fogalmazzuk meg, hogy az értékük csak igen/nem (+/−) lehessen, és az értékek közül szisztematikusan az egyik a jelöltet (szokásosan a +), a másik a jelöletlent (ez általában a −) jelentse, akkor ezen a **bináris** jegyrendszeren jelentős egyszerűsítést tehetünk (l. Kornai 1989). Ha megengedjük további jegyek és hierarchia bevezetését, akkor ezt mindig megtehetjük, hiszen többértékű jegyek esetén ezek értékeit mindig átírhatjuk bináris jeggyé (pl. ilyen a **SZEMÉLY** vagy az **ESET** jegy a főneveknél vagy az **IDŐ** vagy a **MÓD** az igéknél, l. a fenti (5)-öt, ill. (6)-ot).[1]

Lássuk, hogy (5) és (6)-beli példáink milyen bináris jegyszerkezetet kapnak (az újonnan bevezetett jegyek nyelvészeti értelmezéséről l. szintén Kornai 1989-et). Az alábbi (7a)-ban a jelölt főnévi alakot látjuk: itt a legtöbb bináris jegyérték pozitív, míg a (7b)-beli alak esetében az összes másodlagos morfoszintaktikai kategória értéke negatív (az áttekinthetőség érdekében csak a pozitív értékkel bíró jegyek vannak félkövérrel szedve). A (7a) és (7b)-beli AVS-ek ugyanazokat az információkat tartalmazzák, mint az (5a), illetve (5b) szerkezetek. Fontos, hogy a negatív értékkel bíró jegyek alá rendelt jegyeknek semmikor nincs szerepük, ez három esetben állhat elő: (i) az alárendelt jegy negatív értékű domináns jegy esetén nem értelmezhető, vagy (ii) a domináns jegy megfogalmazásából következik, hogy az alárendelt jegy (az adott nyelvben) csak negatív értéket vehet fel, vagy (iii) az adott nyelvben az alárendelt jegy csak a domináns jegy pozitív értéke esetén releváns morfoszintaktikailag. Az (i) esetre példa a **BIRTOKOS** vagy a **BIRTOK** jegyek, amelyek negatív értéke esetén – vagyis ha nincs a főnéven birtokos- vagy birtokjelölés – nincs értelme a birtokos számáról vagy személyéről beszélni (ezt látjuk pl. a *fiú* alak esetén (5b)-ben és (7b)-ben). Egy másik eset a familiáris többes: a *Péterék, szomszédék* stb. alakok morfoszintaktikailag többes számúak, ez azonban egy speciális többes szám: a szóalakban jelölt alakkal familiáris viszonyban álló emberek csoportjára utal; így a **FAMILIÁRIS** jegyet ésszerű a **TÖBBES** jegy alá rendelni (erről részletesen l. Kornai 1989). A (ii) eset akkor áll

[1] Vegyük észre, hogy az alakok helyett a jegyek megcímkézése jelölt és jelöletlen értékekre csak akkor tehető meg, ha egy jegyérték jelöltsége nem függ egy másik jegy értékétől, azaz ebben az értelemben környezetfüggetlen. Ez egyes nyilvánvaló esetekben nem igaz, pl. a jelöltség függhet a lexémától: az ún. relációs főneveknél (pl. *barát, anya* stb.) a birtokos alak jelöletlenebb a nem-birtokos alaknál. További ismert eset a felszólító módú igék: itt a 2. személyű alakok – univerzálisan is – jelöletlenebbek, míg más módban általában a 3. személy jelöletlen (pl. a magyarban is E.2 indefinit alak állhat zérus szám/személyjelölővel (pl. *ad-j*), az E.3 indefinit alak viszont a többi móddal ellentétben toldalékkal áll (*ad-j-**on***). Ezek azonban az egész rendszer szempontjából elhanyagolható mértékű hátrányok: elfogadjuk, hogy a jelöltség jegyértékekre való értelmezésével az alakok jelöltsége jól közelíthető.

elő, ha a jegy megfogalmazásából következik, hogy az alárendelt jegy(ek) negatív értékű domináns jegy esetén egyértelműen csak negatív értékeket vehetnek fel. Ilyen jegy a **NEM-3. SZEMÉLYŰ** és a **NEM NOMINATIVUSI ESETŰ** jegyek, hiszen ha ezek értéke negatív, akkor a birtokos 3. személyű, illetve az eset nominativusi, így az alárendelt jegyeknek (amelyek a további lehetőségeket adják meg) kötelezően negatív értekkel kell rendelkezniük: egy szó a nominativusszal együtt más esettel nem rendelkezhet). A (iii) lehetőségre az igei rendszer bemutatásánál térünk vissza. A (7) ábrában ezeket a „default módon" kitölthető vagy érték nélküli jegy–érték párokat zárójelbe tettük.

(7) Két főnévi alak bináris jegy–érték struktúrája

a. *fiaitokéinak*

fiú	+
FŐNÉV	+
TÖBBES SZÁMÚ	+
FAMILIÁRIS	–
BIRTOKOS	+
TÖBBES SZÁMÚ	+
NEM-3. SZEMÉLYŰ	+
1. SZEMÉLYŰ	–
2. SZEMÉLYŰ	+
BIRTOK	+
TÖBBES SZÁMÚ	+
NEM NOM. ESETŰ	+
ACCUSATIVUS	–
DATIVUS	+
SUPERESSIVUS	–
…	

b. *fiú*

fiú	+
FŐNÉV	+
TÖBBES SZÁMÚ	–
(FAMILIÁRIS	X)
BIRTOKOS	–
(TÖBBES SZÁMÚ	X)
(NEM-3. SZEMÉLYŰ	X)
(1. SZEMÉLYŰ	X)
(2. SZEMÉLYŰ	X)
BIRTOK	–
(TÖBBES SZÁMÚ	X)
NEM NOM. ESETŰ	–
(ACCUSATIVUS	X)
(DATIVUS	X)
(SUPERESSIVUS	X)
(…)	

Vegyük észre, hogy ha a morfoszintaktikai információkat tartalmazó jegyeket rögzítjük, akkor bármilyen negatív értékű jegy redundánssá válik, és elegendő csak a pozitív jegyeket megadnunk. Ezt a tulajdonságot felhasználhatjuk arra, hogy a bináris jegyrendszert egyértékűvé (**unárissá**) tegyük. Ehhez elég a pozitív értékű jegyeket tekintetbe venni, és ha kizárólag ezen jegyek neveit soroljuk fel, akkor teljes értékű annotációt kapunk. Az alábbi (8i) ábrában ez a hierarchikus unáris jegyrendszer látható, amit úgy kaptunk, hogy a (7a,b) bináris jegyrendszerből elhagytuk a negatív értékű jegyeket és a pozitív értékeket. Ezzel az unáris jegyrendszerrel aztán közvetlenül használható annotációs rendszert jön létre: (8ii)-ben a hierarchikus rendszert zárójelek segítségével linearizáltuk (az annotációs formalizmus a következő: a lexémát / jel választja el a morfoszintaktikai annotációtól, ez utóbbi a főkategóriával indul, és utána a további morfoszintaktikai jegyek szerepelnek a hierarchiának megfelelően zárójelezve; az e mögött álló formalizmusról részletesebben a következő részben írunk).

(8) Két főnévi alak elemzése unáris jegyekkel (a redundáns információk nélkül)

a. hierarchikus formában:

fiaitokéinak

fiú
FŐNÉV
  TÖBBES SZÁMÚ (PLUR)
  BIRTOKOS (POSS)
    TÖBBES SZÁMÚ (PLUR)
    NEM-3. SZEMÉLYŰ (--)
      2. SZEMÉLYŰ (2)
  BIRTOK (ANP)
    TÖBBES SZÁMÚ (PLUR)
  NEM NOM. ESETŰ (CAS)
    DATIVUS (DAT)

fiaitokéinak

fiú
FŐNÉV (NOUN)

b. linearizált formában:

fiú/NOUN<PLUR><POSS<PLUR><2>><ANP<PLUR>><CAS<DAT>>

fiú/NOUN

A jegyeknek a legvégső formában látható megnevezései az angol nyelvészeti szakirodalomban elterjedt rövidítéseket követik: PLUR: *plural* (többes szám), POSS: *possessive* (birtokos), ANP: *anaphoric possessive* (birtok), CAS: *case* (eset) stb.). A fent vázolt jegyrendszer úgy van tervezve, hogy a lehető legegyszerűbben feldolgozható formában tükrözze a morfológiai jelöltségi viszonyokat: éppen ezért ahol nem szükséges, ott az adott jegyet elhagytuk; ilyen a **NEM-3. SZEMÉLYŰ** jegy, amelyet a linearizált annotáció nem is jelöl (erre nincs szükség, mert a személyre utaló jegyek amúgy is a **BIRTOKOS** jegy alá vannak rendelve). A birtokos alakok jelölése így egyszerűbbé válik: a POSS jegy alatti személyre utaló jegyek kétfélék lehetnek: <POSS<1>> vagy <POSS<2>>. A 3. személyű birtokos alakokban a POSS jegy az 1 és a 2 jegy nélkül szerepel, azaz jelölése <POSS>, ez egybevág azzal a megfigyeléssel, hogy a három szám/személy közül a 3. a jelöletlen. A birtokosjelölővel ellátott alakok sémája tehát a következő.

(9) Birtokos alakok annotációja

fiam	fiú/NOUN<POSS<1>>
fiad	fiú/NOUN<POSS<2>>
fia	fiú/NOUN<POSS>
fiunk	fiú/NOUN<POSS<PLUR><1>>
fiatok	fiú/NOUN<POSS<PLUR><2>>
fiuk	fiú/NOUN<POSS<PLUR>>

Megemlítjük, hogy a PLUR jegyre a hierarchia három különböző helyén is szükség van: közvetlenül a főkategória-jegy (itt NOUN) alatt (ekkor a lemmában megadott entitás többes számát jelzi), a POSS alatt (ekkor az entitást birtokló birtokos többes számát jelzi), és az ANP alatt (ekkor az entitás által birtokolt birtok többes számát jelzi) – a hierarchikus elrendezés azonban biztosítja, hogy ugyanannak a PLUR jegynek a használata nem vezet félreértéshez, hiszen ezek más jegyek alatt helyezkednek el, amit a linearizált kódban a zárójelezés mutat. Ezt mutatják az alábbi alakok, ahol a PLUR különböző pozíciókban külön-külön és egyszerre is megjelenhet (itt megjegyzendő, hogy a birtok többes számának jelzése a beszélt köznyelvben állítmányi helyzetben nem kötelező, sőt egyes beszélőknél tiltott: pl. *A könyvek a* %*fiúéi*/%*fiúé*).

(10) A PLUR jegy különböző használatai

birtokos és/vagy birtokjelölés		*mi többes számú?*
a. nincs birtokos- és birtokjelölés:		
fiúk	fiú/NOUN<PLUR>	(entitás)
b. csak birtokosjelölés (itt 3. személyű):		
fiai	fiú/NOUN<PLUR><POSS>	(entitás)
fiuk	fiú/NOUN<POSS<PLUR>>	(birtokos)
fiaik	fiú/NOUN<PLUR><POSS<PLUR>>	(entitás és birtokos)
c. csak birtokosjelölés:		
fiúké	fiú/NOUN<PLUR><ANP>	(entitás)
fiúéi	fiú/NOUN<ANP<PLUR>>	(birtok)
fiúkéi	fiú/NOUN<PLUR><ANP<PLUR>>	(entitás és birtok)
d. birtokos- és birtokjelölés is (csak azok, ahol a birtok többes számú):		
fiáéi	fiú/NOUN<POSS><ANP<PLUR>>	(birtok)
fiaiéi	fiú/NOUN<PLUR><POSS><ANP<PLUR>>	(entitás és birtok)
fiukéi	fiú/NOUN<POSS<PLUR>><ANP<PLUR>>	(birtokos és birtok)
fiaikéi	fiú/NOUN<PLUR><POSS<PLUR>><ANP<PLUR>>	(entitás, birtokos és birtok)

Itt kell kitérnünk a többes szám egy speciális használatára: a **familiáris többes** alak morfoszintaktikailag többes számú, de nem a lexémával kifejezett entitás többes számára, hanem az azzal valamilyen „familiáris" viszonyban levők összességére (család, ismerősök stb.) utal: pl. *sógorék, szomszédék* stb. Ez a viszony kombinálódhat a birtokosjelölős alakkal (pl. *sógorodék*) és a birtokjelölős alakkal (*sógoréké*). Ezért az annotáció egy a NOUN alatti PLUR általi dominált FAM jegy segítségével történik (11).

Az esetek kódolása is megfelel a morfológiai jelöltségnek: mivel a jelöletlen eset a nominativus, ezért az alanyesetű alakokat külön nem jelöljük, a többi 17 eset kódolására az elterjedt latin elnevezéseik három betűs rövidítéseit használ-

juk. A CAS jegy azt jelzi, hogy itt egy jelölt (azaz nem nominativusi) alakkal van dolgunk. A 18 eset annotációját és az esetek elnevezését l. (12)-ben.

(11) Familiáris többes alakok

fiúék	fiú/NOUN<PLUR<FAM>>	az entitás familiáris csoportja
fiáék	fiú/NOUN<PLUR<FAM>><POSS>	a birtokolt entitás fam. csoportja
fiúéké	fiú/NOUN<PLUR<FAM>><ANP>	az entitás fam. csoportjának birtoka
fiáéké	fiú/NOUN<PLUR<FAM>><POSS><ANP>	a birtokolt entitás fam. csoportjának birtoka

(12) Az esetek annotációja

a. „strukturális esetek”

fiú	fiú/NOUN	nominativus
fiút	fiú/NOUN<CAS<ACC>>	accusativus
fiúnak	fiú/NOUN<CAS<DAT>>	dativus

b. „lexikális esetek”

b1. „helyhatározói”

i. forrás

fiúról	fiú/NOUN<CAS<DEL>>	delativus
fiúból	fiú/NOUN<CAS<ELA>>	elativus
fiútól	fiú/NOUN<CAS<ABL>>	ablativus

ii. hely

fiún	fiú/NOUN<CAS<SUE>>	superessivus
fiúban	fiú/NOUN<CAS<INE>>	inessivus
fiúnál	fiú/NOUN<CAS<ADE>>	adessivus

iii. cél

fiúra	fiú/NOUN<CAS<SBL>>	sublativus
fiúba	fiú/NOUN<CAS<ILL>>	illativus
fiúhoz	fiú/NOUN<CAS<ALL>>	allativus
fiúig	fiú/NOUN<CAS<TER>>	terminativus

b2. egyéb

fiúval	fiú/NOUN<CAS<INS>>	instrumentalis-comitativus
fiúért	fiú/NOUN<CAS<CAU>>	causalis-finalis
fiúként	fiú/NOUN<CAS<FOR>>	formativus
fiúvá	fiú/NOUN<CAS<TRA>>	translativus-factivus
húsvétkor	húsvét/NOUN<CAS<TEM>>	temporalis

3.4. Bináris és unáris jegyek: igék

Az igék specifikációjában az eddigi elveknek megfelelően a hierarchikus elrendezést a bináris jegyekkel kombináljuk. Az említett *adhattátok* és *ad* igalakok kétértékű jegyekkel való annotációja a következő.

(13) Két igealak bináris jegy–érték struktúrája

a. *adhattátok*

ad +
IGE +
MODÁLIS +
NEM JELEN.KIJ +
MÚLT IDŐ +
FELTÉTELES MÓD −
KÖTŐ-FELSZ. MÓD −
INFINITÍVUSZ −
TÖBBES SZÁM +
NEM 3. SZEMÉLY +
1. SZEMÉLY −
(TÁRGY 2. SZEMÉLYŰ X)
2. SZEMÉLY +
DEFINITSÉG +

b. *ad*

ad +
IGE +
MODÁLIS −
NEM JELEN.KIJ −
(MÚLT IDŐ −)
(FELTÉTELES MÓD −)
(KÖTŐ-FELSZ. MÓD −)
(INFINITÍVUSZ −)
TÖBBES SZÁM −
NEM 3. SZEMÉLY +
(1. SZEMÉLY −)
(TÁRGY 2. SZEMÉLYŰ X)
(2. SZEMÉLY −)
DEFINITSÉG −

(14) Elemzés unáris jegyekkel (a redundáns információk nélkül):

a. hierarchikus formában:

adhattátok

ad
IGE (VERB)
MODÁLIS (MODAL)
NEM JELEN.KIJ (--)
MÚLT IDŐ (PAST)
TÖBBES SZÁM (PLUR)
NEM 3. SZEMÉLY (PERS)
2. SZEMÉLY (2)
DEFINITSÉG (DEF)

ad

ad
IGE (VERB)

b. linearizált formában:

```
ad/VERB<MODAL><PAST><PLUR><PERS<2>><DEF>
ad/VERB
```

Az igék idő/módjának annotációja úgy történik, hogy közvetlenül a VERB jegy alatt szerepel az erre vonatkozó információ (azaz a **NEM JELEN.KIJELENTŐ MÓDÚ** jegy a linearizált formából hiányzik). A jegyrendszer felépítése biztosítja, hogy a zérusmorfémát tartalmazó jelöletlen jelen idő kijelentő módú alakok nem kapnak külön jelölést.

(15) Az igék négy idő/módjának annotációja

ad	ad/VERB	jelen idő kijelentő mód
adott	ad/VERB<PAST>	múlt idő kijelentő mód
adna	ad/VERB<COND>	jelen idő feltételes mód
adjon	ad/VERB<SUBJUNC-IMP>	kötő-felszólító mód

Az igei személyjelölés annotációja a főnévi birtokos mintát követi: a jelöletlen 3. személy jegy nélkül áll, a 1. és 2. személyek jegyei a PERS jegy alatt szerepelnek. A speciális, csak E.1. személyű igéknél megfigyelhető, 2. személyű tárgyra utaló *-lak/lek* toldalékos alakok annotációja egy az <1> jegy alá bevezetett <OBJ2> jeggyel történik. A definit–indefinit (határozott–általános) igealakok megkülönböztetésére a DEF jegy szolgál, hiszen a definit alakok a morfológiailag jelöltek. A számjelölés a független <PLUR> jeggyel történik.

(16) Igei indefinit és definit szám/személyjelölés annotációja

adok	ad/VERB<PERS<1>>	*adom*	ad/VERB<PERS<1>><DEF>
adlak	ad/VERB<PERS<1<OBJ2>>>		
adsz	ad/VERB<PERS<2>>	*adod*	ad/VERB<PERS<2>><DEF>
ad	ad/VERB	*adja*	ad/VERB<DEF>
adunk	ad/VERB<PLUR><PERS<1>>	*adjuk*	ad/VERB<PLUR><PERS<1>><DEF>
adtok	ad/VERB<PLUR><PERS<2>>	*adjátok*	ad/VERB<PLUR><PERS<2>><DEF>
adnak	ad/VERB<PLUR>	*adják*	ad/VERB<PLUR><DEF>

Az infinitívusz szám/személy jelölésének annotációja igen hasonló az igékéhez. Az „infinitivus" igei jegy (VERB<INF>) sajátos jegykombinációkat enged csak meg: az infinitívusznak nincsen idő/módja és definitsége, viszont lehet szám/személye, amit a <PERS> jeggyel fejezünk ki (az %*adhatni* és az ?*adnalak* típusú infinitívuszi alakok is csak periferiálisan léteznek). Az egyetlen jelentős eltérés az igék annotációjához képest, hogy a <PERS> jegy hiánya ebben az esetben nem a 3. személyű alakot (pl. *adnia*), hanem a szám/személyjelölés nélküli alakot (pl. *adni*) kódolja.[2]

(17) A szám/személyjelölővel nem rendelkező és az azzal rendelkező infinitívusz annotációja

adni	ad/VERB<INF>		
adnia	ad/VERB<INF><PERS>	*adniuk*	ad/VERB<INF><PLUR><PERS>
adnom	ad/VERB<INF><PERS<1>>	*adnunk*	ad/VERB<INF><PLUR><PERS<1>>
adnod	ad/VERB<INF><PERS<2>>	*adnotok*	ad/VERB<INF><PLUR><PERS<2>>

Összefoglalva: az unáris jegyekkel való hierarchikus ábrázolás lehetőséget teremt arra, hogy egyszerűen megfogalmazható és nyelvészetileg alátámasztott morfoszintaktikai jegyek segítségével olyan annotációt adjunk, amely teljes, és általában véve tükrözi a morfológiai jelöltségi viszonyokat. Azaz anélkül, hogy közvetlenül hivatkoznunk kellene az elemzett szóalak formai tulajdonságaira (allo-

[2] Ez fontos különbség, mert az infinitívuszt vonzó igék közül azok, amelyek szám/személyjelöléssel rendelkeznek, kizárólag a szám/személyjelölés nélküli infinitívuszt engedik meg: pl. *Dolgozni(*a) akar*.

morfok, szegmentálás stb.), az annotációs kód mégis változó hosszúságú: hossza nagyjából megfelel a szóalak morfológiai komplexitásának. Ez azt is jelenti, hogy zérusmorfémák esetén az annotáció – mivel bináris jegyeik mind negatívak – kizárólag a lexémából és a főkategória címkéjéből áll. Minden további morféma tovább növeli az annotáció bonyolultságát. Az alábbi táblázatban néhány ilyen „monotonon bővülő" komplexitású alaksort adtunk meg az alakok annotációjával együtt a zérustoldaléktól a maximális alakokig (az összehasonlítás kedvéért a hozzávetőleges morfémahatárokat a szóalakokban jelöltük).

(18) Szóalakok és annotációik egy-egy monoton növő komplexitású sora

fiú	`fiú/NOUN`
fiú-k	`fiú/NOUN<PLUR>`
fi-a-i	`fiú/NOUN<PLUR><POSS>`
fi-a-i-d	`fiú/NOUN<PLUR><POSS<2>>`
fi-a-i-tok	`fiú/NOUN<PLUR><POSS<PLUR><2>>`
fi-a-i-tok-é	`fiú/NOUN<PLUR><POSS<PLUR><2>><ANP>`
fi-a-i-tok-é-i	`fiú/NOUN<PLUR><POSS<PLUR><2>><ANP<PLUR>>`
fi-a-i-tok-é-i-t	`fiú/NOUN<PLUR><POSS<PLUR><2>><ANP<PLUR>><CAS<ACC>>`
ad	`ad/VERB`
ad-hat	`ad/VERB<MODAL>`
ad-hat-ott	`ad/VERB<MODAL><PAST>`
ad-hat-t-ak	`ad/VERB<MODAL><PAST><PLUR>`
ad-hat-t-atok	`ad/VERB<MODAL><PAST><PLUR><PERS<2>>`
ad-hat-t-á-tok	`ad/VERB<MODAL><PAST><PLUR><PERS<2>><DEF>`

Fontos rámutatni, hogy az annotáció semmilyen értelemben nem használja az alulspecifikációt (unáris jegyek esetén ez nem is lehetséges), azaz nem lehetséges megadni úgy egy morfoszintaktikai leírást, hogy az valamilyen értékre ne legyen meghatározva – ez bináris vagy többértékű jegyeket alkalmazó rendszerekben egyszerűen a szóban forgó jegy értékének kitöltetlenül hagyásával történhet. Mivel minden annotáció a morfofonológiai értékekre nézve teljesen specifikált, ezért a potenciálisan alulspecifikáltként kezelhető eseteket kétértelműségként kell kezelnünk. Ilyen eset a magyarban meglehetősen ritka. Például az E.1 és E.2 birtokos alakok esetjelölés nélkül jelenthetnek nominativus vagy accusativust; vagy egyes igealakok a definitség mindkét értékét felvehetik. Néhány példa:

(19) Morfológiai kétértelműségek kezelése alulspecifikáció nélkül

fiam	`fiú/NOUN<POSS<1>>`	nominativus (pl. *A fiam látott engem.*)
	`fiú/NOUN<POSS<1>><CAS<ACC>>`	accusativus (pl. *Láttam a fiam.*)
adtam	`ad/VERB<PAST><PERS<1>>`	indefinit (pl. *Egy almát adtam neki.*)
	`ad/VERB<PAST><PERS<1>><DEF>`	definit (pl. *Az almát adtam neki.*)

4. Formalizmus

Az itt következő részben pontosítjuk az inflexiós jegyrendszernek azt a formalizmusát, amelyet az előző részben mutattunk be. Formálisan az inflexiós annotáció két komponensből áll, az egyik komponens a jegy–érték struktúra, amelyben a **bináris morfoszintatikai jegyek** és ezek pozitív vagy negatív **értékei** szerepelnek. A másik komponens a hierarchiáért felelős, ezt a legegyszerűbb egy **irányított körmentes gráfként** (azaz irányított faként) meghatározni, amelyben minden csomóponthoz egy bináris jegy–érték pár van rendelve, az irányított élek pedig megfelelnek a jegy–érték párok közötti dominanciaviszonyoknak. Mivel ez a gráf egy fa, ezért összefüggő és egy csomópont (a gyökércsomópont) kivételével minden csomóponthoz van olyan csomópont, amelyik őt közvetlenül dominálja; a körmentesség pedig azt biztosítja, hogy ne lehessen egy csomópontnak több közvetlenül domináns csomópontja. A jegy–érték párokkal címkézett gráfra egy további feltételnek kell teljesülnie: csak a **pozitív** értékkel rendelkező jegy–érték párok csomópontjai **dominálnak** más csomópontokat (azaz a negatív értékkel címkézett csomópontok a fában levelek lesznek). Ez a feltétel az előző részben elmondottak alapján lehetővé teszi, hogy a bináris jegyes hierarchikus szerkezet unáris jegyessé alakítható legyen a hierarchia megtartásával, és így a(z unáris) jegyek száma tükrözze a morfológiai jelöltséget.[3]

Ahogyan az előző részben láttuk, a **gyökércsomópont** tartalmazza az inflektált szóalak **kategóriáját** (szófaját, POS *(part-of-speech)*-címkéjét): a gyökércsomópont egy olyan jegy–érték párral van címkézve, ahol a jegy valamely főkategória-jegy (az előző részben ezek közül a `NOUN` és a `VERB` szerepelt).[4] Minden inflektálható kategóriához tartozik egy rögzített inflexiós jegy–érték struktúra, azaz bináris jegyértékekkel címkézett csomópontú fagráf. Inflektálható kategória azonban csak öt van: a három névszói és a ragozható determinánsi és az egy igei kategória, ezek jegy–érték szerkezeteiről l. az előző, illetve a következő részt.

Az inflexiós annotáció linearizálása úgy történik, hogy a pozitív értékkel bíró jegyeket írjuk le a megfelelő zárójelezéssel. Mivel egy fában az ugyanazon csomópont által dominált csomópontok (az ún. testvércsomópontok) egymás közötti sorrendje lényegtelen, ilyen esetekben a linearizálás az összes sorrendben lehetséges. Praktikus okokból azonban a jegyek sorrendjét úgy rögzítettük, hogy

[3] Valójában az annotációt közvetlenül unáris jegyekkel címkézett fagráffal is definiálhatnánk, ekkor egy annotáció ennek a jegyekkel címkézett fának olyan részfája lenne, amelynek a gyökércsomópontja megegyezik a bővebb fáéval.

[4] A hunmorph annotációs rendszer aktuális változata által használt főkategória-jegyek listája megtalálható a függelék (A1) ábrájában.

a félreolvasás lehetősége a lehető legkisebb legyen (az inflektálható kategóriákhoz tartozó jegyek kimerítő listáját és sorrendjüket l. a következő részben). Így a linearizált annotáció már egyértelmű, kódokból és zárójelekből álló sztring lesz.

Mivel a linearizált kód – jelöletlen szóalak esetén – egyetlen főkategória-jegyből is állhat, ezért fontos megjegyeznünk, hogy elvi különbség van egy főkategória-jegy és az ilyen „rövid" inflexiós annotáció között. Például a NOUN és a <NOUN> két különböző dologra utal: az első egy **jegy neve**, amely a gyökércsomópontban állhat; a második egy morfoszintaktikailag **teljesen specifikált** alak, azaz jegyekkel címkézett fagráf, amelynek minden főnévi jegye negatív (azaz esetünkben az egyes számú nem-birtokos nem-birtok nominativusi alak, l. (7b), illetve unáris formában (8); hasonlóan igékre, l. (13b), illetve (14)). Ezt a különbséget a végső linearizált kódban azonban nem használjuk: a főkategória (és így az egész morfoszintaktikai jegyrendszer) praktikus okokból mindig külső zárójelek nélkül szerepel – ez nem vezethet félreértéshez, hiszen az annotáció úgyis mindig teljes elemzést ad vissza. Az elemzés általános formája – amely már az előző részből ismerős – a következő:

(20) Az inflexiós annotáció sémája

szóalak
`lemma/FŐKATEGÓRIA<INFL_JEGY_1><INFL_JEGY_2>...<INFL_JEGY_n>`

Morfológiai elemzésnek általánosan egy olyan hozzárendelést nevezünk, amely minden egyes jólformált szóalakhoz (sztringhez) hozzárendel egy lexéma–annotáció párt. Ez a hozzárendelés azonban nem egyértelmű (nem függvény), mivel ugyanahhoz a szóalakhoz több különböző elemzést is rendelhet morfológiai homonímia esetén – l. pl. a (19)-beli eseteket. A hozzárendelés megfordítása (inverze) sem függvény, mert ugyanolyan lexémának ugyanolyan annotációval különböző szóalakok felelhetnek meg: ez a helyzet áll elő morfofonológiai ingadozás esetén (pl. *fotelban – fotelben* vagy *fürdenek – fürödnek*), vagy olyan alakoknál, ahol a szuppletív tő megjelenése nem kötelező (pl. *jöjj – gyere* vagy *volna – lenne*).

(21) Ingadozó alakok azonos annotációt kapnak

fotelban	`fotel/NOUN<CAS<INE>>`
fotelben	`fotel/NOUN<CAS<INE>>`
fürdenek	`fürdik/VERB<PLUR>`
fürödnek	`fürdik/VERB<PLUR>`
gyere	`jön/VERB<SUBJUNC-IMP><PERS>`
jöjj	`jön/VERB<SUBJUNC-IMP><PERS>`
jöjjél	`jön/VERB<SUBJUNC-IMP><PERS>`

A következőkben a korábbi főnévi és igei elemzéseket kiegészítjük a többi inflektálható elem annotációjával.

4.1. Névszói kategóriák

Régi problémája a leíró nyelvtanoknak, hogy be lehet-e (és ha igen, hogyan) sorolni egyértelműen a névszói alakokat valamelyik névszói kategóriába (l. többek között Moravcsik 1997). A melléknevek és a számnevek a főnevekkel átfedő osztályokat alkotnak, és nehéz egyértelmű disztribúciós teszteket adni, amelyeknek alapján ezek a kategóriák egyértelműen megkülönböztethetőek lennének. Ezen a helyzeten a morfológiai vizsgálatok sem segítenek, mivel mind a melléknevek, mind a számnevek felvehetik az összes főnévi inflexiót egyes „elliptikus" és „nominalizáló" kontextusokban: pl. *Nem szeretem a kívánsi**akat**; Ez az én nagy labdám, az meg a te kicsi**d**; Bátr**aké** a szerencse; Összeültek a nyolc**ak**; Négy**et** rendeltem; Az ő öt könyve meg az én hárm**am***. Itt és a többi hasonló példában vitatható, hogy az adott melléknév vagy számnév a saját „prototipikus" mondattani **funkciójában** szerepel-e, de melléknév, illetve számnév voltuk mellett számos érv szól. Nyilvánvaló, hogy a mondatokban különböző funkciókban álló ugyanazon elemek megkülönböztetése nem lehet a feladata egy csak szóalakokat vizsgáló morfológiai elemzőnek, és így az annotációnak sem. Így például a *pék barátom* és a *szomszéd Józsi* típusú szerkezekben az első főnév módosító szerepű (ahogyan tipikusan a melléknevek), a *szépek imádata* és a *kevés is sok* típusú szerkezetekben a melléknév, illetve a számnév főnévi jellegű (birtokos szerkezeten belül, illetve alanyként áll); ezt a tényt azonban nem érdemes az adott alakok többszófajúsága mellett felhozni, mert akkor a névszók jelentős többségével ezt kellene tennünk, és így értelmetlenül sok többszörös annotációt kapnánk. (A kizárólag melléknévinek tartott toldalékok, mint amilyen a közép- és felsőfok jele, sem adnak jobb fogódzót, ezek ugyanis a mellékneveken kívül egyes számnevekkel is lehetségesek (pl. *több, kevesebb, legelső*), és egyes konstrukciókban főnevekhez is járulhatnak: pl. *székebb a széknél*.)

A hunmorph kategóriarendszerének összeállításánál arra is figyelemmel kellett lennünk, hogy az elérhető elektronikus adatbázisok (pl. szótárak) és a rendelkezésre álló elemzett korpuszok (pl. a Szeged Korpusz, l. Csendes et al. 2004) valamilyen módon mégis megkülönböztetik a három fő névszói kategóriát (ezt nagyon sokszor nem formai–disztribúciós, hanem szemantikai–funkcionális alapokon teszik). Ezért az információvesztés elkerülése végett érdemes ezt a kategorizációt megtartani. A három névszói kategória morfoszintaktikai jegyrendszere

viszont azonos lesz: bármely névszó felveheti az összes főnévi inflexiós kategóriát. Az alábbi (22) néhány példát ad inflektált alakokra.

(22) Melléknévi és számnévi alakok névszói inflexiókkal

kíváncsi	`kíváncsi/ADJ`
kíváncsijaitokét	`kívácsi/ADJ<PLUR><POSS<PLUR><2>><ANP><CAS<ACC>>`
kétezer	`kétezer/NUM`
kétezreinkével	`kétezer/NUM<PLUR><POSS<PLUR><1>><ANP><CAS<INS>>`

4.2. Determinánsok

A negyedik inflektálható kategória a determinánsoké (`DET`), l. (23). Pontosabban a determinánsoknak csak egy része inflektálható, az olyan szerkezetekben, mint pl. *ez**eké** a lányoké, ab**ban** a házban*. Más részük viszont nem inflektálható, pl. *e lányoké, ama házban, azon gondolatoknak*. Az inflektálható determinánsok inflexiós jegyszerkezetükben megegyeznek a többi névszóval. (Meg kell jegyeznünk, hogy a szokásosan a determinánsok közé számított névelők a hunmorph-ban külön kategóriát képeznek (`ART`), amit rendkívül gyakori előfordulásuk és speciális funkciójuk indokol – ide csupán három lemma tartozik: *a, az, egy*.) Néhány példa determinánsokra (az utolsóként felsorolt típus – *ezen, azon* stb. – kétértelmű: lehet inflektálhatatlan determináns, de lehet superessivusi esetű inflektálható is: vö. *azon emberekkel* vs. *azon az emberen*):

(23) Inflektált és nem inflektált determinánsok

emez	`emez/DET`
ugyanazokéval	`ugyanaz/DET<PLUR><ANP><CAS<INS>>`
e	`e/DET`
azon	`azon/DET`
	`az/DET<CAS<SUE>>`

4.3. Névmások

4.3.1. Főnévi, melléknévi, számnévi névmások

A névmások a hunmorph rendszerben nem képeznek külön kategóriát (szemben a más alapokon nyugvó annotációkkal, pl. a már említett MSD-kódrendszerrel). A disztribúciós elemzés (és funkcionális meggondolások is) azt az elképzelést támogatják, hogy a névmások szétoszthatók a négy névszói (`NOUN`, `ADJ`, `NUM`, `DET`) és a határozószói (`ADV`) kategóriák között. Hely hiányában a névmások elemzésére itt részletesen nem tudunk kitérni, álljon itt néhány példa a hagyományos besorolásuk szerint:

(24) Főnévi, melléknévi és számnévi névmások annotációja

a. mutató

ez	ez/NOUN
azokéval	az/NOUN<PLUR><ANP><CAS<INS>>
ilyen	ilyen/ADJ
olyanjainak	olyan/ADJ<PLUR><POSS><CAS<DAT>>
ennyi	ennyi/NUM
annyinkat	annyi/NUM<POSS<PLUR><2>><CAS<ACC>>

b. kérdő

micsoda	micsoda/NOUN
kikét	ki/NOUN<PLUR><ANP><CAS<ACC>>
melyik	melyik/ADJ
milyeneken	milyen/ADJ<PLUR><CAS<SUE>>
hány	hány/NUM
mennyivel	mennyi/NUM<CAS<INS>>

c. egyéb (vonatkozó, általános, tagadó)

amely	amely/NOUN
valakijeitekét	valaki/NOUN<PLUR><ANP<PLUR>><CAS<ACC>>
bármelyik	bármelyik/ADJ
semmilyenekkel	semmilyen/ADJ<PLUR><CAS<INS>>
mindahány	mindahány/NUM
akármennyiért	akármennyi/NUM<CAS<CAU>>

4.3.2. Személyes névmások

A hagyományosan személyes és birtokos névmásoknak nevezett szóosztály annotálása érdekében az eddig bemutatott névszói annotációs jegyrendszert kismértékben ki kell bővítenünk. A személyes névmások annotációs rendszerünk szerint speciális főnevek, amelyeknek névszói inflexiós jegyeik lehetnek (alakjuk nagyon gyakran szuppletív, pl. *engem, bennünket, velük, rá*). A különböző személyű személyes névmásokkal való egyeztetési jelenségek indokolják, hogy a névszói jegyrendszert kiegészítsük az igéknél ismert és a személyre utaló PERS jeggyel. Ez a PERS jegy az infinitívuszoknál látott módon jelöli a személyt (l. (17)): magában állva a <PERS> 3. személyre utal, míg az e jegy által dominált személyjegyekkel az 1., illetve 2. személyre. Ekkor a személyes névmások annotációja a következő (a formális – „önöző", illetve „magázó" – személyes névmásokat is szerepeltetjük, ezek morfoszintaktikailag 3. személyűek).

(25) A személyes névmások annotációja

én	én/NOUN<PERS<1>>	*mi*	mi/NOUN<PLUR><PERS<1>>
te	te/NOUN<PERS<2>>	*ti*	ti/NOUN<PLUR><PERS<2>>
ő	ő/NOUN<PERS>	*ők*	ők/NOUN<PLUR><PERS>
ön	ön/NOUN<PERS>	*önök*	önök/NOUN<PLUR><PERS>
maga	maga/NOUN<PERS>	*maguk*	maguk/NOUN<PLUR><PERS>

A személyes névmások esetekkel ellátott alakjai között több morfofonológiailag kivételes, illetve szuppletív alak van (l. pl. (3)), ezen kívül a legtöbb alak a nemformális személyes névmásoknál hiányzik (TRA: **énné*, FOR: **teként*, TER: **őig*, TEM: **önkor*), illetve többszörös alakváltozatok is előfordulnak; néhány példa:

(26) Inflektált személyes névmások

engem engemet	`én/NOUN<PERS<1>><CAS<ACC>>`
neked néked	`te/NOUN<PERS<2>><CAS<DAT>>`
vele véle	`ő/NOUN<PERS><CAS<INS>>`
önhöz	`ön/NOUN<PERS><CAS<ADE>>`
magáig	`maga/NOUN<PERS><CAS<TER>>`
bennünket minket	`mi/NOUN<PLUR><PERS<1>><CAS<ACC>>`
belőletek	`ti/NOUN<PLUR><PERS<2>><CAS<ELA>>`
rajtuk	`ők/NOUN<PLUR><PERS><CAS<SUE>>`
önökké	`önök/NOUN<PLUR><PERS><CAS<TRA>>`
magukként	`maguk/NOUN<PLUR><PERS><CAS<FOR>>`

4.3.3. Birtokos névmások

Az ún. „birtokos" névmások nem birtokosjelölővel, hanem **birtok**jelölővel vannak ellátva, hiszen nem a személyes névmás által kifejezett személy birtokosát, hanem annak birtokát jelölik, és szintaktikai disztribúciójuk is ennek felel meg: *A könyv a fiúé/tied/övé*. Ezért ezek annotációja az ANP jeggyel történik:[5]

(27) A birtokosra utaló névmások annotációja

enyém	`én/NOUN<PERS<1>><ANP>`
tiéd tied tiedé	`te/NOUN<PERS<2>><ANP>`
övé	`ő/NOUN<PERS><ANP>`
öné	`ön/NOUN<PERS><ANP>`
magáé	`maga/NOUN<PERS><ANP>`
miénk mienk mienké	`mi/NOUN<PLUR><PERS<1>><ANP>`
tiétek tietek tieteké	`ti/NOUN<PLUR><PERS<1>><ANP>`
övék övéké	`ők/NOUN<PLUR><PERS><ANP>`
önöké	`önök/NOUN<PLUR><PERS><ANP>`
maguké	`maguk/NOUN<PLUR><PERS><ANP>`

A birtokos névmások viselkedése jól példázza azt, hogy a birtokjelölés bizonyos erősen korlátozott esetekben és módon egy alakon belül megismételhető. A többszörös birtokviszonyok lerövidítésére szolgálnak az olyan szerkezetek, mint *Az én kutyám pórája – Kinek a pórája? – Az én kutyámé/Az* ***enyémé***$^{?\#}$ *Az enyém*.

[5] Megjegyzendő, hogy a POSS jegy főnévi személyes névmásokra nem is használatos, hiszen az ezt kifejező alakok szisztematikusan hiányoznak: **éned, *őm, *önötök, *magánk*.

Ez utóbbi alak nem egyszerű birtokoltságot, hanem a birtok általi újabb birtokoltságát fejez ki, és bizonyos mértékig itt is kifejezhető mindkét birtok többes száma: *A kutyáim pórája – az enyémeké; A kutyám pórázai – ?az enyéméi; A kutyáim pórázai – az ?enyémekéi.*[6] Ezt a szerkezetet az ANP csomópont alatti újabb ANP csomóponttal lehet kódolni.

(28) Többszörös birtokjelölés

enyémé	én/NOUN<PERS<1>><ANP<ANP>>
tieidé	te/NOUN<PERS<2>><ANP<PLUR><ANP>>
miénkéi	mi/NOUN<PLUR><PERS<1>><ANP<ANP<PLUR>>>
tieitekéi	ti/NOUN<PLUR><PERS<2>><ANP<PLUR><ANP<PLUR>>>

A birtokos névmások szintén felvehetnek további főnévi inflexiós jegyeket (a POSS jegy kivételével). Így az előbbiekkel és az esetjelöléssel megkaphatjuk a maximálisan komplex PERS jegyet tartalmazó főnévi alakot (az alábbi listában az összehasonlítás kedvéért a szóalakokban bejelöltük a morfémák hozzávetőleges határait):

(29) Inflektált birtokos névmások növekvő komplexitású sora

(*ő*	ő/NOUN<PERS>)
öv-é	ő/NOUN<PERS><ANP>
eny-é-m	én/NOUN<PERS<1>><ANP>
ti-e-i-d	te/NOUN<PERS<2>><ANP<PLUR>>
mi-e-i-nk	mi/NOUN<PLUR><PERS<1>><ANP<PLUR>>
ti-e-i-tek-et	ti/NOUN<PLUR><PERS<2>><ANP<PLUR>><CAS<ACC>>

4.4. Névutók

A névutók külön főkategória (POSTP), de bizonyos névutós alakok érintik a főnévi annotációt is. Az ún. „személyragozott névutók" (*utánam, eléd, nélküle* stb.) olyan elemek, amelyeknek a formája névutó + szám/személyjelölő, de szintaktikai viselkedésük a főnév+névutó sémát követi (pl. *A fiú nélkül/Nélkülem jött el*). Ez problematikus, mivel a morfológiai elemző csak egyes szavakat képes annotálni, a szavak között fennálló konstrukciók elemzése közvetlenül nem feladata. Vegyük azonban észre, hogy a főnév+névutó szerkezetek szintaktikailag az esetjelölős főnévi szerkezetekkel rokoníthatók (vö. az előző mondatokat a következőkkel: *A fiúval/Velem jött el*). Így a személyragozott névutókat tekinthetjük

[6] Ugyanez a szerkezet nem-névmási alakokkal csak nagyon nehezen fogadható el: ?**Páléé*, talán kicsit grammatikusabbá válik, ha az első birtokjelölő után többesszámjelölés van: ??*Páléié*.

speciális névutós személyes névmásoknak, és mivel a névmások PERS jeggyel ellátott főnevek, ezért érdemes bevezetni a főnévi inflexiós rendszerbe a POSTP jegyet (amely a POSTP főkategória-jegytől különbözik). A főnév alatti POSTP jegynek aljegye lesz az összes névutós jegy, amelyeknek a nevei a névutó lemmájával azonosak. Ekkor a személyragozott névutók annotációja az esetjelölős személyes névmásokéval rokonítható. Figyeljük meg, hogy a személyrag nélküli névutók és a személyraggal ellátott névutók más-más főkategóriához tartoznak: az első névutó (POSTP), a második főnév (NOUN), ami természetes, hiszen szintaktikai disztribúciójuk nagyban eltér: a POSTP kategóriát közvetlenül megelőzi egy főnév (pl. *fiú mögött*), míg a névutós NOUN kategória határozóként vagy vonzatként áll a mondatban (pl. *mögöttünk*). Megfigyelhetjük azt is, hogy a 3. személyű személyragozott névutók lemmája sohasem lehet a formális önözés/magázás *ön,* illetve *maga* lemmája: *előtte* értelmezése 'őelőtte', és nem 'ön/maga előtt', és hasonlóan a T.3 *közéjük* 'őközéjük' és nem 'önök/maguk közé'.

(30) Személyragozott és sima névutók annotációja

nélkül	nélkül/POSTP
nélküled	te/NOUN<POSTP<NÉLKÜL>><PERS<2>>
közül	közül/POSTP
közülük	ők/NOUN<POSTP<KÖZÜL>><PLUR><PERS>
mellé	mellé/POSTP ő/NOUN<POSTP<MELLÉ>><PERS>
melléje	ő/NOUN<POSTP<MELLÉ>><PERS>
számára	számára/POSTP ő/NOUN<POSTP<SZÁMÁRA>><PERS>
számunkra	mi/NOUN<POSTP<SZÁMÁRA>><PLUR><PERS<1>>

Figyeljük meg, hogy egyes névutók kétértelműek: egyrészt vannak olyanok, amelyeknek E.3 alakja megegyezhet az inflektálatlan alakkal (pl. *mellé, mögé*); másrészt azok, amelyek birtokjelölős alakúak, az E.3 személyben kétértelműek, pl. a *számára* funkciója kettős: egyrészt névutó (pl. *Pál számára*), másrészt az E.3 névutós névmási alak: 'az ő számára' (pl. *Számára nincs kegyelem*). Bizonyos esetekben a mutató névmás és a névutó egybeírandó (pl. *anélkül, ezelőtt, amiatt*). Ekkor az annotáció szintén a főnévi kategóriát használja a POSTP jeggyel.

(31) Mutató névmási névutók annotációja

anélkül	az/NOUN<POSTP<NÉLKÜL>>
ezelőtt	ez/NOUN<POSTP<ELŐTT>>

A névszói alakok teljes inflexiós specifikációja a függelék (A2) táblázatában található.

4.5. Igék

Az igei alakok jelentős része morfológiailag defektív, azaz egyes lemmák bizonyos – egyébként megengedett – jegykombinációkkal nem állnak. Ilyenek többek között a személytelen igék, amelyeknek minden idő/módjuk megvan, de csak E.3 indefinit alakban állhatnak: ilyen például a *hajnalodik* ige vagy a *kell, lehet* segédigék.A *rejlik, történik* stb. igéknek vannak többes számú alakjaik is, de csak 3. személyben (és indefinitként) állnak. Az intranzitív igéknek nem vagy csak periferikusan van definit alakjuk (pl. *kimosakodja magát, lejárja a távot* stb.), de van néhány ige, amelyeknek egyáltalán nincs (pl. *jön, megy, van*). Van olyan ige, amelynek formailag csak múlt ideje van, más idő/módban, infinitívuszban nem állhat, ilyen a *szokott* segédige, vagy pedig csak kijelentő mód 3. személyben áll: pl. *nincs, nincsenek.* A *fog* segédigének viszont nincs sem kifejezett idő/módja, sem infinitívusza. A legtöbb ige azonban minden idő/módban, szám/személyben és definit, illetve indefinit alakban is állhat. Az igék teljes annotációs specifikációja a függelék (A3) ábrájában látható.

5. Képzés és szóösszetétel

5.1. A képzés annotációja

Bizonyos feladatok (tartalomelemzés, gépi fordítás) megoldásához szükség lehet az inflexiós réteg kielemzésén túli morfológiai elemzésre is. Tipikus példa erre a szintaktikai elemzés *(parsing),* hiszen a melléknévi igeneves szerkezetek – pl. a *Koreában terjedő bankók* főnévi csoport – helyes elemzése megköveteli, hogy a *terjedő* ne csak mint melléknév kerüljön elemzésre (hisz ezt az elemző az őt követő főnévhez csatolná) hanem ki tudjuk elemezni a *terjed* igetövet is, hogy a *Koreában* alakot mint ennek bővítményét tudjuk elemezni. Képzőnek azt a toldalékot tekintjük, amely főkategóriához egy másik főkategóriát rendel. Ennek az elvnek a következménye, hogy a képzés bemenete nem érinthet inflektált alakot, morfotaktikai megfogalmazásban: inflektált alakok nem képezhetők tovább. Képzett alakoknak viszont kötelező inflexiót felvenniük (ha a kimeneti kategória inflektálható). Ezért a képzés annotációját formálisan olyan irányított gráfként tudjuk megadni, amelynek csomópontjai a főkategóriák: egy **képző** két (nem feltétlenül különböző) **kategóriacímke közötti irányított él**. Az elemzésben a képzés alapjául szolgáló lemma és a lemma kategóriája mellett meg kell adnunk a képző elnevezését és a kimeneti kategóriát, amely a következő módon történik.

(32) Az elemzés formalizmusa képzett alak esetén

szóalak
`lemma/LEMMA_KATEGÓRIA[KÉPZŐ]/VÉGSŐ_KATEGÓRIA<INFL_ANNOTÁCIÓ>`

Az alábbiakban néhány példát adunk meg különböző kategóriákból történő képzések elemzésére.

(33) Példák inflektált képzésekre

terjedő	`terjed/VERB[IMPERF_PART]/ADJ`	(foly. mn-i igenév)
csináltatna	`csinál/VERB[CAUS]/VERB<COND>`	(kauzatív)
székestül	`szék/NOUN[COM]/ADV`	(társ)
lábúak	`láb/NOUN[INAL_ATTRIB]/ADJ<PLUR>`	(elidegeníthetetlen tul.)
oroszul	`orosz/ADJ[MANNER]/ADV`	(mód)
okosabbjai	`okos/ADJ[COMPAR]/ADJ<PLUR><PERS>`	(középfok)
sokszor	`sok/NUM[MULTIPL-ITER]/ADV`	(multiplikatív–iteratív)
nyolcadikak	`nyolc/NUM[ORD]/NUM<PLUR>`	(sorszám)
szembeni	`szemben/POSTP[ATTRIB]/ADJ`	(tulajdonság)
utániról	`után/POSTP[ATTRIB]/ADJ<CAS<DEL>>`	

Többszörös képzés esetén a teljes képzési gráf kerül linearizálásra az előbbihez hasonló módon, az alábbiakban az általános sémát követően példákat adunk meg.

(34) Többszörös képzést annotáló elemzés

szóalak
`lemma/KATEGÓRIA_1[KÉPZŐ_1]/...KATEGÓRIA_n[KÉPZŐ_n]/`
`VÉGSŐ_KATEGÓRIA<INFL_ANNOTÁCIÓ>`

(35) Példák többszörös képzésekre

faxolgatás	`fax/NOUN[ACT]/VERB[FREQ]/VERB[GERUND]/NOUN`
lányosabban	`lány/NOUN[ATTRIB]/ADJ[COMPAR]/ADJ[MANNER]/ADV`
láthatósági	`lát/VERB[MODAL_PART]/ADJ[ABSTRACT]/NOUN[MET_ATTRIB]/ADJ`
tárgyiasít	`tárgy/NOUN[MET_ATTRIB]/ADJ[ATTRIB]/ADJ[TRANS_RESULT]/VERB`

A `VERB`, `NOUN`, `ADJ`, `NUM` kategóriák között minden irányban lehetséges képzés (a denumerális igeképzés kivételével). Az `ADV` kategória nem igazán képezhető, a `POSTP` kategóriából egyedül melléknévképzés lehetséges. A függelék (A4) táblázatában megtalálható a jelenleg használt képzők listája.

5.2. Szóösszetételek formalizmusa

A szóösszetételek annotációjához szükség van az összetételi tagok elemzésére. Szerencsére inflektált összetételi tag tipikusan csak az összetétel utolsó eleménél fordul elő, ezért az összetételi tagokat elegendő csak az (esetlegesen előforduló képzési annotációval együtt) kategóriájukkal elemezni. A formalizmust követően néhány példát az alábbiakban találunk.

(36) Az elemzés formalizmusa szóösszetétel esetén

Szóalak
`lemma_1/KÉPZÉS_ELEMZÉSE1+...+lemma_n/KÉPZÉS_ELEMZÉSE<INFL_ANNOTÁCIÓ>`

(37) Példák összetételek elemzésére

a. inflektált összetételek

eladja	`el/PREV+ad/VERB<DEF>`
vérfarkasok	`vér/NOUN+farkas/NOUN<PLUR>`
sötétzöldje	`sötét/ADJ+zöld/ADJ<POSS>`
kanárisárgáé	`kanári/NOUN+sárga/ADJ<ANP>`

b. képzett összetételi tagok

zúzottkő	`zúz/VERB[PERF_PART]/ADJ+kő/NOUN`
háromszínű	`három/NUM+szín/NOUN[INAL_ATTRIB]/ADJ`

c. többszörös összetétel

birsalmasajt	`birs/NOUN+alma/NOUN+sajt/NOUN`

6. Alkalmazások

A BME MOKK kutatócsoportjának irányításával számos olyan nyelvtechnológiai eszköz fejlesztése történt meg, amelyek a cikkünkben bemutatott morfológiai reprezentációt felhasználva végeznek automatikus szövegfeldolgozási műveleteket. Az alábbiakban bemutatjuk ezek közül a cikkünk szempontjából a legfontosabbakat.

6.1. Morfológiai leírás

A **hunlex** formalizmus egy morfológiai leírások reprezentálására alkalmas formális nyelv. A nyelvész szakértő ebben írhatja le egy adott nyelv morfológiai szabályait, illetve az ezekhez kapcsolódó morfológiai szótárakat. A hunlex reprezentáció specifikálja egy morfológiai elemző (vagy generáló) program viselkedését

az adott nyelvre, azaz szóalakokhoz morfológiai annotációkat rendel. Munkatársaink számos nyelvre dolgoztak ki morfológiai leírást ebben a formalizmusban, ezek közül a **morphdb.hu** erőforrás a jelen cikkben bemutatott magyar nyelvű annotációs rendszer megvalósítása (Trón et al. 2006). A szintén hunlex névre hallgató szoftver a nyelvészek által könnyen olvasható és módosítható magas szintű morfológiai erőforrásból egy úgynevezett aff/dic (affixumlista/szótár) formátumú, „gépközelibb" erőforrást állít elő. Ez az egyszerű formalizmus szabványnak tekinthető, a Firefox és OpenOffice szoftverek komponenseként több száz millió számítógépre telepített **hunspell** helyesírásellenőrző rendszer is ezt használja a morfológiai információ leírására.

6.2. Morfológiai elemzés

Egy morfológiai elemző szoftver feladata egy szóalakhoz megadni az összes legális morfológiai elemzést. Több olyan morfológiai elemző szoftverimplementáció is létezik (**rfst**, **ocamorph**, **jmorph**), amely az aff/dic formátumú morfológiai erőforrásra támaszkodva elvégzi ezt a feladatot (Trón et al. 2005).

6.3. Morfológiai egyértelműsítés

A **hunpos** morfológiai egyértelműsítő azt a problémát orvosolja, hogy egy adott szóalakhoz a morfológiai elemző gyakran több legális elemzést is talál (Halácsy et al. 2007). Az egyértelműsítő ezek közül választja ki a mondatbeli szövegkörnyezet alapján legvalószínűbbet. Ennek a feladatnak a hibátlan megoldásához a mondat teljes megértése lenne szükséges, amely a technológia mai állása mellett természetesen nem lehetséges. Ugyanakkor magyar nyelvre a morfológiai egyértelműsítés feladata – automatikusan feltárt egyszerű statisztikai szabályszerűségek kiaknázásával – elfogadható (tokenenként 95% feletti) pontossággal megoldható.

6.4. Tulajdonnév-felismerés (NER) és a főnévi csoportok felismerése (NP-chunking)

A morfológiai egyértelműsítés hasznos előfeldolgozási lépésként szolgálhat magasabb szintű nyelvfeldolgozási műveletek elvégzése előtt. Ilyen rendszerre példa a **hunner** tulajdonnév-felismerő, amely a szövegben automatikusan azonosítja

és klasszifikálja a személyneveket, helyneveket, szervezetneveket és egyéb tulajdonneveket (Varga–Simon 2007). A jelen kötetben bemutatott **hunchunk** NP-daraboló (főnévicsoport-felismerő) is támaszkodik működése során a morfológiailag feldolgozott szövegre (l. Recski–Varga 2012).

6.5. Gyakorisági korpusz

Korpusznyelvészeti és pszicholingvisztikai kutatásokhoz hasznos segédeszköz a **Szószablya** webes **gyakorisági szótár** (Szalai–Halácsy 2008). Ennek alapja a magyar nyelvű webről reprezentatív módon kigyűjtött több mint félmilliárd tokennyi szöveg automatikus morfológiai egyértelműsítése nyomán épített gyakorisági táblázat. A keresések eredményei a legkülönbözőbb morfológiai és fonológiai szempontok alapján szűrhetők és rendezhetők. A morfológiai reprezentáció hierarchikus mivolta megkönnyíti, hogy a gyakorisági szótár nyelvész felhasználója kereséseihez a legkülönbözőbb módon specifikálhasson morfoszintaktikai relációkat.

Irodalom

Csendes, Dóra – János Csirik – Tibor Gyimóthy 2004. The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. Lecture Notes in Artificial Intelligence 3206: 41–48.

Erjavec, Tomaž – Monica Monachini 1997. Specifications and notation for lexicon encoding. Deliverable D1.1 F. Multext-East Project COP-106. http://nl.ijs.si/ME/CD/docs/mte-d11f/

Halácsy, Péter – András Kornai – Csaba Oravecz 2007. HunPos – An open source trigram tagger. In: Sophia Ananiadou (szerk.): Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume. Proceedings of the Demo and Poster Sessions. Prague: Association for Computational Linguistics. 209–212.

Kornai András 1989. A főnévi csoport egyeztetése. Általános Nyelvészeti Tanulmányok 17: 183–211.

Melcsuk, Igor 1965. A magyar főnév birtokos személyragjainak morfológiai felépítéséről. Magyar Nyelv 61: 264–275.

Moravcsik, Edith 1997. Hungarian adjectives from a typological point of view. Kézirat. University of Wisconsin-Milwaukee, Milwaukee.

Recski Gábor – Varga Dániel 2012. Magyar főnévi csoportok azonosítása. Általános Nyelvészeti Tanulmányok 24: 81–95.

Szalai András – Halácsy Péter 2008. Szószablya gyakorisági szótár. http://szotar.mokk.bme.hu/szoszablya/

Trón Viktor 2002. Attribútum–érték struktúrák. In: Kálmán László – Trón Viktor – Varasdi Károly (szerk.): Lexikalista elméletek a nyelvészetben. Budapest: Tinta Könyvkiadó. 333–344.

Trón, Viktor – László Németh – Péter Halácsy – András Kornai – György Gyepesi – Dániel Varga 2005. Hunmorph: Open source word analysis. In: Martin Jansche (szerk.): Proceedings of the ACL Workshop on Software. Stroudsburg, PA: Associaton for Computational Linguistics. 77–85.

Trón, Viktor – Péter Halácsy – Péter Rebrus – András Rung – Péter Vajda – Eszter Simon 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In: Nicoletta Calzolari – Khalid Choukri – Aldo Gangemi – Bente Maegaard – Joseph Mariani – Jan Odijk – Daniel Tapias (szerk.): Proceedings of 5th International Conference on Language Resources and Evaluation. Genoa: European Language Resources Association (ELRA). 1670–1673.

Varga, Dániel – Eszter Simon 2007. Hungarian named entity recognition with a maximum entropy approach. Acta Cybernetica 16: 293–301.

Függelék

(A1) A hunmorph főkategória-jegyei

a. inflektálható (és továbbképezhető)

névszóként:

NOUN	főnév *(noun)*
ADJ	melléknév *(adjective)*
NUM	számnév *(numeral)*
DET	determináns *(determiner)* [8]

igeként:

VERB	ige *(verb)*

b. nem inflektálható (és nem továbbképezhető)

ADV	határozószó *(adverb)*
POSTP	névutó *(postposition)*
	– marginálisan esetjelölés és mn-képzés lehetséges
ART	névelő *(article)* [3]
CONJ	kötőszó *(conjunction)* [198]
PREP	prepozíció *(preposition)* [5]
PREV	igekötő *(preverb)* [105]
UTT-INT	mondatszó/indulatszó *(utterance/interjection)*
ONO	hangutánzó *(onomatopoeic)*

c. egyéb, írott nyelvi kategória

PUNCT	központozási jel *(punctuation)*

(A2) A névszói inflexiós jegyhierarchia

```
{NOUN ADJ NUM DET}
 <POSTP
    <ALATT>
    <...>>                                          (az összes névutó)
 <PLUR
    <FAM>>
 <PERS
    <1>
    <2>>
 <POSS
    <PLUR>
    <1>
    <2>>
 <ANP
    <PLUR>
    <ANP
       <PLUR>>>
 <CAS
    <ACC>
    <DAT>
    <...>>                                          (az összes [17] jelölt esetrag)
```

(A2′) Nem létező névszói jegykombinációk

a. egymást kizáró testvérjegyek

személyes névmás /birtokos nem lehet egyszerre 1. és 2. személyű:

*<PERS<<1><2>>>

*<POSS<<1><2>>>

egyszerre két esetjelölés/névutójelölés nem lehetséges:

*<CAS<<*A*><*B*>>>

*<POSTP<<*A*><*B*>>>

b. kötelezően folytatandó jegyek
nem-nominativusi esetnél/névutónál meg kell jelölni a konkrét esetet

*<CAS>

*<POSTP>

(A3) Az igei inflexiós jegyhierarchia

```
VERB
  <MODAL>
  <PAST>
  <COND>
  <SUBJUNC-IMP>
  <INF>
  <PLUR>
  <PERS
    <1
      <OBJ2>>
    <2>>
  <DEF>
```

(A3′) Nem létező igei jegykombinációk

a. egymást kizáró testvérjegyek
a jelölt idő/módok és az inf. páronként kölcsönösen kizárók:

```
<PAST>|<COND>|<SUBJUNC-IMP>|<INF>
```

ige nem lehet egyszerre 1. és 2. személyű:

```
*<PERS<<1><2>>>
```

b. kötelezően folytatandó jegyek
véges ige nem-3. személye esetén a személyt meg kell jelölni:

```
*<-INF><PERS>
```

(A4) A hunmorph képzői (a más elemzőkben gyakran inflexióként elemzettek kiemelve)

	VERB	NOUN	ADJ	ADV	NUM
VERB	*-gat* [**FREQ**] *-tat* [**CAUS**] *-ódik* [MEDIAL]	*-ás* [GERUND]	*-ó* [**IMPERF_PART**] *-ott* [**PERF_PART**] *-andó* [**FUT_PART**] *-ható* [MODAL_PART] *-hatatlan* [NEG_MODAL_PART] *-atlan* [NEG_PERF_PART]	*-va* [**PART**] *-ván* [**PERF_PART**]	—
NOUN	*-ozik* [ACT] *-ol* [ACT2] *-kodik* [REG_ACT]	*-né* [MRS] *-cska* [DIMIN]	*-os* [ATTRIB] *-i* [MET_ATTRIB] *-jú* [INAL_ATTRIB] *-tlan* [NEG_ATTRIB] *-mentes* [NEG_ATTRIB2] *-nyi* [QUANTITY] *-szerű* [TYPE1] *-féle* [TYPE2] *-beli* [LOC_INE]	*-stul* [COM] *-képpen* [ESS-FOR] *-nként* [PERIOD]	—
ADJ	*-ít* [TRANS_RESULT] *-ul/-odik* [INTRANS_RESULT]	*-ság* [ABSTRACT]	*-bb* [**COMPAR**] *-bbik* [**COMPAR_DESIGN**] *leg- -bb* [**SUPERLAT**] *leg- -bbik* [**SUPERLAT_DESIGN**] *-cska* [DIMIN] *-as* [ATTRIB]	*-an/-ul* [**MANNER**]	—
NUM	—	*-dika* [DATE]	*-as* [ATTRIB] *-szori* [ITER_ATTRIB] *-szoros* [MULTIPL_ATTRIB]	*-an* [**AGGREG**] *-szor* [MULTIPL-ITER] *-adszor* [ORD-ITER]	*-ad* [FRACT] *-adik* [ORD]
POSTP	—	—	*-i* [ATTRIB]	—	—

A general-purpose morphological annotation system

Abstract: Ideally, a morphological annotation scheme should encode all and only morphosyntactic information, abstracting away from spoken and written form. Annotation schemes based directly on the component morphs inherit the difficulties of segmentation, in particular for fused morphemes and suppletive forms. In this paper we propose an indirect annotation scheme based on binary morphosyntactic features (attribute–value pairs) arranged in a tree in such a manner that only positive (marked) features can dominate other nodes. This restriction makes it possible to convert the structure to unary features, thereby directly reflecting the markedness of a form. In the last part of the paper we describe some applied computational systems that employ this annotation scheme.

Keywords: Hungarian, morphology, annotation, tagging, inflection

Magyar főnévi csoportok azonosítása

Recski Gábor[1] – Varga Dániel[2]

[1]*MTA SZTAKI Nyelvtechnológiai Kutatócsoport, Budapest*
[2]*BME Média Oktató és Kutató Központ, Budapest*
recski@sztaki.hu; daniel@mokk.bme.hu

Cikkünkben bemutatjuk a hunchunk eszközt, amellyel főnévi csoportok (NP) azonosítása végezhető magyar nyelvű szövegeken. Az NP-azonosítás hasznos összetevő olyan magasabb szintű gépi szövegfeldolgozási feladatok megoldásakor, mint az információkinyerés vagy a gépi fordítás. A feladatot felügyelt gépi tanulási módszerrel oldjuk meg, tanítókorpuszunk alapja a Szeged Treebank, egy szintaktikailag annotált 1.2 millió szavas korpusz. A szakirodalomban elterjedt megoldással az NP-azonosítás feladatát először szekvencia-címkézési feladattá fogalmazzuk át. Ezután egy maximum entrópia (ME) modellt tanítunk, jegyekként a szóalakból, szófajból és morfológiai jegyekből kinyert információkat alkalmazva. Az ME modell által hozott lokális döntéseket a címke-szekvenciákon épített nyelvmodell segítségével harmonizáljuk. Rendszerünk 90.28%-os F-mértéket ér el a Szeged Treebank kiértékelésre elkülönített részén. A rendszert a CoNLL 2000 Shared Task-on is betanítottuk és kiértékeltük, ahol versenyképes 93.79%-os F-mértéket ért el az angol alap-NP-k azonosításának feladatán. Az implementált algoritmus nyelv- és feladatfüggetlen, így sikeresen alkalmaztuk a fenti feladatokon túl tulajdonnév-felismerésre és magyar mondatok közvetlen összetevőinek azonosítására is.

Kulcsszavak: NP-felismerés, mondattani elemzés, felügyelt tanulás, maximum entrópia, rejtett Markov-modellek

1. Chunk, chunkolás, NP-felismerés

1.1. Meghatározás

Sem a *chunk*, sem a *chunkolás* kifejezés nem rendelkezik általánosan elfogadott definícióval a számítógépes nyelvészeti szakirodalomban. A *chunk* kifejezést elsőként Abney (1991, 257) használja, aki a mondat olyan, egymással át nem fedő egységeit nevezi így, melyek „egyetlen tartalmas szóból és az őt körülvevő funkciószavakból állnak". Elsősorban Gee–Grosjean (1983) munkájára alapozva, akik a mondat pszicholingvisztikai egységeinek megnevezésére a **performancia-szerkezet** *(performance structure)* fogalmát vezetik be, Abney a chunkokat olyan egységekként kezeli, amelyek nem szükségszerűen esnek egybe mondattani összetevőkkel. A chunkolás témakörében végzett újabb kutatások ezzel szemben egyöntetűen olyan meghatározásokkal élnek a chunk fogalmára,

amelyek lehetővé teszik, hogy egy mondat chunkjait annak teljes elemzési fájából egyértelműen előállíthassuk. A gyakorlatban ez azt jelenti, hogy valamennyi chunk-típust szintaktikai frázisok valamely csoportjának feleltetünk meg és ezzel a chunkolás a teljes mondattani elemzés részfeladatává válik. Egy minden szintaktikai kategóriára kiterjedő definícióhalmazt találunk a CoNLL 2000 (Tjong Kim Sang–Buchholz 2000) versenyfeladat[1] leírásában, ennek alapján nyerik ki a chunkolási mintát a Penn Treebank (Marcus et al. 1994) szintaktikailag annotált korpuszból.

A chunkolás legelterjedtebb formája az NP-chunkok azonosítása. Az NP-chunkolás témakörében született egyik legismertebb mű szerzői Ramshaw és Marcus (1995), akik az **alap-NP-k** (más néven baseNP-k, nem-rekurzív NP-k vagy minimális NP-k) azonosításával foglalkoztak, értve ezen azokat a főnévi csoportokat, amelyek nem tartalmaznak másik főnévi csoportot. A CoNLL 2000 feladatához is ezt a meghatározást használta fel Tjong Kim Sang és Buchholz, chunkolási feladatukat az eredeti, Ramshaw és Marcus által megadott módon tűzték ki, és ez a feladat szolgál mindmáig terepként a különböző gépi tanulási algoritmusok összehasonlítására.

Magyar nyelvű NP-chunkok azonosításával foglalkozik Váradi (2003) és Prószéky et al. (2004). Az általunk bemutatandó rendszer ezektől egyrészt abban különbözik, hogy nem szabályalapú, hanem statisztikai módszerekkel azonosítja a chunkokat (e módszereket a 3. pontban részletesen bemutatjuk), másrészt pedig az NP-chunknak a szokásostól eltérő definícióját használja (ezt a 2.1. pontban részletezzük). A magyar NP-chunkok azonosítására alkalmas eszközöket – beleértve az itt bemutatandó `hunchunk` eszközt is – Miháltz (2011) sztenderdizált körülmények között hasonlítja össze.

1.2. Alkalmazások

A teljes mondattani elemzés *(parsing)* sokkal alacsonyabb pontossággal végezhető, mint a chunkolás, így az utóbbi lehetővé teszi, hogy kevesebb, de megbízhatóbb információt nyerjünk ki egy mondat szerkezetéről. Az NP-chunkok azonosítása egyes információ-kinyerési feladatokhoz is hozzájárul, amennyiben a főnévi csoportokkal egyszersmind a mondatban szereplő entitásokat is azonosítja.

[1] A versenyfeladat *(shared task)* azt jelenti, hogy egy konkrét feladatot, jelen esetben az angol NP, VP, PP stb. chunkok megtalálását, egységes adathalmazon és sztenderdizált kiértékelési eljárás mellett a tudományos közösség elé tárnak, így az annak megoldására javasolt különböző eljárások pontosan összemérhetővé válnak.

2. A tanulóadatok előállítása

A legelterjedtebb NP-chunkoló eszközökhöz hasonlóan az általunk készített rendszer is **felügyelt tanulásra** *(supervised learning)* épül, azaz az alkalmazás egy manuálisan előállított tökéletes minta alapján, statisztikai módszerekkel tárja fel a szavak egyes tulajdonságai és chunkhoz tartozásuk közti összefüggéseket. (A rendszer működésének részleteit a 3. pontban mutatjuk be.) Ezért első lépésként szükségünk van egy ilyen ún. tanulóadat-halmaz létrehozására egy mondattanilag annotált korpusz segítségével.

2.1. A feladat

Bár a szakirodalomban NP-chunkoláson általában az alap-NP-k megtalálását értik, mi egy ezzel gyakorlatilag ellentétes definíciót választunk, amennyiben NP-chunknak tekintünk minden olyan szósorozatot, amely a mondat elemzési fájában NP-t alkot és ezt az NP-t nem tartalmazza magasabb szintű főnévi csoport (ezeket fogjuk **maximális NP**-knek nevezni). Ez a definíció lehetővé teszi, hogy a chunkolással a mondat közvetlen összetevőit különítsük el és a mondatban szereplő igék vonzatkeretét feltérképezzük, ami a gépi fordításban különös jelentőséggel bír. Ezen túl a maximális NP-k azonosítása az információ-kinyerésben is hasznos lehet, amennyiben a mondatokban szereplő főneveket összes bővítményükkel együtt nyerjük ki. Fontosnak tartjuk megemlíteni, hogy az NP-chunk itt használt definíciója csupán a korpuszt előállító rendszer beállításaitól függ, így amennyiben eltérő egységeket tekintünk chunknak – például a fent említett módon az alap-NP-ket szeretnénk azonosítani – úgy ahhoz egyszerűen állítható elő megfelelő tanítókorpusz. A jelen cikkben bemutatott rendszer tehát bármilyen módon definiált NP-chunk azonosítására alkalmas, választásunk jelentősége abban rejlik, hogy a rendszer – a későbbiekben részletesen ismertetett – paramétereit, így különösen a tanításhoz használt jegyek összetételét úgy választottuk meg, hogy a maximális NP-k azonosításában a lehető legjobb eredményt érje el.

Tanulóadataink forrása az 1,43 millió szóból álló Szeged Treebank korpusz (Csendes et al. 2005), mely különböző műfajú (szépirodalom, újságcikkek, jogszabályi szövegek, szoftverdokumentációk stb.), morfológiailag annotált és mondattanilag elemzett szövegekből áll. Egy Treebank a benne szereplő mondatok teljes elemzési fáját tartalmazza, így az NP-k azonosításához szükségesnél bővebb szerkezeti információkat is, amelyekre az NP-korpusz előállításakor nincsen szükségünk. Az általunk elvégzett eljárás lényege, hogy a Treebank-ben található elemzési fákat bejárjuk, az azokban található szavakat pedig a rendel-

kezésre álló morfológiai információkkal együtt a korpuszhoz adjuk, feljegyezve azt is, hogy részét képezik-e maximális NP-nek, azaz a mi definíciónk szerinti NP-chunknak.

Az NP-felismerési feladatot címkézési feladatként oldjuk meg, ami azt jelenti, hogy egy chunkolás alapján a mondat minden szavához címkét rendelhetünk, amely azt írja le, hogy az adott szó részét képezi-e NP-nek. A címkézés legelterjedtebb – és a CoNLL 2000 feladatban is használatos – módja az ún. IOB konvenció, mely három címkét különböztet meg: az NP-k első szava a `B-NP`, többi szava az `I-NP` címkét, az NP-hez nem tartozó szavak pedig az `O` címkét kapják (Tjong Kim Sang–Veenstra 1999).

Ez a jelölés nem különbözteti meg az NP végén álló szavakat az NP közepén állóktól, az önmagukban főnévi csoportot alkotó szavakat pedig ugyanúgy jelöli, mint az NP-t kezdő szavakat. Feladatunkban célszerűbbnek bizonyult az 5 címkét használó Start-End konvenciót használni, mely a fentieket pontosítva az NP-végző tokeneket `E-NP` címkével, az egyszavas NP-ket pedig `1-NP` címkével jelöli (Uchimoto et al. 2000). Az effajta címkézést megvalósító rendszerünknek biztosítania kell, hogy szabálytalan – azaz NP-chunkolásnak meg nem feleltethető – címkesorok ne jöhessenek létre; például az `I-NP` címkét nem követheti legálisan a `B-NP` címke.

Az adatok kinyerésekor feljegyezzük azt is, hogy az adott NP-be milyen mélyen ágyazódnak be további NP-k, így lehetőségünk nyílik egyfajta komplexitásfogalom alapján több chunk-típust megkülönböztetni. Így például az (1) alatti frázisokat, ha maximális NP-t alkotnak, rendre `N_1`, `N_2`, `N_3` típusú chunknak címkézzük.

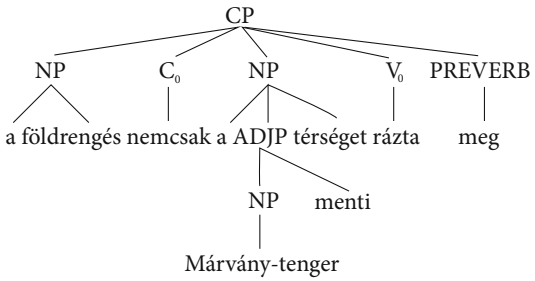
(1) a. [$_{NP}$ Az elnök]

b. [$_{NP}$ [$_{NP}$ Az Egyesült Államok] [$_{NP}$ elnöke]]

c. [$_{NP}$ [$_{NP}$ [$_{NP}$ Az Egyesült Államok] [$_{NP}$ elnökének]] irodája]]

Ezen információ kinyerését nem tekintjük a címkéző feladatának, csupán a gépi tanulási feladatot könnyítjük meg vele: az NP-felismerés pontosságának javulását várjuk attól, hogy a modellnek lehetősége van olyan szabályszerűségeket is tárolni, melyek csak bizonyos „mélységű" NP-kre jellemzőek. Végül a legjobb eredményeket azzal a címkézéssel értük el, ahol csupán a legalacsonyabb – tehát további NP-t nem tartalmazó – főnévi csoportokat különböztettük meg (`N_1`) a komplexebbektől (ezeket `N_2+`-szal jelöltük).

A fenti definíció és címkézés eredményeképpen a Szeged Treebank 1. ábrán látható mondata az NP-korpuszban az 2. ábrán látható módon jelenik meg.



1. ábra. A Szeged Treebank egy mondatának elemzése

A	földrengés	nemcsak	a	Márvány-tenger	menti	térséget	rázta	meg
B-N_1	E-N_1	O	B-N_2+	I-N_2+	I-N_2+	E-N_2+	O	O

2. ábra. A Szeged Treebank egy mondatának chunk-címkézése

3. A hunchunk rendszer

3.1. Felügyelt tanulás

Az NP-felismerő eszközök többsége **felügyelt tanulásra** épül. Ez a kifejezés azt jelenti, hogy a számítógéppel elvégezni kívánt feladathoz (általában emberi munkával) **tanulóadat-halmazt** hozunk létre, amely nagyszámú bemenet-kimenet párból áll. Egy felügyelt tanulóalgoritmus feladata, hogy a tanulóadat alapján automatikusan tárjon fel összefüggéseket a bemenet és kimenet között, és az így létrehozott **modell** segítségével később új, korábban nem látott bemenetre meghatározza a legvalószínűbb kimenetet. **Címkézésről** beszélünk, ha a kimenet egy rögzített véges halmazból kerül ki.

Szekvenciális címkézési feladatról beszélünk, ha a feladat struktúrája olyan, hogy a bemenet egy sorozat (pl. egy mondat tokenjeinek sorozata), a kimenet pedig a sorozat minden eleméhez címkét rendel. Felhívjuk a figyelmet arra, hogy a szekvenciális címkézés szigorú értelemben nem alesete a címkézésnek, hiszen kimenetének hossza függ a bemenet hosszától. Az NP-felismerésen kívül számos más nyelvtechnológiai feladat is megoldható felügyelt szekvenciális címkézési módszerekkel, így például a szófaji címkézés *(part-of-speech tagging, POS-tagging)* és a tulajdonnév-felismerés *(named entity recognition, NER).*

A felügyelt tanulás feladatát megoldó matematikai módszereknek a bemenetet valamilyen strukturált formában kell megkapniuk. Ennek egy standard módja, hogy a bemenetet (vagy szekvenciális címkézésnél annak egyes tokenjeit) ún. **jegyek** halmazával írjuk le; ilyen jegyre lehetséges példa az, hogy a token

nagy betűvel kezdődik-e, ige-e, többes számban áll-e stb. A tanulóalgoritmus ezek alapján modellépítéskor olyan statisztikai összefüggéseket tár fel, mint például hogy a token nagy betűvel kezdődése hogyan befolyásolja annak az esélyét, hogy a token a `B-N_1` címkét kapja. Címkézéskor az újonnan érkezett bemenet jegyei alapján megállapítja, hogy mi a modell által legvalószínűbbnek ítélt címkézés. Noha az itt bemutatott rendszer kizárólag bináris jegyeket használ – azaz egy jegynek csak két értéke lehet, például egy szó vagy nagybetűs, vagy nem –, a tanulóalgoritmus lehetővé teszi, hogy egy-egy jegy tetszőleges valós értéket felvegyen.

A felügyelt tanítás módszere mögött ott van az alapfeltevés, hogy a tanítóadat-halmazon feltárt összefüggések relevánsak lesznek az új adatok feldolgozásakor is. Ez az alapfeltevés csak akkor tekinthető kellően megalapozottnak, ha az új szöveg jellege (zsánere) nem tér el nagyon a tanulóadatétól. Ugyanakkor megállapítható, hogy a rendszerünk által automatikusan feltárt milliónyi statisztikai összefüggés közül a statisztikai értelemben legerősebbek zsánertől független, általános jelenségeket számszerűsítenek.

A 3. pont hátralévő részében részletesen bemutatjuk, hogy rendszerünk számára hogyan reprezentáljuk jegyekkel a bemenetet, majd ismertetjük magát a tanuló- illetve címkézőalgoritmust.

3.2. Jegyek

Egy token legfontosabb jegyei a szóalak és annak valamennyi morfológiai jegye. A szóalak jegyként való felvétele szükséges ahhoz, hogy a tanulóalgoritmusnak lehetősége legyen olyan statisztikai összefüggések megtalálására, mint például: „a *ha* szó ritkán áll NP belsejében". Egy szó morfológiai jegyeinek reprezentációjára számos különböző konvenció létezik. A Szeged Treebank az MSD-kódolást követi (Erjavec 2004), ezt az NP-korpusz létrehozásakor azonban átalakítottuk az ún. KR-formalizmusra (Kornai et al. 2004), mivel az általunk használt morfológiai címkéző, elemző és egyértelműsítő egyaránt ezt a formátumot követi. A KR-formalizmus előnye, hogy egy szó valamennyi morfológiai jegyét külön karaktersorozatnak felelteti meg, így a KR-kódok jelentése kompozicionális (például az *asztalát* szó a `NOUN<POSS><CAS<ACC>>` kódot kapja). A KR-konvencióra való áttérés lehetővé tette, hogy ezen kódok valamennyi, egy-egy morfológiai jegynek megfeleltethető összetevőjét jegyként vegyük fel (l. 1. táblázat).

Fontos megemlítenünk, hogy míg a tanulókorpuszban a morfológiai jegyek készen rendelkezésre állnak, addig nem látott szöveg címkézésekor más a helyzet. Hogy korábban nem látott mondatok NP-chunkolását is elvégezhesse a rendszer,

jegytípus	jegyek
szóalak	`mesélte`
n-gram	`mes`, `esé`, `sél`, `élt`, `lte`
KR	`VERB`, `PAST`, `DEF`

1. táblázat. A *mesélte* szó jegyei

a nyers szöveg szavaihoz előfeldolgozási lépésként morfológiai címkéket kell rendelnünk. A `hunpos` morfológiai címkéző (Halácsy et al. 2007), maga is felügyelt címkézési módszereket alkalmazva, megoldja ezt a feladatot.

Mivel feltételezzük, hogy egy szó címkéjének valószínűségét a környezetében található szavak tulajdonságai is befolyásolják, ezért a jegyeket minden tokenre annak 5 szavas környezetében értékeljük ki, tehát egy-egy token jegyei közt szerepel az önmagára, valamint az őt megelőző és követő 5-5 tokenre vonatkozó információ is. A `-1_form=Az` jegy például azt jelenti, hogy az ezzel a jeggyel rendelkező tokent megelőző szó alakja *Az*, a `2_kr=PLUR` jegy azt jelenti, hogy a kettővel utána következő szó többesszámú stb.

Bevezettük továbbá az ún. **szófajminta-jegyet**, mely egy szó adott hosszúságú környezetében az egymást követő szavak szófaji cimkéinek sorozatait írja le a következő módon: ha a jegy sugarát r-el, egy mondat i-edik pozíciójában álló szót w_i-vel, szófaji címkéjét pedig p_i-vel jelöljük, úgy bármely w_i szóra jegyként vesszük fel a $p_{i-r} \dots p_{i+r}$ sorozat összes összefüggő részintervallumát. Más szavakkal minden szóhoz feljegyezzük, hogy valamekkora hosszúságú környezetében milyen szófajú (kategóriájú) szavak sorozatait találjuk. Minden jegyérték elején feltüntetünk két számot, amelyek azt jelölik, hogy a szóban forgó tokenhez képest az adott szófaj-sorozat hol helyezkedik el. Így a 2. táblázatban szereplő mondat *csapos* szava a 3. táblázatban szereplő jegyeket veszi fel a szófajminta jegy értékeként.

A	csapos	mesélte	,	hogy	milyen	szép	kés	van	bennem	.
`ART`	`NOUN`	`VERB`	`PUNCT`	`CONJ`	`ADJ`	`ADJ`	`NOUN`	`VERB`	`NOUN`	`PUNCT`

2. táblázat. A *Szeged Treebank* egy mondatának szófaji címkézése

A KR-mintákat kiválasztó jegy sugarát növelve a chunkolás minősége is nő, 3-nál nagyobb sugár mellett azonban a jegyek túl magas száma nem teszi lehetővé a modell tanítását.

Az alábbiakban bemutatjuk azt a modellezési eljárást, amellyel a szekvenciális címkézési feladatot megoldjuk. A matematikai részletekben legkevésbé sem elmerülve inkább arra törekszünk, hogy a statisztikai módszerekben nem feltétlenül járatos olvasó képet kapjon arról, hogy milyen heurisztikák vezérlik egy,

-1_1_ART+NOUN
-1_2_ART+NOUN+VERB
-1_3_ART+NOUN+VERB+PUNCT
-1_4_ART+NOUN+VERB+PUNCT+CONJ
0_2_NOUN+VERB
0_3_NOUN+VERB+PUNCT
0_4_NOUN+VERB+PUNCT+CONJ
1_3_VERB+PUNCT
1_4_VERB+PUNCT+CONJ
2_4_PUNCT+CONJ

3. táblázat. Egy token szófajminta-jegyei ($r = 4$)

a miénkhez hasonló statisztikai modell megalkotását, és milyen standard technikákra támaszkodhat a modellező kutató. Módszerünk részletesebb leírása megtalálható Recski et al. (2009)-ben.

3.3. Maximum Entrópia modell

A jegyek és címkék közötti összefüggések feltárásához a Maximum Entrópia módszert (MaxEnt) választottuk (Ratnaparkhi 1998). Ezen eljárást már sikerrel alkalmaztuk összetevőként többek között morfológiai címkézés (Halácsy et al. 2005) és tulajdonnév-felismerés (Varga–Simon 2007) feladatának megoldásakor. A módszer fontos tulajdonsága, hogy felhasználásakor egy-egy szóra – annak jegyei alapján – nem csupán a legvalószínűbb címkét adja meg, hanem valamennyi lehetséges címkéhez valószínűségeket rendel; a modell ezen képessége elengedhetetlen ahhoz, hogy egy mondat legvalószínűbb címkézését megtaláljuk (ez utóbbi folyamatot a későbbiekben ismertetjük majd).

A Maximum Entrópia módszer pontos ismertetésére itt nem nyílik lehetőségünk, de egy példán keresztül mutatjuk be alkalmazásunk szempontjából legfontosabb sajátosságait. Tegyük fel, hogy a tanulókorpuszt vizsgálva megállapítjuk, hogy a `kr=PLUR` jeggyel rendelkező tokenek körében kétszer akkora az `E-N_1` címke esélye, mint az ilyen jegyet nem kapott tokenek körében. Ekkor első közelítésben kiindulhatunk abból, hogy a jegy jelenléte kétszeresére növeli a `E-N_1` címke esélyét a nem látott adatokon is. Ezt úgy mondjuk, hogy a `kr=PLUR` jegy jelenlétére nézve a `E-N_1` címkének kettő az **esélyhányadosa** *(odds ratio).* Ha több jegy áll rendelkezésünkre a döntéshez, akkor az egyes jegyekhez tartozó esélyhányadosokat összeszorozva kiszámolhatjuk az egyes címkék egymáshoz viszonyított esélyeit. Így működik az ún. Naiv Bayes módszer, ahol a modell egyszerűen ezeknek az esélyhányadosoknak a tanulókorpusz alapján épített táblázata. A Naiv Bayes módszer nyilvánvaló hibája, hogy figyelmen kívül hagyja, hogy

a jegyek által adott esélyhányadosok valójában függenek attól, milyen más jegyek vannak jelen. Például a `kr=PLUR` jegyből kinyerhető információ redundáns akkor, ha már kinyertük a `form=azok` jegyből megszerezhető információt. Ezt a problémát orvosolja a Maximum Entrópia módszer. Egy MaxEnt-modell ugyanúgy a jegyekhez rendelt esélyhányadosok táblázata, mint egy Naiv Bayes modell, de olyan módon határozza meg az esélyhányadosokat, hogy az ilyen kettős számlálásokat lehetőség szerint kiküszöbölje.

3.4. Átmenetmodell

Egy szekvenciális címkézési feladat megoldásának egy egyszerű, kevéssé kifinomult módja, hogy eltekintünk a szekvencialitástól: olyan modellt tanítunk, amely minden egyes tokenre külön-külön hozza meg a döntését, a token jegyei alapján, például az előző alpontban bemutatott MaxEnt módszerrel. Már egy ilyen modell is képes jó minőségű chunkolásra (l. a 4.2. pontot), de nem használja ki explicit módon a tokenek címkéi közötti összefüggéseket. Ennek következtében például az is előfordulhat, hogy a kimenetben egy `I-NP` címkét `B-NP`, egy `O` címkét `E-NP` stb. követ, pedig ezek szabálytalan, chunkolásként értelmezhetetlen címkesorozatok, amelyek a tanulókorpuszban nem is fordultak elő. A szabálytalan szekvenciák kiszűrésén túl azt is figyelembe szeretnénk venni, hogy a szabályos címkesorozatok sem egyformán gyakoriak: Az `E-NP` címkét például csak akkor követi `B-NP` vagy `1-NP`, ha egy NP-t közvetlenül követ egy másik – ez valószínűtlenebb, mint az `E-NP O` szekvencia, amikor egy NP-t NP-n kívüli szó követ, például egy ige.

A címkék közötti korrelációk figyelembevételének talán legegyszerűbb módja az **átmenetmodell**, amely csak a szomszédos címkék közti összefüggést modellezi, pontosabban azt, hogy egy adott címke után milyen eséllyel következik egy adott másik. Ennek megépítéséhez a tanítókorpuszban megszámoljuk valamennyi, két címke hosszúságú sorozatot (bigramot), majd minden címkére feljegyezzük, hogy azon bigramok közül, melyeknek ő az első tagja, milyen arányban szerepelnek a különböző címkék a második helyen.

Az átmenetmodell segítségével egyszerű szorzat formájában írhatjuk fel egy adott címkesorozat megfigyelésének valószínűségét:

$$P(t_1 t_2 \dots t_n) = P(t_n|t_{n-1})P(t_{n-1}|t_{n-2}) \dots P(t_2|t_1)P(t_1)$$

Ez a képlet várakozásainknak megfelelően nulla valószínűséget rendel a szabálytalan címkesorozatokhoz.

3.5. Címkézés

Ahogy az előző két alpontban bemutattuk, a tanítókorpusz alapján két modellt építettünk, a jegyek és egyetlen címke kapcsolatát feltáró Maximum Entrópia modellt, illetve a címkék egymáshoz való viszonyáról nyilatkozó átmenetmodellt. Ebben az alfejezetben azt tárgyaljuk, hogy egy új, a tanulókorpuszban feltehetőleg nem szereplő mondathoz hogyan találhatjuk meg a fenti két modell ismeretében legvalószínűbb címkézést. Az alább ismertetett módszert elsőként McCallum et al. (2000) alkalmazta címkézési feladatok megoldására.

A két modell kombinálásához – az itt most nem tárgyalt rejtett Markov-folyamatok (*Hidden Markov Model, HMM,* l. pl. Rabiner 1989) elmélete által motivált módon – leegyszerűsítő feltételezéssel élünk: azt tételezzük fel, hogy az az információ, amit egy címkéről az előző címke ad, független attól az információtól, amit a jegyek alapján a MaxEnt modell ad. Ennek a feltételezésnek a matematikailag pontos megfogalmazásához észre kell vennünk, hogy a függetlenség valószínűségszámítási értelemben nem teljesül, amennyiben mindkét modell előnyben részesíti a gyakoribb címkéket. Ez a megfontolás vezet az alábbi képlet végső formájához, amely (fix, csak w-től függő szorzótényezőtől eltekintve) megadja, hogy egy t címkesorozat mennyire teszi valószínűvé egy w mondat megfigyelését.

$$P(w|t) \sim \prod_i \frac{P(t_i|w)P(t_i|t_{i-1})}{P(t_i)}$$

A valószínűséget ($P(w|t)$) tehát úgy kapjuk meg, ha minden címke esetében összeszorozzuk a MaxEnt modell által szolgáltatott valószínűséget ($P(t_i|w)$) az átmenetvalószínűséggel ($P(t_i|t_{i-1})$), korrigálva a címke relatív gyakoriságával ($P(t_i)$). A képlet levezetését l. Recski et al. (2009).

A fenti képlet egy adott címkézés valószínűségét értékeli ki, de a mi feladatunk nem ez, hanem az összes lehetséges címkézés közül a legvalószínűbbet megtalálni. Az elvben lehetséges címkézések rendkívül nagy száma miatt az egyenkénti kiértékelésük nem lenne kivitelezhető. A fenti képlet speciális formája miatt azonban létezik hatékony algoritmus annak a címkézésnek a megtalálására, amely a képletet maximalizálja. Ez a Viterbi algoritmus, amelynek részleteit itt szintén nem ismertetjük, de lényege, hogy a mondat szavain balról jobbra végighaladva a mondat minden kezdőszeletéhez és minden lehetséges címkéhez meghatározza, hogy mi a kezdőszelet legvalószínűbb címkézése azok közül, amelyek az adott címkével végződnek.

A címkézőeszközünknek szabadon változtatható paramétere, hogy a címkemodellt – azaz az egyes chunk-címkék közötti átmenetvalószínűségeket tartalmazó modellt – a MaxEnt modellhez képest milyen súllyal vegye figyelembe (ez

bevett eljárás a rejtett Markov-modellezés területén). A fenti képletet ez úgy általánosítja, hogy a $P(t_i|t_{i-1})$ és a $P(t_i)$ kifejezéseket valamely pozitív λ kitevővel hatványozzuk. A λ paraméter legjobb eredményt biztosító értékét úgy kerestük meg, hogy a tanítókorpusz egy kisméretű részén megmértük, melyik beállítás adja a legjobb eredményt.

4. Kiértékelés

4.1. Módszertan

NP-felismerőnk kiértékeléséhez a korpusz tanításra fel nem használt, kb. százezer tokenből álló részét használtuk fel. A tesztkorpuszon lefolytatott cimkézések kimenetét a CoNLL 2000 feladat szigorú előírásait követve értékeltük ki, tehát akkor és csak akkor tekintettünk egy főnévi csoportot helyesen azonosítottnak, ha az eszközünk az NP minden tokenjét – és csak azokat – ismerte fel NP részeként.

Az eszköz teljesítményét – a szakirodalomban megszokott módon – mind **pontosság**, mind **fedés**, mind pedig a kettőből kiszámítható **F-pontszám** segítségével jellemezzük. A pontosság azt mutatja meg, hogy a címkéző által azonosított NP-k hány százaléka helyes, a fedés ezzel szemben annak mérőszáma, hogy a tényleges NP-k közül mennyit találtunk meg. A legtöbb osztályozási feladat során könnyen lehet e két értékből valamelyiket a másik rovására magasan tartani, ezért szokásos az ilyen eszközök teljesítményét az F-pontszámmal jellemezni, amely a pontosság és fedés harmonikus közepeként áll elő. (A harmonikus közép azt az elvárásrendszert formalizálja, amikor a tévesen NP-ként azonosított szövegrészek *(false positive)* ugyanolyan súlyú hibának minősülnek, mint az észre nem vett NP-k *(false negative)*.)

4.2. Eredmények

Az osztályozási feladatokon elért eredményeket szokásos eljárás szerint egy ún. **baseline** módszerhez hasonlítjuk, amely általában valamely egyszerű heurisztikát jelent, amelyhez képest jobb teljesítmény elérése a bemutatott módszer minimális célkitűzése. Az általunk választott baseline módszer lényege, hogy az egyszerűbb IOB címkézési konvenciót követve (l. a 2. pontot) minden szóhoz azt a címkét rendeljük, amely a szó kategóriája (szófaja) alapján a legvalószínűbb, tehát amely címke az adott kategóriájú szavak mellett a leggyakrabban figyelhető meg a tanítókorpuszban. E módszer, valamint a `hunchunk` rendszer eredménye

a 4. táblázatban látható. Tájékoztatásul feltüntetjük a címkék közti átmenetvalószínűségeket figyelmen kívül hagyó, csupán a MaxEnt-modell kimenetére támaszkodó rendszer eredményét is. Végül a táblázat azon rendszer teljesítményét is megmutatja, amelynek a tesztkorpusz kézzel készült morfológiai elemzése nem állt rendelkezésére, csupán a `hunmorph` morfológiai elemző kimenetére támaszkodhatott.

	Pontosság	Fedés	F-pontszám
baseline	60.24%	60.50%	60.37%
csak MaxEnt	88.32%	87.54%	87.93%
hunchunk	**90.58%**	**89.98%**	**90.28%**
hunchunk+hunpos	87.27%	86.32%	86.79%

4. táblázat. A `hunchunk` rendszer teljesítménye

4.3. A CoNLL feladat

Felhívjuk a figyelmet arra, hogy az NP-chunk általunk adott, a szakirodalomban legelterjedtebbtől eltérő definíciója jelentősen hosszabb és szerkezetüket tekintve komplexebb NP-ket eredményezett, mint az alap-NP-k. Ez magyarázza a szakirodalomban szokásosan láthatónál alacsonyabb pontszámokat. Noha figyelmünk középpontjában a maximális NP-k azonosítása volt, algoritmusunk teljesítményét a state-of-the-art statisztikai szegmentálóalgoritmusokéval is össze kívántuk vetni, ezért a már említett angol nyelvű CoNLL 2000 feladaton is kipróbáltuk. A CoNLL 2000 feladat tanuló- és tesztadata rögzített, ezáltal szolgálhat a különböző szegmentálóalgoritmusok összehasonlításának standard terepeként. Eszközünk 93.79%-os F-pontszámot ért el a feladaton, míg a legmagasabb publikált eredmények között szerepel például 94.34% (Sun et al. 2008) és 94.29% (Sha–Pereira 2003) is. Bár ez utóbbi eredményektől rendszerünk kb. fél százalékponttal elmarad, fontosnak tartjuk megemlíteni, hogy azoknak a komplexebb modelleknek a tanítása, amelyeknek a segítségével ezek az eredmények születtek (*Conditional Random Field, CRF,* l. Lafferty et al. 2001) akár egy nagyságrenddel hosszabb időt vesz igénybe, mint az általunk bemutatott modellezési eljárás.

5. További tervek

A 2.1. pontban már megemlítettük, hogy az általunk választott feladatnak – a maximális NP-k azonosításának – egyik célja, hogy a későbbiekben lehetőségünk

nyíljon mondatok közvetlen összetevőinek, elsősorban az igék vonzatainak feltérképezésére, aminek hasznát vehetjük gépi fordítási feladat során. Ennek első lépéseként szükséges elkészítenünk egy magyar–angol párhuzamos, azaz ugyanazon szövegeket két nyelven tartalmazó NP-korpuszt, amelyhez kiindulópontként szolgált a Hunglish magyar-angol párhuzamos korpusz (Varga et al. 2005). Megfelelő tanulóadat előállítását követően a `hunchunk` eszközt alkalmassá tettük angol nyelvű NP-felismerésre is (bővebben l. Recski et al. 2009), így elvégezhettük a Hunglish korpusz NP-chunkolását mindkét nyelven. További terveink között szerepel a korpusz NP-szintű párhuzamosítása – azaz a megfelelő NP-k összerendelése –, később egy felügyelt tanulásra épülő magyar-angol NP-fordító létrehozása. Elképzeléseink szerint az NP-fordítás képességével jelentősen közelebb kerülnénk egy jó minőségű angol-magyar fordítórendszer létrehozásához is. A bemutatott rendszert továbbá – megfelelő új tanítókorpusz létrehozásával – alkalmassá tettük tetszőleges kategóriájú maximális mondattani összetevők azonosítására (Recski 2011), ezzel újabb lépést téve a magyar mondatok felszíni szerkezetének gépi azonosítása felé.

Irodalom

Abney, Steven P. 1991. Parsing by chunks. In: Robert Berwick – Steven Abney – Carol Tenny (szerk.): Principle-based parsing. Dordrecht: Kluwer. 257–278.

Csendes, Dóra – János Csirik – Tibor Gyimóthy – András Kocsor 2005. The Szeged Treebank. Lecture Notes in Computer Science: Text, Speech and Dialogue 3658: 123–131.

Erjavec, Tomaž 2004. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In: Nicoletta Calzolari (szerk.): Fourth International Conference on Language Resources and Evaluation, LREC, volume 4. ELRA. 1535–1538.

Gee, James Paul – François Grosjean 1983. Performance structures: A psycholinguistic and linguistic appraisal. Cognitive Psychology 15: 411–458.

Halácsy, Péter – András Kornai – Csaba Oravecz 2007. HunPos – An open source trigram tagger. In: Sophia Ananiadou (szerk.): Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume. Proceedings of the Demo and Poster Sessions. Prague: Association for Computational Linguistics. 209–212.

Halácsy Péter – Kornai András – Varga Dániel 2005. Morfológiai egyértelműsítés Maximum Entrópia módszerrel. In: Alexin Zoltán – Csendes Dóra (szerk.): A III. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 180–189.

Kornai András – Rebrus Péter – Vajda Péter – Halácsy Péter – Rung András – Trón Viktor 2004. Általános célú morfológiai elemző kimeneti formalizmusa. In: Alexin Zoltán – Csendes Dóra (szerk.): A II. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 172–176.

Lafferty, John – Andrew McCallum – Fernando Pereira 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Carla E. Brodley – Andrea Pohoreckyj Danyluk (szerk.): Machine learning – International Workshop and Conference. Morgan Kaufmann. 282–289.

Marcus, Mitchell P. – Beatrice Santorini – Mary Ann Marcinkiewicz 1994. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19: 313–330.

McCallum, Andrew – Dayne Freitag – Fernando Pereira 2000. Maximum Entropy Markov Models for information extraction and segmentation. In: Pat Langley (szerk.): Proceedings of the seventeenth international conference on machine learning. Stanford: Stanford University. 591–598.

Miháltz Márton 2011. Magyar NP-felismerők összehasonlítása. In: Tanács – Vincze (2011, 333–335).

Prószéky, Gábor – László Tihanyi – Gábor Ugray 2004. Moose: A robust high-performance parser and generator. In: John Hutchins – Michael Rosner (szerk.): Proceedings of the 9th Workshop of the European Association for Machine Translation. La Valletta, Malta: Foundation for International Studies. 138–142.

Rabiner, R. Lawrence 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77: 257–286.

Ramshaw, Lance A. – Mitchell P. Marcus 1995. Text chunking using transformation-based learning. In: David Yarowsky – Kenneth Church (szerk.): Proceedings of the Third Workshop on Very Large Corpora, Cambridge, MA. Cambridge MA: MIT Press. 82–94.

Ratnaparkhi, Adwait 1998. Maximum entropy models for natural language ambiguity resolution. Doctoral dissertation, University of Pennsylvania.

Recski Gábor 2011. A sekély mondattani elemzés további lépései. In: Tanács – Vincze (2011, 113–118).

Recski Gábor – Varga Dániel – Zséder Attila – Kornai András 2009. Főévi csoportok azonosítása magyar–angol párhuzamos korpuszban. In: Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): A VI. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 3–13.

Sha, Fei – Fernando Pereira 2003. Shallow parsing with conditional random fields. In: NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Morristown, NJ: Association for Computational Linguistics. 134–141.

Sun, Xu – Louis-Philippe Morency – Daisuke Okanohara – Jun'ichi Tsujii 2008. Modeling latent-dynamic in shallow parsing: A latent conditional model with improved inference. In: COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics. Morristown, NJ: Association for Computational Linguistics. 841–848.

Tanács Attila – Vincze Veronika (szerk.) 2011. A VIII. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem.

Tjong Kim Sang, Erik F. – Sabine Buchholz 2000. Introduction to the CoNLL shared task: Chunking. In: Walter Daelemans – Miles Osborne (szerk.): Proceedings of CoNLL 2000 and LLL 2000. Association for Computational Linguistics. 127–132.

Tjong Kim Sang, Erik F. – Jorn Veenstra 1999. Representing text chunks. In: Henry S. Thompson – Alex Lascarides (szerk.): EACL. Association for Computational Linguistics. 173–179.

Uchimoto, Kiyotaka – Qing Ma – Masaki Murata – Hiromi Ozaku – Hitoshi Isahara 2000. Named entity extraction based on a maximum entropy model and transformation rules. In: Hitoshi Iida (szerk.): ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Morristown, NJ: Association for Computational Linguistics. 326–335.

Váradi, Tamás 2003. Shallow parsing of Hungarian business news. In: Dawn Archer – Paul Rayson – Andrew Wilson – Tony McEnery (szerk.): Proceedings of the Corpus Linguistics 2003 Conference, Lancaster. UCREL. 845–851.

Varga, Dániel – László Németh – Péter Halácsy – András Kornai – Viktor Trón – Viktor Nagy 2005. Parallel corpora for medium density languages. In: Nicolas Nicolov – Kalina Bontcheva – Galia Angelova – Ruslan Mitkov (szerk.): Proceedings of the Recent Advances in Natural Language Processing 2005 Conference. Amsterdam & Philadelphia: John Benjamins. 590–596.

Varga, Dániel – Eszter Simon 2007. Hungarian named entity recognition with a maximum entropy approach. Acta Cybernetica 16: 293–301.

An NP chunker for Hungarian

Abstract: We introduce the `hunchunk` tool for detecting noun phrases (NPs) in Hungarian text, a key technology in various tasks in natural language processing such as information extraction or machine translation. We employ supervised machine learning methods, training our system on the Szeged Treebank, a syntactically annotated corpus of over 1.2 million words. As common in the literature, we convert the detection task into a sequence labeling task. We then train a Maximum Entropy (ME) model using features describing the form, category and morphological features of a word and its neighbours. When tagging sentences, local decisions made by the ME model are harmonized by a language model trained on the sequence of tags in the training corpus. Our system achieves an F-score of 90.28% on a section of the Szeged Treebank reserved for evaluation purposes. The system was also evaluated on the dataset of the CoNLL 2000 Shared Task and achieved a competitive F-score of 93.79% on the task of finding base NPs in English text. The algorithm we implemented is language- and task-independent and was successfully used to handle various sequence labeling tasks such as named entity recognition, detection of English noun phrases and general chunking.

Keywords: NP-chunking, shallow parsing, supervised learning, Maximum Entropy, Hidden Markov Models

Tulajdonnevek a számítógépes nyelvészetben*

Vincze Veronika[1] – Farkas Richárd[2]

[1]*MTA-SZTE Mesterséges Intelligencia Kutatócsoport, Szeged*
[2]*Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged*
vinczev@inf.u-szeged.hu; rfarkas@inf.u-szeged.hu

A tanulmányban a tulajdonnevek számítógépes nyelvészeti kezelése során felmerülő problémákat tekintjük át. Megvitatjuk, hogy lehetséges-e nyelvi és formai kritériumok alapján meghatározni a tulajdonneveket, majd megtárgyaljuk a tulajdonnevek terjedelmével kapcsolatban felmerülő kérdéseket. Míg a nyelvészet abszolút, addig a számítógépes nyelvészet relatív tulajdonnévosztályokat állít fel, hiszen alkalmazásonként más és más kategóriák, illetve terjedelmi szabályok felállítása szükséges. A tulajdonnevek metonimikus használatának számítógépes nyelvészeti vonatkozásairól is szót ejtünk, végül a tulajdonnevek normalizálásának és gépi fordításának kérdéseit vesszük sorra.

Kulcsszavak: tulajdonnév, névelem, tulajdonnév-felismerés, osztályozás, normalizálás

1. Bevezetés

Névelem-felismerésen (angolul *Named Entity Recognition,* a továbbiakban NE-felismerés) a nyers szövegből a minket érdeklő információk, jelen esetben a tulajdonnevek és különféle azonosítók kigyűjtését és osztályokba sorolását értjük. Így a nyelvészetben is tulajdonnévnek tekintett elemek (például személynevek, földrajzi nevek, címek, márkanevek stb.) azonosításán kívül más szövegbeli entitások felismerése is beletartozik a számítógépes névelem-felismerés körébe. Utóbbiakra jellemző példák az e-mail címek, weblapok, rendszámok, telefonszámok, dátumok stb., de különféle szakszövegekben más típusú névelemek azonosítása is szükséges lehet: orvosi-biológiai szövegekben például előfordulhatnak fehérjenevek, génnevek *(citochrom-c, CD8 antigén),* kémiai témájú szövegekben a vegyületek nevének és képletének felismerése is lényeges *(nátrium-klorid, NaCl).* Bizonyos esetekben pedig szükséges lehet egy adott névelemosztály további szűkítése: egy bírósági jegyzőkönyvben nem elég azt megjelölni, hogy az adott névelem a PERSON (személy) kategóriába sorolandó, hiszen az is igen lényeges, hogy a tanúról, a felperesről, az alperesről, egy ügyvédről vagy esetleg a bíróról van-e szó.

* A tanulmány létrejöttét (részben) a MASZEKER kódnevű projekt keretében a Nemzeti Fejlesztési Ügynökség támogatta.

Ilyenkor célszerű az adott névelemosztályt további alkategóriákra bontani. A felismerendő névelemek osztálya tehát alkalmazásfüggő: mindig az adott feladat és szövegtípus határozza meg, hogy milyen típusú névelemeket keresünk.

Míg az azonosítókat általában könnyű felismerni reguláris kifejezések (minták vagy sémák) alkalmazásával (például egy magyarországi személygépkocsi rendszáma jelenleg a 3 ékezet nélküli betű + kötőjel + 3 számjegy sémára épül), addig a tulajdonnevek azonosítása számos nyelvészeti problémát rejt magában. Ennek megfelelően a továbbiakban elsődlegesen a hagyományos értelemben vett tulajdonnevek számítógépes felismerése során felmerülő problémákra összpontosítunk. Elsőként a tulajdonnevek lehetséges definícióit tekintjük át, majd a formai jellemzőikre összpontosítunk. A tulajdonnevek osztályozásának bemutatása után pedig néhány speciális számítógépes nyelvészeti problémát veszünk sorra.

2. Mi a tulajdonnév?

Ebben a szakaszban a tulajdonnevek definícióit tekintjük át, illetve megvitatjuk a problémás eseteket.

2.1. Definíciók

Hagyományos definíciók szerint a tulajdonnév „olyan nyelvi elem, amely a létezők […] bizonyos kategóriáiba tartozó azonos fajú egyedek egymástól való megkülönböztetésére és önmagukkal való azonosítására szolgál" (J. Soltész 1979, 173). *A mai magyar nyelv rendszere* című nyelvtan alapján a tulajdonnév „bizonyos élőlénynek vagy élettelen dolognak saját, külön neve. Elsősorban a más lényektől vagy dolgoktól való megkülönböztetést segíti" (MMNyR. 215). A *Magyar grammatika* című egyetemi tankönyvben olvasható: „A tulajdonnév […] szerepe az identifikáció" (MGr. 127). Kiefer (2007, 141) szerint a tulajdonnevek „szemantikailag címkék, amelyeket egyedek azonosítására használnak". Kripke a tulajdonnevet merev jelölőnek tekinti, mely konstans módon ugyanazt az egyedet azonosítja (Kripke 1980). A számítógépes nyelvészetben leggyakoribb meghatározás szerint a névelemek „a világ valamely entitására egyedi módon (unikusan) referálnak" (Simon 2008, 181).

A fenti definíciók alapján tehát a tulajdonnevek legjellemzőbb tulajdonságának az azonosítás, illetve elkülönítés tekinthető: számos hasonló létező közül választanak ki egy adott lényt vagy dolgot, azaz egyedítő funkciójuk van. Ezért olvashatjuk többek között az alábbiakat: (a tulajdonnév) „nem válhat tehát osz-

tályfogalommá (semmi más nem tartozik alá), nincs rokonértelmű párja" (Martinkó 1956, 192).

A nyelvhasználatban azonban ez a kérdés („Mi számít tulajdonnévnek?") nem annyira egyértelmű. Az alábbiakban sorra vesszük azokat a tipikus eseteket, amelyek problémákat vetnek fel a fenti elméletek és definíciók számára.

2.2. Típusjelölés

A tulajdonnevek többsége nem minden előfordulásában azonosít egyedeket, hanem használható típusjelölőként is. Nézzük az alábbi (magyar és angol) példákat!

(1) Négy **Péter** volt az 1/a. osztályban.

(2) There are two **Portlands** in the United States.
'Az Egyesült Államokban két Portland is található.'

(3) A **Fritzek** lerohanták Lengyelországot.

Az első két mondatban a tulajdonnevek nem elkülönítenek, hanem – ellenkezőleg – kiemelik a közös vonást az adott egyedekben (nevezetesen, hogy mindegyiket Péternek hívják, illetve mind a Portland nevet viselik), azaz általánosítanak. Ez pedig jellegzetesen köznévi funkció (vö. MMNy., Martinkó 1956). A köznévi jelleget az angolban az is mutatja, hogy többes számba tehető a földrajzi név (sőt – ebben a mondatban – nyelvtani okok miatt abba is kell tenni). Azt mondhatjuk tehát, hogy alkalmilag viselkedhetnek köznévszerűen a tulajdonnevek is, amikor is egy adott típust jelölnek, és ezt a típust különítik el, illetve állítják szembe az összes többi típussal (tehát az egyedítési funkció a csoportok szintjén érhető tetten).

A harmadik példa azt illusztrálja, hogy a létezők egy adott típusát gyakran egy rájuk jellemző (vagy annak vélt) tulajdonnévvel jelöljük (J. Soltész 1979), jelen esetben a *Fritz* a tipikus német katonát testesíti meg. Ez abban tér el az előző esettől, hogy míg ott a tulajdonnévi címke hozzátartozott az egyedekhez (mindegyik gyereket ténylegesen Péternek hívják, illetve a városok neve is Portland), itt az is könnyen előfordulhat, hogy egyik katonának sem Fritz a neve. A Fritz név típusra vonatkozó használata tehát egy sztereotípia nyelvi kifejeződéseként is felfogható.

Az ebbe a körbe tartozó példák egy részénél azt láthatjuk, hogy mind a kis-, mind a nagybetűs írásmód is megtalálható: a *Fritz* mellett *fritz* helyesírású alakot is látunk – utóbbi alak található meg az Osiris Kiadó széles körben használt helyesírási kézikönyvében (Laczkó–Mártonfi 2004), de 1979-ben még *Fritz*ként

szerepel (J. Soltész 1979, 110), és ugyanez érvényes például az angol rendőrök *Bobby/bobby* megnevezésére. Az általánosító funkció, a terjedő kisbetűs írásmód és a többes számú használat mind arra utalnak, hogy ezek a szavak a köznevesülés felé haladnak.

2.3. Köznevesülés és tulajdonnévvé válás

Mint az előző példából is látszik, további gondot jelent a köznév és tulajdonnév határmezsgyéjén mozgó szavak kezelése is. A köznevesülés felé tartó tulajdonnevekre a fenti példákon kívül megemlítjük az *Internet – internet* esetét, mely szó az Osiris helyesírási kézikönyve szerint kisbetűvel írandó, de néhány évvel ezelőtti szakszövegekben inkább nagybetűvel fordul elő (az 1999-es kiadású helyesírási szótár (Deme et al. 1999) íráskép szempontjából is elkülöníti az Internetet mint intézményt és az internetet mint távközlési rendszert). A tudósokról elnevezett mértékegységek is a köznevesülés klasszikus példái: *Coulomb – coulomb, Watt – watt, Tesla – tesla* stb., illetve a feltalálójukról elnevezett találmányok is ide sorolhatók: *Röntgen – röntgen*.

Bizonyos köznevek pedig tulajdonnévvé válnak (például állatok, tárgyak, cégek elnevezése során), ezért sokszor kétféle (kis- és nagybetűs) helyesírással is előfordulnak. Néhány példa:

(4) A vizsláját **Fügének** nevezte el. (kutya)

(5) Végül **Eperrel**, a piros Suzukival mentek el nyaralni. (autó)

(6) A **Pajtás** elsüllyedése volt a balatoni hajózás történetének legtöbb áldozatot követelő katasztrófája. (hajó)

A fenti tulajdonnevek természetesen közszóként kisbetűvel írandók, de ezekben az esetekben tulajdonnévnek tekintendők, így nagybetűvel írjuk őket.

2.4. Metafora és metonímia

A köznevesülés kérdésköre összefügg azzal a ténnyel, hogy a tulajdonnevek használhatók metaforikusan és metonimikusan is, ezt Kiefer kontextuális jelentéseltolódásnak nevezi (2007, 141). A metaforikus használat során az adott tulajdonnév legismertebb viselőjének egy jellegzetes vonását kapcsoljuk a névhez:

(7) Szóval tőled tudta meg? Te **Júdás**! (= áruló)

(8) Nem volt egy **Adonisz**. (= férfiideál)

Tulajdonképpen a fentiekben említett *Fritz*, illetve *Bobby* példák is a tulajdonnevek metaforikus használatra mutatnak példát, azonban meg kell említeni, hogy míg az előbbi példamondatokban szereplő nevek eredeti tulajdonosait könnyen lehet azonosítani (a bibliai, illetve a mitológiai alakként), addig a Fritz vagy Bobby megnevezés hátterében álló eredeti személyt gyakorlatilag lehetetlen megtalálni, következésképpen nem is jut eszünkbe konkrét személy a megnevezések hallatán.

A metonímia esetében referenciaátvitel történik: egy adott tulajdonnév az eredetileg jelölthöz képest más dologra vagy fogalomra utal. A metonímiának számos altípusa létezik, melyeket igen részletesen tárgyal Simon (2008), így itt most csak néhány példát mutatunk be:

(9) **Waterloo** volt az egyik legnagyobb hatással Európa XIX. századi politikai életére. (helynév = a csata, illetve annak végkimenetele)

(10) Tegnap este végig **Hemingwayt** olvasott. (személynév = regény)

(11) A **Barcelona** legyőzte a **Manchestert**. (helynév = sportegyesület)

A metonímia kérdésére még visszatérünk a névelemek osztályozásának kérdéskörе kapcsán (l. 4.3.).

2.5. Nyelvek közti eltérések

A különféle nyelvek nem egységesek abból a szempontból, hogy mi számít tulajdonnévnek. Angolul például a napok *(Monday, Saturday),* hónapok *(March, December)* és ünnepek nevei *(Christmas, Pentecost)* nagybetűvel írandók – e tulajdonságot hagyományosan a tulajdonnevek sajátjának tekintik (vö. 3., illetve az angolról Ehrlich 2000), így például a PENN Treebank annotációjában a fentieket tulajdonnévként jelölik (Marcus et al. 1993). A magyarban azonban ezek egyértelműen köznevek.

A nép-, illetve nemzetségnevek külön figyelmet érdemelnek (J. Soltész 1979). Az angolban ezek mind főnévi, mind melléknévi alakban[1] nagybetűvel írandók:

[1] Ehrlich (2000, 102) ezekre a *proper adjectives* terminust alkalmazza, és a CoNLL-2003 versenyre használt adatbázisban (Tjong Kim Sang–De Meulder 2003) is ezek az MISC (egyéb)

(12) It is widely known that **Hungarians** are very proud of their cuisine. (főnév)
'Széles körben ismert, hogy a magyarok nagyon büszkék a konyhaművészetükre.'

(13) It is widely known that **Hungarian** meals are usually spicy but tasteful. (melléknév)
'Széles körben ismert, hogy a magyar ételek általában fűszeresek, de ízletesek.'

Megjegyezzük, hogy egyes nyelveken a népekre vonatkozó főnévi és melléknévi hivatkozások felszíni jegyeikben is eltérnek, és például a franciában a melléknévi alak egyértelműen nem tulajdonnév:

(14) C'est bien connu que les **Hongrois** sont très fiers de leur cuisine.
'Széles körben ismert, hogy a magyarok nagyon büszkék a konyhaművészetükre.'

(15) C'est bien connu que les plats **hongrois** sont typiquement épicés mais savoureux.
'Széles körben ismert, hogy a magyar ételek általában fűszeresek, de ízletesek.'

A franciában a városok, illetve egyéb földrajzi területek lakosságának elnevezései (francia terminológiával a *gentilé*k) is tulajdonnévnek számítanak, részben az elnevezések látszólagos önkényessége miatt:[2]

(16) Angers – les Angevins

(17) Cahors – les Cadurciens

(18) Pont-à-Mousson – les Moussipontains

A magyarban ezekkel szemben egyik esetben sem kezeljük a fentieket tulajdonnévként. Egyetlen kivétel van: a honfoglaláskori törzsnevek *(Nyék, Megyer* stb.), melyek nagybetűvel írandók, ezért J. Soltész (1979, 79) tulajdonnévnek tekinti, és a népnevek közé sorolja be őket.

Itt említjük meg a rendszertani nevek problémáját is. Ezek latin változata kötelezően nagybetűs (pontosabban a megnevezés első tagja kezdődik nagybetűvel), míg a magyarban kisbetűvel írandók: *Canis lupus – farkas*.

Mindezek a példák azt mutatják, hogy az adott nyelvtől is függ, hogy mi esik a tulajdonnév kategóriájába és mi nem.

---

kategóriába sorolt tulajdonnevek. Más korpuszok azonban ezeket nem tekintik tulajdonnévnek: a PENN Treebankben például melléknévként szerepelnek (Marcus et al. 1993).

[2] Mind a képzők, mind a szótövek terén számos rendhagyó formával találkozhatunk, azonban bizonyos tendenciák megfigyelhetők – Eggert (2005) például a gentilék helynévből való képzésének automatizálására tesz kísérletet e tendenciák alapján.

2.6. Mi a tulajdonnév a számítógépes nyelvészet szempontjából?

A fentiek alapján levonhatjuk azt a következtetést, hogy nyelvészeti értelemben sem mindig lehetséges egyértelműen tulajdonnévnek tekinteni egy adott nagy kezdőbetűvel írt egységet a következők miatt:

- típusjelölés,
- köznevesülés és tulajdonnevesülés,
- metaforikus, illetve metonimikus használat,
- nyelvek közti eltérések.

E problémák időnként még az ember számára is megnehezítik annak eldöntését, hogy az adott egység tulajdonnév-e vagy sem. A számítógépes NE-felismerő alkalmazások számára azonban mindenképpen érdemes kidolgozni egy konkrét, jól definiált szempontrendszert, amely irányadónak tekinthető a vitás esetekben. Ennek fényében a fenti esetekre az alábbiakat javasoljuk.

Noha a tulajdonnévnek lehet típusjelölő funkciója, ettől még tulajdonnév marad – a helyesírása változatlan (ha eredetileg is nagybetűvel kezdődött, akkor típusjelölőként is azzal fog), illetve a tulajdonnévre jellemző egyedítési funkció is megmarad (a csoportok szintjén), ezért véleményünk szerint indokolható a típusjelölők tulajdonnév, azaz névelem volta. Így a számítógépes alkalmazásoktól is elvárható ezek jelölése.

A köznevesülés és tulajdonnévvé válás kérdése esetében látható, hogy a helyesírás sem teljesen mérvadó. Ha elfogadjuk, hogy elsődlegesen a nagybetűs írásmód jelzi az egység tulajdonnév voltát, akkor feltételezhetjük, hogy ha a szöveg alkotója tulajdonnévnek érezte, akkor nagybetűvel írta, ha pedig köznévnek tekinti, akkor kisbetűvel kezdte az adott szót. E szabályt követhetjük a számítógépes alkalmazások esetében is: olyan szavaknál, amelyek előfordulnak kis- és nagybetűs változatban is, csak a nagybetűseket jelöljük névelemnek.

A tulajdonnevek metaforikus használatakor a legtöbbször szintén nagybetűvel kezdődik a szó, ráadásul absztrakt síkon az azonosító funkció is jelen van (egyértelműen tudjuk azonosítani a tulajdonnév eredeti viselőjét), így az előbbiek értelmében ilyenkor is tekinthetjük tulajdonnévnek. Különféle NE-felismerők a tulajdonnevek metonimikus használatát eltérőképpen jelölhetik a normális használathoz képest (például a magyarra Simon et al. 2006, erről l. bővebben a 4.3. részben).

A nyelvek közti eltérések arra mutatnak rá, hogy míg egy NE-felismerő rendszer alapjaiban alkalmazható több nyelvre is (pl. Farkas–Szarvas 2006), bizonyos finomítások minden nyelv, illetve szövegtípus esetében szükségesnek bizonyulnak. Így például az angolban érdemes lehet külön-külön NE-kategóriát bevezetni az időt (napot, hónapot) és a nemzetiséget jelölő kifejezésekre (vö. 4.2.), utóbbi a franciában is hasznos lehet, esetleg kibővítve a városlakókkal. A latin rendszertani nevek esetében pedig szintén a külön kategória jelenthet megoldást. Mindegyik esetben az NE-felismerő jelöli, hogy névelemről (időt jelentő kifejezés, nemzetiség, latin rendszertani név) van szó. Ezt a megoldást követve tehát nem kell feltétlenül állást foglalni arról, hogy ezek a névelemek nyelvészeti értelemben véve tulajdonnevek-e, vagy pedig inkább az azonosítókra hasonlítanak (amelyek névelemek, de hagyományos értelemben nem tulajdonnevek): a rendszer egységesen névelemként kezeli őket.

3. A tulajdonnevek formai jellemzői

A legtöbb nyelvben a tulajdonnevek formai sajátosságokkal is elkülönülnek a köznevektől: a nagybetűs kezdet az esetek túlnyomó többségében a szóalak tulajdonnév voltára utal.[3] Vannak azonban olyan esetek is, amikor a nagybetűs kezdet nem jelenti azt, hogy a szóban forgó egység tulajdonnév. Egyrészről, a mondatkezdő szavak is nagybetűvel kezdődnek, mégsem mindig tulajdonnevek: ebben a mondatban például az *Egyrészről* nem tekinthető tulajdonnévnek. Másrészről, előfordulhatnak olyan köznevek (vagy jelszerű rövidítések, l. Laczkó–Mártonfi 2004) is, amelyek (csupa) nagybetűvel írandók:

(19) Karácsonyra **PDA**-t kapott a szüleitől.

(20) Az elemzők szerint a negyedik negyedévben 4,5 százalékkal eshet Magyarország **GDP**-je.

(21) A számlán szereplő végösszeg tizenötezer **Ft** volt.

Harmadrészről, nem minden tulajdonnév kezdődik nagybetűvel: kezdődhetnek számmal vagy egyéb speciális karakterrel, illetve kisbetűvel is. Ezt az alábbi példák mutatják:

[3] Ezen általánosítás alól természetesen vannak kivételek: egy jól ismert példa a német nyelv, ahol a helyesírási szabályok szerint minden főnevet nagybetűvel kell kezdeni. Emiatt a német tulajdonnév-felismerők nem építhetnek erre a formai jegyre (Rössler 2002).

(22) Az **eBay** nyereményjátékot hirdetett meg vásárlói között.

(23) A szemináriumra **ee cummings** munkásságából készült fel.

(24) Végre eljutott a **4 Non Blondes** koncertjére.

A fentiekből arra következtethetünk, hogy a nagybetűs írásmód megléte vagy hiánya önmagában véve nem árul el semmit az adott egység státuszáról: noha a nagybetűs írásmód az esetek többségében tulajdonnévre utal, általános érvényű szabálynak mégsem tekinthetjük ezt, csak tendenciának.

3.1. A tulajdonnevek terjedelme

A többtagú tulajdonnevek esetében felmerülő probléma a következő: mettől meddig tart a tulajdonnév? Laczkó–Mártonfi (2004) négy lehetőséget vázol fel a többelemű tulajdonnevek terjedelmének jelölésére nézve. E lehetőségek igen gyakran adott tulajdonnévosztályokhoz kapcsolódnak, melyek helyesírását legtöbbször az adott lehetőség szabja meg:

1. egybeírás (jellemzően földrajzi nevek)

   Budapest

2. kötőjelezés (például földrajzi nevek, kettős nevek, díjak…)

   Fehér-tó
   Kis-Kovács
   Oscar-díj

3. csupa nagybetűs kezdés (jellemzően intézménynevek, szervezetnevek, személynevek…)

   Móra Ferenc Általános Iskola
   Magyar Olimpiai Bizottság
   Horváthné Szabó Mária

4. tipográfiai eszközök (dőlt betű, ritkítás, idézőjel) használata, *nevű/című* stb. jelzők kitétele (elsősorban címek)

   A gyerekek elszavalták az *Anyám tyúkját*.
   A gyerekek elszavalták az Anyám tyúkja című verset.
   A gyerekek elszavalták az „Anyám tyúkjá”-t.

Természetesen e formai jegyek használata sem mondható abszolút érvényűnek, hiszen például a *Széchenyi fürdő* és a *Vígszínház* helyesírása eltér, noha mindkettő intézménynév, illetve egy-egy lehetőség is több tulajdonnévosztályra terjed ki (l. kötőjelezés).

Külön problémát jelent az az eset, amikor két (vagy több) azonos típusú tulajdonnév kerül egymás mellé a szövegben:

(25) **Gyurcsány Orbán** gazdaságpolitikájáról mondott véleményt.

Mivel mind a *Gyurcsány*, mind az *Orbán* személynév, elképzelhető a mondatnak olyan értelmezése is, hogy valaki egy Gyurcsány Orbán nevű személy[4] gazdaságpolitikájáról nyilvánított véleményt. A megfelelő háttérinformációk ismeretében azonban ezt az értelmezést felváltja a sokkal valószínűbb 'Gyurcsány (Ferenc) véleményt mondott Orbán (Viktor) gazdaságpolitikájáról' változat – vegyük azonban észre, hogy ehhez nem elég pusztán a mondatra hagyatkozni, a kontextusra, illetve a világtudásra is szükségünk van a legvalószínűbb értelmezés megtalálásához.

3.2. Névtartozékok

Meg kell említenünk az ún. névtartozékok problémáját is (Deme 1989). Bizonyos tulajdonneveket gyakran követ egy köznévi utótag, amely időnként azonosító, máskor magyarázó funkciójú:

(26) Kovács néni

(27) Bükk hegység

(28) Fekete-tenger

(29) Adria tenger

(30) Széchenyi tér

(31) Berlin város

(32) New York város

(33) New York városa

A (26) példában az utótag egyértelműen az azonosítást szolgálja, hiszen nélküle nem tudnánk megmondani, melyikről van szó az esetlegesen szóba jöhető Kovács nevű személyek közül. Ehhez hasonlít a *Fekete-tenger* példája is: a *Fekete*

[4] Mivel az *Orbán* keresztnév és vezetéknév is lehet, a *Gyurcsány Orbán* frázis is többértelmű: (1) vezetéknév + keresztnév; illetve (2) kettős vezetéknév. A két lehetőség közti választásban a kontextus vagy a háttértudás segíthet.

név önmagában nem jelöli ki a tengert, hiszen lehetne szó személyről *(Fekete* mint vezetéknév), hegy(ség)ről *(Fekete-hegy*, sőt *Fekete-erdő*), folyóról *(Fekete-ügy, Fekete-Körös*) stb., ezért bír nagy jelentőséggel az utótag használata is. A névtartozék és az előtag szoros viszonyát jelzi ez esetben a helyesírás is, hiszen a két elem kötőjellel kapcsolódik össze. Noha a *Széchenyi tér* esetében nincs ilyen helyesírási megszorítás, a példa mégis hasonló: névtartozék nélkül a személyre (illetve a személyről elnevezett tárgyakra, például gőzhajó) gondolnánk, illetve más utótaggal akár iskoláról, alapítványról vagy egyéb tulajdonnévosztályba tartozó elemről is lehetne szó. Így a névtartozék a tulajdonnév része (vö. Várnai 2005; Simon 2008).

Ezzel szemben a (27) és (29) példákban a köznévi utótag szerepe pusztán a magyarázat, hiszen a *Bükk* tulajdonnév önmagában is meghatározza a hegységet, illetve az *Adria* is a tengerre utal. Ilyenkor a tulajdonnév szintaktikailag minőségjelzőként szerepel, és a szerkezet feje a névtartozék (Deme 1989), így egy jelzői mellékmondattal is helyettesíthetjük a minőségjelzőt:

Adria tenger = az a tenger, amelynek Adria a neve

Ezekben az esetekben nyelvészetileg nem indokolt a névtartozékot a tulajdonnév részének tekinteni.

Fontos észrevenni, hogy a névtartozékról önmagában nem dönthető el, hogy azonosító vagy magyarázó jellegű – ez mindig csak kontextusban, tehát az adott tulajdonnév + névtartozék szókapcsolat ismeretében határozható meg. Egy ilyen jellemző utótag a *város*. Míg a *Berlin város* szókapcsolatban magyarázó jelleggel bír (hiszen *Berlin* önmagában is várost jelöl), addig a *New York város* esetében információt hordoz az utótag, hiszen lehetne szó New York államról is. Ezért utóbbi esetben a név része az utótag, míg az előbbiben nem.

Sajátos szintaktikai szerkezetben is előfordulhat a *város* utótag: *New York városa*, illetve *Róma városa*. Ez szemantikailag redundánsnak tűnik, hiszen – ha a szerkezetet elemeire bontjuk – tulajdonképpen egy város városáról beszélünk. A szerkezet használatának azonban inkább pragmatikai okai vannak, például a beszélő beszédhelyzethez (esetleg a városhoz) való viszonyulását fejezi ki. Így célszerű az egész egységet tulajdonnévként kezelni, nem pedig csak a városnevet az utótag nélkül.

3.3. Névelővel kezdődő tulajdonnevek

Bizonyos tulajdonnevek – leginkább címek, esetleg intézmények nevei – gyakran kezdődnek névelővel, például: *az MTA Nyelvtudományi Intézete* vagy *A kő-*

szívű ember fiai. Ilyenkor felmerül az a kérdés, hogy a névelő hozzátartozik-e a tulajdonnévhez vagy pedig attól külön kezelendő. A szakirodalom példái gyakran ad hoc jellegűek, bizonytalanok (vö. Várnai 2005), így egyik eljárás sem tekinthető abszolút érvényűnek.

Amennyiben a névelő vitathatatlanul a tulajdonnév része (például műcímek esetében), a névelőzhetőség problémája is felbukkan: kaphat-e és milyen névelőt a tulajdonnév? A lehetőségeket két Jókai-regény példáján szemléltetjük:

(34) *Vettem egy A kőszívű ember fiait.

(35) Vettem egy Kőszívű ember fiait.

(36) *Vettem egy Egy magyar nábobot.

(37) ?Vettem egy Magyar nábobot.

(38) ?Vettem Egy magyar nábobot.

(39) *Elfogyott az A kőszívű ember fiai.

(40) Elfogyott A kőszívű ember fiai.

(41) Elfogyott a Kőszívű ember fiai.

(42) Elfogyott az Egy magyar nábob.

Azt látjuk tehát, hogy két azonos típusú névelő nem fér össze (vö. (36) és (39)), ilyenkor az egyik (vagy a mondatbeli, vagy a tulajdonnév része) törlődik, továbbá a határozatlan névelő nem előzheti meg a határozottat (l. (34)), fordítva viszont lehetséges (l. (42)).

Külön problémát jelentenek az idegen nyelvű címek, szervezetek nevei; itt időnként törlődik az idegen névelő, máskor meg nem:

(43) Los Angelesből **az Offspring**, Glasgowból a Snow Patrol, Düsseldorfból **a Die Toten Hosen** és a világ számos pontjáról további zenekarok jelezték a napokban, hogy elfogadják a Sziget szervezők meghívását.

Érdekesség, hogy az angol zenekarnévből *(The Offspring)* törlődött az angol névelő, míg a németben *(Die Toten Hosen)* megmaradt, így voltaképpen két – egy magyar és egy német – névelő áll előtte. Nádasdy (2005) szerint a magyar névelő idegen nyelvű címek, nevek esetén sem hagyható el, de az idegen *the, die, la* stb. könnyebben elmaradhat, főleg közismert nevek esetében (például *Beatles*).

3.4. Egyeztetés

A szintaktikai egyeztetés terén is sajátosan viselkednek egyes tulajdonnevek. Többes számú, csoportot jelölő tulajdonnevek a magyarban egyes számú állítmánnyal szerepelnek:

(44) Fellépett a Vágtázó Halottkémek.

Ezzel szemben a brit angolban a formailag egyes számú, de (metonimikusan) csoportot jelölő tulajdonnevek kötelezően többes számú igével állnak (Fowler 1965):

(45) Arsenal have won ten FA Cups.
'Az Arsenal tízszer nyerte meg az FA-kupát.'

(46) Iron Maiden have sold more than 100 million albums worldwide.
'Az Iron Maiden világszerte több mint 100 millió albumot adott el.'

A névelemek automatikus azonosítása során érdemes felfigyelni erre a szintaktikai furcsaságra.

3.5. Számítógépes nyelvészeti szempontok

Az előbbiekben bemutattuk azokat a tulajdonságokat, amelyek általában jellemzik a tulajdonnevek formáját és terjedelmét. Az is egyértelmű azonban, hogy e jellemzők nem abszolút érvényűek, hiszen egyrészt nyelvenként különböző szabályok érvényesülhetnek (l. az angol és a magyar szintaktikai környezet eltérését), másrészt adott nyelven belül is számos kivételes esettel szembesülhetünk. Az automatikus NE-felismerők ezért igen sokszor statisztikai alapokra épülnek (például Farkas–Szarvas 2006; Varga–Simon 2006). A mondatkezdő szavakat így nem minősítik tulajdonnévnek, hiszen a legtöbb esetben a kisbetűs (mondatbeli) előfordulások jóval meghaladják a nagybetűs (mondatkezdő) előfordulások számát. A nagybetűvel kezdődő rövidítéseket vagy közneveket *(GDP, GPS)* pedig külön listaként szokás beépíteni a rendszerbe. A speciális karakterek jelenléte a szóban pedig szintén tulajdonnévre utal.[5]

Az NE-felismerés minél pontosabb eredménye érdekében célszerű a tulajdonnevek terjedelme problémakörében is egységes szabályokat követni, azonban

[5] Meg kell azonban említenünk, hogy manapság terjedőben van például a számok használata a számnévvel megegyező hangsorok helyettesítésére az elektronikus kommunikációban: *5let, +6ározás*, illetve angolul *sk8ing*.

mindig az adott alkalmazás függvényében célszerű az adott szabályokat lefektetni. Általában az egybeírt és kötőjelezett alakok esetén a teljes alakot tulajdonnévnek tekinthetjük. A különírt névtartozékoknál láttuk, hogy nyelvészeti szempontok alapján el lehet őket különíteni magyarázó és azonosító utótagokra; azt azonban az adott feladat ismeretében kell mérlegelni, hogy ezt a nyelvészeti megkülönböztetést mikor érdemes megtenni a számítógépes elemzés szintjén is. A névelővel kezdődő tulajdonnevekből – mint fent bemutattuk – időnként törlődik a névelő, máskor viszont a mondatbeli névelő esik ki. A következetes jelölés szempontjából ilyen esetekben legjobb a helyesírásra támaszkodni: amennyiben a névelő is nagybetűs, a tulajdonnév része lesz *(az **Egy magyar nábob***), de ha kisbetűs, akkor nem *(a **Kőszívű ember fiai***). Idegen neveknél pedig amennyiben szerepel az eredeti névelő, mindenképpen a név részének tekintendő *(a **Die Toten Hosen***), a magyar névelő viszont semmiképpen *(az **Offspring***).

Az egymást követő, azonos típusú névelemek automatikus azonosítására egy webes keresésre épülő, statisztikai módszer is hatékonynak bizonyul, itt a névelemek együttes (egymást követő) és külön-külön történő előfordulási gyakorisága játszik döntő szerepet a névelem(ek) határainak megállapításában (Farkas et al. 2008).

A szintaktikai környezet is utalhat arra, hogy a szóban forgó elem tulajdonnév: amennyiben az alany és az állítmány számbelileg nem egyezik, akkor erősen valószínű, hogy csoportot jelölő tulajdonnév az alany. Az NE-felismerők tehát a szövegen végrehajtott morfológiai és szintaktikai elemzés eredményéből is profitálhatnak.

4. A tulajdonnevek osztályozása

A nyelvészeti szakirodalmi hagyományok és a számítógépes nyelvészeti alkalmazások általában eltérő osztályokba sorolják a tulajdonneveket, ezenkívül más-más szerzők mind a számítógépes nyelvészeten, mind a nyelvészeten belül különböző számú és jellegű csoportokat állítanak fel. A továbbiakban a különféle osztályozásokat vetjük össze.

4.1. A tulajdonnevek kategóriái a nyelvészetben

A következőkben a nyelvészetben használatos tulajdonnévosztályokat ismertetjük az Osiris Kiadó helyesírási tanácsadója alapján (Laczkó–Mártonfi 2004). Az

alábbi főkategóriákat találjuk meg a szabályzatban (zárójelben tüntetjük fel a főkategóriák alá tartozó kategóriák és alkategóriák számát):

1. Személynevek (4 kategória, 6 alkategória)
2. Állatnevek és tárgynevek (2 kategória, 4 alkategória)
3. Földrajzi nevek (7 kategória, 35 alkategória)
4. Intézménynevek (2 kategória)
5. Csillagászati elnevezések (4 kategória)
6. Márkanevek (2 kategória)
7. Címek (2 kategória, 10 alkategória)
8. Kitüntetések és díjak elnevezései (2 alkategória)

Az igen részletes osztályozás 8 főkategóriát, ezeken belül 25 kategóriát és további 59 alkategóriát tartalmaz.[6] Ha az osztályozás minden szintjét fel akarjuk használni a tulajdonnevek annotációjához, akkor összesen 70 tulajdonnévtípust kellene feltételeznünk. A gyakorlatban azonban többnyire nincs szükség ilyen részletes osztályozásra, sőt, a túl részletes osztályozás igencsak megnehezíti a kategorizációt: minél több osztályunk van ugyanis, annál nagyobb a hibázás lehetősége. Ezt mutatja az is, hogy magában a kézikönyvben is találunk következetlenséget: a templomok, pályaudvarok és éttermek nevei egyrészt földrajzi (179. o.), másrészt intézménynévként (220. o.) is szerepelnek. Erről l. bővebben 4.3. alatt.

4.2. A tulajdonnevek kategóriái a számítógépes nyelvészetben

A számítógépes nyelvészeti megközelítésben több próbálkozás is született a tulajdonnevek általános és teljes kategorizálására. A legismertebb Sekine hierarchikus rendszere (Sekine et al. 2002), amely 140 tulajdonnév-kategóriát különböztet meg. Ez elsősorban a személyneveken, szervezetneveken és földrajzi neveken kívüli *egyéb* kategória tovább-bontására, az alkategóriák felsorolására koncentrál. Noha a legtöbb alkalmazás a fenti négy kategóriát használja – például a 2003-as CoNLL-verseny során is az ezekbe tartozó tulajdonnevek azonosítása volt a feladat (Tjong Kim Sang–De Meulder 2003) –, más rendszerek ezektől eltérő osztályokba sorolják a tulajdonneveket: az ACE projektben például a személynév,

[6] Vegyük észre, hogy az intézménynevek nincsenek is részletezve a szabályzatban, pusztán néhány altípusuk van megemlítve a teljesség igénye nélkül (cégnevek, iskolák nevei stb.).

szervezetnév, földrajzi-politikai név, földrajzi név és létesítménynév kategóriákat alkalmazzák (Doddington et al. 2004). A számítógépes nyelvészetben tehát nincsen elfogadott tulajdonnév-kategóriarendszer, sőt törekvés sincs egy ilyen felállítására. Az elmúlt évtized hatékonyan működő nyelvtechnológiai rendszerei a felismerendő elemosztályokat példák segítségével írják le. Már Sekine sem kísérelte meg 140 kategóriájának szemantikai és szintaktikai definiálását, csak tipikus példákat sorolt fel.

A számítógépes nyelvészet ún. felügyelt gépi tanulási (példaalapú) megközelítésben (Manning–Schütze 1999) minden egyes feladathoz rendelkezésre áll egy tanító korpusz (magyar nyelvre jelenleg egyetlen tulajdonnévkorpusz létezik, a SzegedNE korpusz[7] (Szarvas et al. 2006), amelyben a tulajdonnevekre referáló frázisokat szakértők – kategóriába sorolva – bejelölték. A cél ennek felhasználásával olyan gépi modell megtanulása, amely ismeretlen szövegen is hatékonyan felismeri az adott kategóriákat. Fontos megjegyeznünk, hogy ez a tanult modell csak a tanítóhalmazzal megegyező tulajdonságú szövegeken működik pontosan. Azaz ebben a megközelítésben a tulajdonnév-kategóriák definíciói a korpuszépítést végző (és az építéshez készített annotálási útmutatót megfogalmazó) szakértők fejében implicite vannak jelen, az automatikus azonosítást végrehajtó módszerek csak a példákon keresztül ismerik azt meg.

A számítógépes nyelvészeti megközelítés egy másik jellegzetessége, hogy nem célozza meg egy egységes kategóriarendszer és szabályrendszer megalkotását, minden feladathoz elkészül egy korpusz. Ezek különböző irányelveket követhetnek, sőt különböző típusú szakszövegek esetén különbözniük is kell. A legkutatottabb szakszövegek az újsághírek (általában gazdasági, politikai és sport) és az utóbbi években a biológiai szövegek – ahol a cél gének, fehérjenevek felismerése (Ohta et al. 2004). Az orvosi szövegekben érdekességként jelentkezik, hogy itt meg kell különböztetni a személyneveken belül az orvosok és páciensek nevét (Uzuner et al. 2007). Amellett, hogy az egyes szakterületeken jelentősen eltérhet a felismerendő tulajdonnév-kategóriák köre, azok szövegei eltérő jellegzetességekkel bírnak, ami a szövegkörnyezetet kiaknázó automatikus módszerek feladatát megnehezíti (Farkas–Szarvas 2006).

A nyelvészeti és a számítógépes nyelvészeti osztályozás összevetéséből kiderül, hogy míg a nyelvészeti kategóriarendszer abszolút (azaz nem változik szövegről szövegre), a számítógépes nyelvészeti kategorizálás relatív: mindig az adott alkalmazástól függ a kategóriarendszer, és más feladat szövegeire nem is lenne feltétlenül adaptálható.

[7] http://www.inf.u-szeged.hu/rgai/corpus_ne

4.3. A tulajdonnevek metonimikus használata

Az osztályozás körében felmerül a tulajdonnevek metonimikus használatának problémája is. Bizonyos tulajdonnevek a jelentésátvitelnek (metonímiának) köszönhetően többjelentésűnek tűnnek – hol helyet, hol eseményt, hol szervezetet stb. jelöl ugyanaz a névelem. Simon (2008) részletesen ismerteti a tulajdonnevek metonimikus használatának számos esetét, így mi pusztán néhány példával illusztráljuk a jelenséget:

(47) Elutazott Pekingbe. (hely)

(48) Peking után rögtön összeült a MOB. (esemény)

(49) Peking hírzárlatot rendelt el. (szervezet)

Mind a nyelvészeti, mind a számítógépes osztályozás szempontjából igen lényeges eldönteni, hogy az aktuális kontextus szerint osztályozzuk a metonimikus viselkedést mutató névelemet *(Peking* háromféle osztálycímkével rendelkezik), vagy állandó jelleggel egy osztályba soroljuk őket kontextustól függetlenül (tehát *Peking* hely lesz a fenti példák mindegyikében). A számítógépes nyelvészetben mindkét módszert alkalmazzák: előbbi *tag-for-meaning* (kontextusnak megfelelő), utóbbi *tag-for-tag* (állandó) annotáció néven ismert. A tag-for-meaning annotáció nyelvileg pontosabb ugyan, számítógépes szempontból viszont nehezebben kezelhető, hiszen nemcsak magát a névelemet, de a kontextust is figyelembe kell venni a megfelelő osztálycímke megtalálásához.

Bizonyos típusú tulajdonnevek metonimikus használata morfológiai és szintaktikai változásokkal jár együtt. Egyrészt, noha a tulajdonnevek általában nem tehetők többes számba, a többes számú szervezetnevek gyakran márkanévre utalnak (l. alább), hasonlóan a típusjelöléshez (vö. 2.2.). Másrészt, a városnevek nem névelőzhetők (a magyarban legalábbis), így amennyiben a szövegben egy névelő+városnév képletet találunk, az valójában gyakran az adott város egy sportegyesületét jelenti, és nem az eredeti földrajzi helyet jelöli:

(50) A Manchester vereséget szenvedett a Barcelonától.

Ezekben az esetekben az eredetileg földrajzi névként használatos frázis szervezetre utal. A tulajdonnevek felismerését és azok osztályokba sorolását tehát a morfológiai és szintaktikai elemzés is segítheti.

5. Speciális számítógépes nyelvészeti problémák

A következőkben néhány speciális számítógépes nyelvészeti problémát veszünk sorra: a tulajdonnevek normalizációja, illetve gépi fordítása során felmerülő kérdéseket tekintjük át.

5.1. Normalizáció

A tulajdonnévként viselkedő frázisok szövegbeli azonosításán és szemantikai kategorizálásán felül a számítógépes nyelvészeti rendszereknek meg kell oldaniuk a felismert tulajdonnevek normalizációját is. Normalizáción a tulajdonnevek szótári alakjának meghatározását, illetve annak egyedi azonosítóra való leképezését értjük.

A tulajdonnevek morfológiai elemzése nyelvtechnológiai szempontból igen fontos. Egyrészt a keresés (például információ-visszakeresés) és tárolás miatt ragozatlan formában kell eltárolni, másrészt a szótári alak felhasználása segíthet a tulajdonnév szemantikai kategóriájának meghatározásában, sőt a ragok azonosítása is hasznos lehet például metonimikus jelentések felismerésénél. A következő példában a többes szám felismerése implikálhatja a tulajdonnév termékre vonatkozó szervezetnév voltát:

(51) A Volvó**k** a legmegbízhatóbb autók.

A tulajdonnevek szótári alakjának gépi azonosítása nehéz feladat. A köznevek automatikus szótövezése általában felhasznál egy listát, amely tartalmazza az adott szófaj összes lehetséges szótövét. A tulajdonnevek esetében ez nem kivitelezhető, hiszen ezek nyílt osztályt képeznek. A szótövezés feladatkörébe tartozik annak eldöntése is, hogy a felismert frázis szótári alakban szerepel-e már, például a *Pannon* szóalaknál a *Pann*-t, a *Philips*-nél a *Philip*-et gondolhatnánk szótőnek (az *-on* és *-s* toldalékokkal), ha az adott márkanevek nem lennének közismertek.

A lemmatizálás gépi megoldásai két különböző megközelítést követhetnek. Vagy a szuffixumok és a hangrend vizsgálata felől közelítenek (Trón et al. 2005), vagy korpuszstatisztikákra építenek. Ez utóbbi esetben az a hipotézis, hogy elég nagy korpuszt vizsgálva egy tulajdonnév ragozatlan alakjának gyakorisága szignifikánsan nagyobb, mint bármely ragozott alakjáé (Farkas et al. 2008).

A normalizáció másik feladata a tulajdonnevek és egyedek összerendelése. Egy tulajdonnév nem feltétlenül hivatkozik egyértelműen egy egyedre, hiszen például több *Kovács János* is létezhet (poliszémia). A személynevek egyértelműsítése napjaink nyelvtechnológiájának egyik intenzíven kutatott területe (Artiles

et al. 2009), de emellett például a biomedikai célú szövegfeldolgozásban is kiemelt szerepe van. Chen et al. (2005) empirikusan megmutatta, hogy a fehérjenevek előfordulásainak 13 százaléka többértelmű. A probléma megoldásához először az egyedek halmazának azonosítása szükséges, amelyekre a szóban forgó tulajdonnév hivatkozhat (jelentések klaszterezése), majd a létező gépi megoldások (Farkas 2008) az egyes jelentések közti döntést a szövegkörnyezet elemzése révén hozzák meg.

A tulajdonnév-egyed összerendelést a másik irányból vizsgálva, ugyanazon egyedre több tulajdonnévvel is hivatkozhatunk (szinonímia), például *Ferencváros = Fradi = FTC*. A szinonimák egy speciális esete az akronimával történő hivatkozás, ami elsősorban szervezetneveknél előforduló jelenség (például *Magyar Nemzeti Bank (MNB)*), azonban más kategóriáknál is megjelenhet (például *Kovács J.*). A normalizációnak a szinonimák problémáját is fel kell oldania, hiszen a számítógépes nyelvészet egyik legfontosabb alkalmazásának (az információkinyerésnek) célja a szövegből nyerhető információ összegyűjtése adott egyedekről – és nem tulajdonnevekről. A feladat megoldásához a gépi rendszerek szinonimalistákat gyűjtenek nagy korpuszokból egyszerű mintázatok felismerésével (például „C-vitamin, *vagy más néven* aszkorbinsav").

5.2. Gépi fordítás

A tulajdonnevek (gépi) fordíthatóságának kérdése (Li et al. 2009) is érdekes vizsgálati terület. A tulajdonnevek egy részét le kell (illetve le szokás) fordítani, ilyen a legtöbb intézménynév és földrajzi név stb. Néhány példa:

(52) United Nations Organization – Egyesült Nemzetek Szervezete

(53) Bécs – Wien – Vienna

Ezzel szemben más osztályok különleges bánásmódot igényelnek: a személyneveket nem fordítjuk: *Kovács János*ból nem lesz *Johann Schmidt* egy német nyelvű szövegben. Ez alól kivételt jelentenek az uralkodók, pápák stb. nevei, melyeknek léteznek bevett magyar megfelelőik *(János Károly, II. János Pál),* sőt a XIX. században az írók, gondolkodók stb. neveit is magyarították *(Verne Gyula, Marx Károly).* A személynevekkel kapcsolatban fontos megemlíteni, hogy a magyar sorrendben írt nevet más nyelvekben meg kell fordítani *(János Kovács),* erre tehát külön fel kell készíteni a fordítóprogramot.

Érdekes a címek esete: az irodalmi, képzőművészeti alkotások és bizonyos zeneművek címeit lefordítjuk:

(54) As You Like It – Ahogy tetszik

(55) Dama con l'ermellino – Hermelines hölgy

(56) Le quattro stagioni – A négy évszak

Ezzel szemben az újságcímeket és a zeneszámok címeit a mai gyakorlat szerint nem:

(57) The Times – ??Az Idők

(58) November Rain – ??Novemberi eső

Hasznosnak tűnik tehát a fordítóprogramokba beépíteni a fordítandó tulajdonnevek idegen nyelvű megfelelőit, a nem fordítandók esetében pedig külön szabályokkal biztosítani a különleges kezelést.

6. Összegzés

A tanulmányban a tulajdonnevek számítógépes nyelvészeti kezelése során felmerülő problémákat tekintettük át. Megvitattuk, hogy lehetséges-e nyelvi és formai kritériumok alapján meghatározni a tulajdonneveket, megtárgyaltuk a tulajdonnevek terjedelmével kapcsolatban felmerülő kérdéseket. A tulajdonnevek nyelvészeti és számítógépes osztályozásával kapcsolatban kiderült, hogy míg a nyelvészet abszolút, addig a számítógépes nyelvészet relatív kategóriákat állít fel, hiszen alkalmazásonként más és más osztályok, illetve terjedelmi szabályok felállítása szükséges. A tulajdonnevek metonimikus használatának számítógépes nyelvészeti vonatkozásairól is szót ejtettünk, majd a tulajdonnevek normalizálásának és gépi fordításának kérdéseit vettük sorra. A nyelvészeti szempontok számítógépes nyelvészeti elvárásokhoz hangolása, majd azok beépítése a számítógépes NE-alkalmazásokba igen hasznosnak bizonyulhat, mivel azok pontosabb működéséhez vezet, így a tulajdonnevek számítógépes nyelvészeti kezelése is könnyebbé válhat.

Irodalom

Alexin Zoltán – Csendes Dóra (szerk.) 2006. A IV. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem.

Artiles, Javier – Julio Gonzalo – Satoshi Sekine 2009. WePS 2 Evaluation campaign: Overview of the web people search clustering task. In: Javier Artiles – Julio Gonzalo – Satoshi Sekine (szerk.): Proceedings of the 2nd Web People Search Evaluation Workshop at the 18th WWW Conference. Madrid: Association for Computing Machinery. 1–9.

Chen, Lifeng – Hongfang Liu – Carol Friedman 2005. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics 21: 248–256.

Deme László 1989. Névterjedelem és névtartozékok. In: Balogh Lajos – Ördög Ferenc (szerk.): Névtudomány és művelődéstörténet. Zalaegerszeg: Zalaegerszeg Város Tanácsa. 282–286.

Deme László – Fábián Pál – Tóth Etelka 1999. Magyar helyesírási szótár. Budapest: Akadémiai Kiadó.

Doddington, George – Alexis Mitchell – Mark Przybocki – Lance Ramshaw – Stephanie Strassel – Ralph Weischedel 2004. The Automatic Content Extraction (ACE) Program – Tasks, data and evaluation. In: Maria Teresa Lino – Maria Francisca Xavier – Fátima Ferreira – Rute Costa – Raquel Silva (szerk.): Proceedings of LREC 2004. Lisbon: ELRA. 837–840.

Eggert, Elmar 2005. Bisontins ou Besançonnais? A la recherche des règles pour la formation des gentilés pour une application au traitement automatique. Tübingen: Günter Narr Verlag.

Ehrlich, Eugene H. 2000. Schaum's outline of theory and problems of English grammar. New York: McGraw-Hill.

Farkas, Richárd 2008. The strength of co-authorship in gene name disambiguation. BMC Bioinformatics 9: 69.

Farkas Richárd – Szarvas György 2006. Nyelvfüggetlen tulajdonnév-felismerő rendszer és akalmazása különböző domainekre. In: Alexin – Csendes (2006, 22–31).

Farkas, Richárd – Veronika Vincze – István Nagy – Róbert Ormándi – György Szarvas – Attila Almási 2008. Web-based lemmatisation of named entities. In: Aleš Horák – Ivan Kopeček – Karel Pala – Petr Sojka (szerk.): Proceedings of the 11th International Conference on Text, Speech and Dialogue. Berlin/Heidelberg: Springer Verlag. 53–60.

Fowler, Henry Watson 1965. A dictionary of modern English usage (Second ed.). Oxford: Oxford University Press.

J. Soltész Katalin 1979. A tulajdonnév funkciója és jelentése. Budapest: Akadémiai Kiadó.

Kiefer Ferenc 2007. Jelentéselmélet. Budapest: Corvina.

Kripke, Saul 1980. Naming and necessity. Cambridge MA & Oxford: Blackwell.

Laczkó Krisztina – Mártonfi Attila 2004. Helyesírás. Budapest: Osiris Kiadó.

Li, Haizhou – A. Kumaran – Vladimir Pervouchine – Min Zhang 2009. Report of NEWS 2009 Machine Transliteration Shared Task. In: Haizhiu Li – A. Kumaran (szerk.): Proceedings of the 2009 Named Entities ACL Workshop. Singapore: Association for Computational Linguistics. 1–18.

Manning, Christopher D. – Hinrich Schütze 1999. Foundations of statistical natural language processing. Cambridge MA: MIT Press.

Marcus, Mitchell P. – Beatrice Santorini – Mary Ann Marcinkiewicz 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19: 313–330.

Martinkó András 1956. A tulajdonnév jelentéstanához. In: Bárczi Géza – Benkő Loránd (szerk.): Emlékkönyv Pais Dezső hetvenedik születésnapjára. Budapest: Akadémiai Kiadó. 189–195.

MGr. = Keszler Borbála (szerk.) 2000. Magyar grammatika. Budapest: Nemzeti Tankönyvkiadó.

MMNy. = Bencédy József – Fábián Pál – Rácz Endre – Velcsov Mártonné (szerk.) 1968. A mai magyar nyelv. Budapest: Tankönyvkiadó.

MMNyR. = Tompa, József 1961/1962. A mai magyar nyelv rendszere I–II. Budapest: Akadémiai Kiadó.

Nádasdy Ádám 2005. A The Game. Magyar Narancs 2005.09.15.

Ohta, Tomoko – Yoshimasa Tsuruoka – Yuka Tateisi 2004. Introduction to the BioEntity recognition task at JNLPBA. In: Nigel Collier – Patrick Ruch – Adeline Nazarenko (szerk.): Proceedings of International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Genova: Association for Computational Linguistics. 70–75.

Rössler, Marc 2002. Using Markov models for named entity recognition in German newspapers. In: Erhard W. Hinrichs – Sandra Kübler (szerk.): Proceedings of the Workshop on Machine Learning Approaches in Computational Linguistics. Trento: ESSLII. 29–37.

Sekine, Satoshi – Kiyoshi Sudo – Chikashi Nobata 2002. Extended named entity hierarchy. In: M. Gonzáles Rodriguez – C. Paz Suárez Araujo (szerk.): Proceedings of Third International Conference on Language Resources and Evaluation (LREC'02). Canary Islands: ELRA. 1818–1824.

Simon Eszter 2008. Nyelvészeti problémák a tulajdonnév-felismerés területén. In: Sinkovics Balázs (szerk.): LingDok 7. Nyelvész-doktoranduszok dolgozatai. Szeged: JATEPress. 181–196.

Simon Eszter – Farkas Richárd – Halácsy Péter – Sass Bálint – Szarvas György – Varga Dániel 2006. A HunNER Korpusz. In: Alexin – Csendes (2006, 373–376).

Szarvas, György – Richárd Farkas – László Felföldi – András Kocsor – János Csirik 2006. A highly accurate Named Entity corpus for Hungarian. In: Nicoletta Calzolari – Khalid Choukri – Aldo Gangemi – Bente Maegaard – Joseph Mariani – Jan Odijk (szerk.): Proceedings of LREC 2006. Genoa: ELRA. 1957–1960.

Tjong Kim Sang, Erik F. – Fien De Meulder 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Walter Daelemans – Miles Osborne (szerk.): Proceedings of CoNLL-2003. Edmonton: Association for Computational Linguistics. 142–147.

Trón, Viktor – László Németh – Péter Halácsy – András Kornai – György Gyepesi – Dániel Varga 2005. Hunmorph: Open source word analysis. In: Martin Jansche (szerk.): Proceedings of the ACL Workshop on Software. Stroudsburg, PA: Associaton for Computational Linguistics. 77–85.

Uzuner, Özlem – Yuan Luo – Peter Szolovits 2007. Evaluating the state-of-the-art in automatic de-identification. Journal of American Medical Informatics Association 14: 550–563.

Varga Dániel – Simon Eszter 2006. Magyar nyelvű tulajdonnév-felismerés maximum entrópia módszerrel. In: Alexin – Csendes (2006, 32–38).

Várnai Judit Szilvia 2005. Bárhogy nevezzük… A tulajdonnév a nyelvben és a nyelvészetben. Budapest: Tinta Könyvkiadó.

Proper nouns in natural language processing

Abstract: In this paper we give an overview of the problematic issues concerning the natural language processing of proper nouns. We examine whether it is possible to give a definition of proper nouns on the basis of linguistic and formal criteria, then we discuss what is considered to be a part of a proper noun. While linguistic categories of proper nouns are absolute, natural language processing works with relative categories since different categories or different ways of determining the boundaries of a proper noun may be necessary for different applications. We also discuss the metonymic usage of proper nouns from the aspect of natural language processing and finally we present the questions of normalization and machine translation of proper nouns.

Keywords: proper noun, named entity, named entity recognition, classification, normalization

Analógiás tanulás asszociatív memóriamodellel

Kálmán László

Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Budapest
kalman@nytud.hu

Az írás három nagyobb alpontból áll. Az elsőben röviden összefoglalom a nyelvtudomány történetének azt a részét, amely a generativista nyelvmodelltől eltérő modelleket dolgozott ki vagy feltételezett, vagyis azt, amely a nyelvet nem a „jólformált mondatok halmazával" azonosítja. A második pontban konkrét javaslatot teszek arra, hogy milyen nyelvmodellek bontakoznak ki ezekből az egymástól is eléggé eltérő nem generativista felfogásokból. Az alternatív nyelvmodellt „memóriaalapúnak" nevezem, mert a lényege az, hogy a nyelvet emléknyomok és a köztük levő asszociációk és gátlások összességének tekinti. Végül a harmadik pontban leírom azt a kísérleti számítógépes rendszert, amelynek segítségével egy memóriaalapú modell működésének legfőbb aspektusait modellálni tudtam, bemutatva azokat a kísérleti eredményeket, amelyeket a segítségével elértem.

Kulcsszavak: számítógépes szimuláció, memóriaalapú modellek, nem generatív nyelvészet, analógia, nyelvelmélet

1. Bevezetés

A számítógépes nyelvészet általában a saját útját járta, csak néhány vonatkozásban hatott rá komolyabban az általános nyelvészet. Ilyen például az amerikai deskriptív iskola és az orosz strukturalizmus disztribúciós elemzése – a számítógépes nyelvészet máig az ezen alapuló módszert használja az elemek viselkedésének jellemzésére, osztályaik felfedezésére. A generatív nyelvészet is csak rövid időre és csak viszonylag felszínesen hatott a számítógépes nyelvészetre (például a kontextusfüggetlen nyelvtanok levezetési fáit, szerkezeti ábrázolásait máig is használják).

Érdekes kontrasztban áll ezzel az a tény, hogy a hasonlóságon (analógián) alapuló nyelvszemléletnek, amely legalább a 19. századig, de egyes értelmezések szerint az ókorig nyúlik vissza (l. pl. Esper 1973), a számítógépes nyelvészet megszületése óta kitapintható hatása van a számítógépes nyelvészekre, beleértve az említett deskriptivistákat is. A 20. század végétől kezdve azonban az ún. esetalapú, példányalapú és más analógián alapuló algoritmusok szinte a főáramává váltak a számítógépes nyelvészetnek.

Az alábbiakban az elméleti nyelvészet legutóbbi évtizedekben látott fejlődését foglalom össze ebből a szempontból (l. 2.). Felfogásom szerint ezek a fejlemények nem forradalmian újak, hanem inkább régóta létező és fel-felbukkanó

gondolatok újrafogalmazásai, egyúttal új sikerei is. A következő, 3. pontban felvázolom azt az elméleti nyelvmodellt, amelyet mindezekből az előzményekből kibontakozni vélek, és amelynek lényege, hogy a nyelvi viselkedés modellálása elsősorban az **emlékezeti** mechanizmusok modellálásán alapul, ezért a vizsgált felfogást **memóriaalapúnak** fogom nevezni. Végül a 4. pontban a modell egy gyakorlati, számítógépes megvalósításáról számolok be, annak illusztrálására, hogy a közeli jövőben ilyen természetű rendszerek gyakorlati nyelvi szimulációs feladatok megoldására is alkalmasak lehetnek.

2. Előzmények

Az 1950-es évektől, a generatív nyelvtan térhódítása nyomán szinte egyeduralkodóvá vált a **nyelvnek** az a szemlélete, hogy az nem más, mint a „jólformált" („grammatikus") mondatok halmaza (akár a jelentésükkel együtt, akár azoktól függetlenül). Ennek megfelelően a **nyelvtan** olyan szabályrendszert kezdett jelenteni, amely ezt a halmazt kimerítően jellemzi.

Ezt a felfogást megelőzően a nyelvről való gondolkodásban jelen volt egy másik szemlélet is, amely a 20. század közepén visszaszorult, majd a 20. század végétől kezdve ismét fel-felbukkant. E szerint a felfogás szerint az, hogy a beszélők bizonyos megnyilatkozásokat bizonyos helyzetekben elfogadhatónak („jólformáltnak") fogadnak el, másokat pedig nem, epifenomenálisnak tekintendő. Igaz, mindenki nagyjából meg tudja ítélni, hogy a beszélő az ő anyanyelvén szándékozott-e egy megnyilatkozást kimondani, és arról is tud véleményt alkotni, hogy ez többé-kevésbé sikerült-e neki, de ez csak következménye, nem pedig a lényege anyanyelvi ismereteinek. Még az is lehet, hogy az anyanyelvi ismeretünkre valamennyire vissza lehet következtetni ezekből az ítéletekből, de ez közel sem jelenti azt, hogy az elfogadhatónak ítélt mondathalmaz alkalmas meghatározása egy-egy emberi nyelvnek.

Akkor mit is érthetünk ebben a másik szemléletben **nyelven**? Sokféle alternatív meghatározást találhatunk a nyelvfilozófia, a pszichológia és a nyelvészet történetében, és ezek legtöbbje egyáltalán nem elég pontos ahhoz, hogy tudományos, sőt természettudományos értelemben vett modellezés alapjául szolgáljon. Abban a hagyományban, amelyet itt némileg önkényesen emelek ki a tudománytörténet hosszú folyamából, pontosabban a saját szemszögemből vélek abból kibontakozni, a nyelv egyénenként változó, de nagyon sok vonásában mégis hasonló **emlékekből** áll: azokból az emlékekből, amelyek a korábbi nyelvhasználatunk nyomait őrzik (még ha nem felidézhető formában is), a használt formákat, alakokat és a használat körülményeit is beleértve. (Természetesen nem bármi-

lyen körülmény egyformán fontos: bár jelentéktelen részletek is nyomot hagynak bennünk, nyilván vannak köztük olyanok, amelyek tartalmasabb módon összekapcsolódnak azzal, hogy az emlék keletkezésekor éppen mi hangzott el.)

A nyelvnek ezt a felfogását a legváltozatosabb formákban találjuk meg a szakirodalomban, és különböző szerzők különböző aspektusokat hangsúlyoznak. Hogy csak néhány példát emeljek ki, Locke (1700), Paul (1880) vagy Saussure (1916) nyelvfelfogása is nagyjából beleillik abba az általános jellemzésbe, amelyet fent adtam. A modern, 20–21. századi nyelvészetben a konstrukciós nyelvészek (Fillmore 1988; Goldberg 1997; Kay 1995; Kay–Fillmore 1999; Sag 1997; 2001), a kognitív nyelvtan képviselői (Langacker 1987; 1990; 1991; 2008; Taylor 2002), az analógiás és evolúciós nyelvészek többsége (Bybee et al. 1994; Bybee 2001; 2005; 2006; Blevins 2004; 2006; Blevins–Garrett 2008; Blevins–Blevins 2009), a számítógépes nyelvészetben az esetalapú és példányalapú rendszerek (Schank 1982; Kolodner 1983; 1993; Lebowitz 1983; Skousen 1989; 1992; Schank et al. 1994; Aamodt–Plaza 1994; Skousen et al. 2002) is nagyjából ilyen meghatározást tételeznek fel, bár a részletekben persze sok eltérés van köztük. A továbbiakban nem foglalkozom azzal, hogy melyik megközelítés mely tényezőket tart fontosabbnak, és ennek megfelelően milyen elméleti keretet javasol. Ehelyett a fent leírt általános meghatározás szellemében fogom ismertetni először elméleti, majd gyakorlati szempontból az általam javasolt modellt. Nehéz lenne nevet adni ennek a nyelvfelfogásnak, mert a legtöbb ilyen értelmű elnevezés már foglalt (ilyen elnevezések pl. a *pszichologista, mentális, kognitív, strukturalista, memóriaalapú*). Talán a *memóriaalapú* a legkevésbé elméletileg terhelt, bár a mesterséges tanulási algoritmusok egyik családját is így hívják (l. Russell–Norvig 1995).

3. Az elméleti modell

Akár sajátos agyi alapjai vannak a nyelvi viselkedésnek, akár más kognitív képességekből következik a nyelvhasználatra való képesség, a memóriaalapú felfogásban azon alapul, hogy mind a beszéd, mind a megértés során a nyelvi emlékeinket úgy idézzük fel és kombináljuk újra, hogy ennek segítségével korábbi nyelvi tapasztalatainkhoz a lehető leghasonlóbban tudjunk viselkedni. Ezért hívják ezt a megközelítést **analógiásnak** is: a kölcsönös megértés – már amennyire lehetséges – azon alapul, hogy hasonló dolgokat hasonló eszközökkel fejezünk ki, és ezt feltételezzük beszélőtársainkról is. Ezért minden megnyilatkozásunkat korábbiak hasonlatosságára, analógiájára alkotjuk meg, és ugyanilyen elven értelmezik őket üzenetünk címzettei.

A kulcskérdés természetesen az, hogy mik is a nyelvi természetű emlékek felidézésének és újrakombinálásának törvényszerűségei és korlátai, tudatos és önkéntelen módszerei, stratégiái. Egyelőre szinte semmit sem tudunk arról, milyen nyomokat hagynak a tapasztalatok az agyban, tehát arról, hogy miben is áll pontosan egy emlék(nyom). Ugyanakkor rengeteg mindennapi és kísérleti tapasztalat áll a rendelkezésünkre ahhoz, hogy az emlékezetnek néhány olyan vonását feltételezhessük, amelyek a nyelv szempontjából fontosak. Ezeknek a megfigyeléseknek a nagy része réges-régen ismert, és több-kevesebb súllyal általában szerepelnek mindazokban a művekben, amelyekre előzményekként hivatkoztam.

3.1. Az emléknyomok erőssége

Régi megállapítás, hogy annál élesebben emlékszünk valamire, illetve annál mélyebb, nagyobb hatású nyomot hagy, minél **gyakrabban** tapasztaltuk, vagy minél **fontosabbnak** ítéljük. A gyakoriságba nemcsak az érzékszervek által való tapasztalás számít bele, hanem az **endogén,** vagyis belső, asszociációkon keresztül való aktiváció is, tehát az, amikor más emléknyomok felidézéséről jut eszünkbe az illető emlék. (A fontosság éppen ezért bizonyos értelemben szintén összefügg a gyakorisággal: azt szokták fontosnak tekinteni, ami kellemes dolgok eléréséhez vagy kellemetlenek elkerüléséhez kapcsolódik, és éppen emiatt az ilyenek viszonylag gyakran foglalkoztatják az embert.)

Nyelvi szempontból is nagyon fontos a tapasztalatok gyakorisága: nemcsak egyénileg (a gyakrabban hallott és használt, a fontosabb szavakat és jelentéseket könnyebben hívjuk elő), hanem az egész beszélőközösség szempontjából is, hiszen a nyelvet igen gyakran használjuk kommunikációra, ezért a nyelvi elemek, eszközök gyakorisága nagy mértékben hasonlít a különböző beszélők esetében. Mivel az új kifejezéseket a régebben tapasztaltak hasonlatosságára alkotjuk és értjük meg, a gyakoribb szavak, szerkezetek, nyelvi tulajdonságok nagyobb hatást gyakorolnak, nagyobb szerepük van az alkotási és megértési folyamatokban, nagyobb az analógiás vonzóerejük. Ezt figyelték meg és modellálták például Daelemans et al. (1997); Daelemans–van den Bosch (2005), amikor a holland kicsinyítő alakok képzéséről próbáltak számot adni. Ha az X alak kicsinyítő formáját még nem hallottuk (vagy nem emlékszünk rá), akkor a sok lehetőség közül azt az X' alakot fogjuk a legnagyobb valószínűséggel választani, amelyik a legjobban hasonlít a leggyakoribb Y' kicsinyítő alakokhoz, feltéve, hogy az Y' nem kicsinyítő Y alakja a lehető legjobban hasonlít X-hez. Ez a bonyolult optimalizálási eljárás minden nyelvi viselkedés alapja.

3.2. Általánosítások

Mint az előző alpontból is kiderült, a **hasonlóság** fogalma az egyik legfontosabb kelléke minden memóriaalapú felfogásnak. Gondolkodásunkban minden bizonnyal hatalmas szerepet játszik, hogy szoros kapcsolatban vannak egymással azok az emlékeink, amelyek valamiben hasonlítanak egymásra. Ez a képzettársításnak, az **asszociációnak** az egyik legfontosabb fajtája (a nyelvre vonatkozóan ezt hívta Saussure **paradigmatikus** kapcsolatnak). A hasonlóság nem más, mint bizonyos tulajdonságok közös volta: *X* és *Y* abban és csak abban hasonlítanak egymásra, ami a közös tulajdonságuk. Vagyis a hasonlóságok azonosítása feltételezi az **általánosítást,** a részletektől való eltekintést, elvonatkoztatást: *Z* akkor és csak akkor hasonló vonása *X*-nek és *Y*-nak, ha mind *X*-nél, mind *Y*-nál általánosabb (vagyis ha *Z* mind *X*-ből, mind *Y*-ból megkapható úgy, hogy bizonyos tulajdonságaiktól eltekintünk).

Az általánosítások a tapasztalatokból automatikusan bontakoznak ki (manapság úgy is mondják ezt, hogy **emergálnak**), vagyis alapvető mechanizmusa az érzékelésnek és az emlékezetnek, hogy minden új tapasztalatot összevetünk korábbi emlékekeinkkel, és megtaláljuk a hasonlóságokat. A hasonlóságok ilyen módon ugyanúgy tapasztalatnak és emléknyomnak minősülnek, mint a közvetlenül az érzékelésből származók, csak éppen endogén eredetűek. És ugyanúgy, mint azok, elhomályosulnak, ha nem érnek el egy bizonyos gyakoriságot.

Az általánosításhoz hasonló spontán folyamat az, hogy tulajdonságokról eszünkbe jutnak olyan konkrétabb emlékek, amelyekben az illető tulajdonságok jelen vannak. Ez az általánosítással ellenkező irányú folyamat azonban sokkal kevésbé hatékony és kiszámítható, mint a fordítottja. Például a *fehér* szó hallatán nemcsak valamiféle általános 'fehér' fogalom vagy érzet jelenik meg a tudatunkban, hanem legalább néhány általános fajtája a fehérség különböző megnyilvánulásainak (mint amilyen a fehér lepedő, fal, papír, haj), és teljesen kiszámíthatatlanul esetleg ezek még konkrétabb emlékei is (mint akár egy konkrét személy fehér haja). Tehát míg az általánosítás irányában minden esetben szabad az út, az általánosabb emlékek felől a konkrétabbak felé sokkal kevésbé erősek és általános érvényűek az asszociációk.

3.3. Érintkezésen alapuló asszociáció

Az emléknyomok közötti másik fontos kapcsolat az, amelyik nem hasonlóságon, hanem „érintkezésen", együttes előforduláson alapul: az egymáshoz kapcsolódóan tapasztalt emléknyomok kapcsolatban maradnak egymással, az egyikről

eszünkbe jut a másik. Ennek a fajta kapcsolatnak is döntő szerepe van a nyelvi emlékezetben (Saussure a nyelvvel kapcsolatban **szintagmatikus** kapcsolatnak nevezte), hiszen eleve ez kapcsolja össze a közvetlenül nyelvi tapasztalatainkat a használat körülményeivel (Saussure szavával ez a kapcsolat maga a **jel**), de a nyelvi tulajdonságok jellegzetes együttes előfordulásait (mint amilyen a nyelv hangkészlete) és időbeli egymásutánjait is (ilyenek például a nyelv fonotaktikája vagy jellegzetes szórendi mintázatai).

Az általam javasolt elméleti modellben az asszociációt nem önálló egységként ábrázolom, például úgy, ahogy az ún. szemantikai hálókban (Quillian 1966; Collins–Quillian 1969; Schank 1975; Fahlman 1979; Sowa 1991) szokás. A szemantikai hálókban és más asszociációs modellekben a csomópontok „fogalmaknak" (vagy gyakran „szavaknak") felelnek meg, és az asszociációkat a csomópontok közti kapcsolatok ábrázolják. Ehelyett én azt javaslom, hogy az alapegységek (a „csomópontok") különböző (egyedi vagy általánosítás révén létrejövő) emléknyomoknak felelnek meg. Ennek az a hatalmas előnye, hogy **részleges** (csak halványan meglévő) emlékekről is értelmes beszélni, ezek ugyanolyan alapegységek, mint a konkrét egyedi emlékek. Másrészt egy-egy ilyen emléknyom eleve komplex, több tulajdonság együttes jelenlétének felel meg, és ez ábrázolja azt, hogy az érzetek egymással asszociálódnak, képzettársítás alakul ki közöttük, vagyis hogy együttes, egyidejű megtapasztalásuk nyomot hagy az emlékezetben.

3.4. A modularitás hiánya

Az emlékek egyik legfontosabb sajátossága a **heterogeneitásuk**: nemcsak nehéz megmondani, hogy egy-egy emlék inkább vizuális, inkább motoros, inkább verbális stb. természetű-e, hanem feltételezhetjük, hogy ezek általában együtt vannak jelen. Nincs meggyőző bizonyíték arra, hogy elkülönült vizuális, motoros stb. emléknyomok léteznek a tudatunkban, legfeljebb arra utaló jelek vannak, hogy különböző emlékek esetében más-más modalitás játszik uralkodó szerepet (l. Barsalou 1999; 2005). Például az 'elefánt' esetében legtöbbünknél a vizuális és a verbális emlékek a dominánsak (mert az elefántnak leginkább a látványára és róla szóló állításokra emlékezhetünk), míg az 'úszik' esetében a motoros és a taktilis emlékek dominálnak (legalábbis annál, aki szokott úszni).

Mindez a nyelvi természetű emlékekre nézve azt jelenti, hogy ezek formai (hangtani szintű) és tartalmi (használati) vonatkozásai összefonódnak bennük, és bár külön-külön általánosíthatók, valószínűleg az általánosításaik többsége is heterogén. Vagyis nem beszélhetünk egy olyan kognitív szintről, amelyet „a gondolkodás nyelvének" lehetne nevezni, amelyre a különböző forrásokból származó

emlékek „lefordítódnak", és amely modalitástól függetlenül tárolja ezeket (Barsalou 2005 például egy sor olyan kísérleti bizonyítékot mutat be, amely arra utal, hogy „le nem fordított" emlékek léteznek, „lefordítottak", egységes formában tároltak azonban nem). Ezért minden bizonnyal a valóságtól igen távol álló, és sok szempontból alkalmatlan modell az olyan, amely különböző nyelvi „modulok" létezését feltételezi, még ha ezek bonyolult együttműködését elismeri is.

3.5. Gátlások

Az asszociatív memóriamodellek legproblematikusabb része a **gátlások** problémája. Egyrészt nyilvánvaló (és az ilyen modellek működésének elengedhetetlen eleme), hogy az egymásnak ellentmondó, egymással összeférhetetlen „fogalmak", „képzetek" gátolják egymást, vagyis ha az egyik aktív, akkor megakadályozza a többi aktiválódását. Másrészt egyáltalán nem világos, hogyan épülnek ki, hogyan tanulhatók a gátló kapcsolatok. A jelenleg létező tanulási modellek mindegyikében probléma ez, és ez az alábbiakban ismertetendő modellre is igaz: az általam javasolt modellben is külső, emberi beavatkozás szükséges a gátló kapcsolatok létrejöttéhez.

A gátlások modellálásához valószínűleg kettős mechanizmust kell majd feltételezni: az egyik típus az inherensen összeférhetetlen érzékletek, emlékek között működő gátlás, amelynek létrejötte automatikus (a jelenlegi gyakorlati modell, amelyet ismertetni fogok, csak ezt a fajta gátlást ragadja meg), míg a másik az **elvárások** be nem igazolódása révén alakul ki. Akkor, ha több erős emléknyom alapján azok kombinációját várnánk egy bizonyos helyzetben, és mégis mást tapasztalunk, akkor közöttük gátlás épül ki. Például egy bizonyos szó kicsinyítő alakja a mi tapasztalataink alapján $a + b$ alakú lenne, és mégis $a + c$ alakban halljuk, akkor az a és a b között épül ki gátlás. Ehhez a mechanizmushoz azonban azt kell feltételeznünk, hogy a nyelvi tapasztalatot mindig ellenkező irányú aktivitás kíséri, vagyis ha egy bizonyos helyzetben hallunk valamit, akkor összevetjük azzal, amit mi abban a helyzetben mondtunk volna. (Azért beszéltem ehelyett **elvárásról,** mert az esetek többségében valószínűleg meg is előzi a tapasztalatot a befogadó által végzett szimuláció.) Ez az a mechanizmus, amely az általam elkészített modellben nem szerepel.

4. A modell gyakorlati megvalósítása

Az alábbiakban a fenti elképzeléseknek egy nagyon leegyszerűsített, gyakorlati meggondolásokból egyelőre igen korlátozott megvalósítását ismertetem. A gya-

korlati modell konkrétan egy sor olyan számítógépes programban testesül meg, amelyek szimulált nyelvi tapasztalatoknak egy asszociatív memóriamodellbe való beépülését, illetve a rájuk adott reakciókat modellálják.

4.1. A rendszer áttekintése

Annak ellenére, amit fentebb a modularitás hiányáról mondtam (l. 3.4. pont), a gyakorlati modell semennyire sem modellálja a különböző modalitásokból eredő információk heterogeneitását és sokaságát. A rendszer által kezelt adattípus olyan hálózat, amelynek minden csomópontja valamilyen közelebbről meg nem határozott **általánosításnak** felel meg. Az általánosításoktól csak annyit követelünk meg, hogy az általánosság szempontjából legyenek **részben rendezve,** de egyelőre semmi sem felel meg annak, hogy az egyes csomópontok milyen érzékelési, motoros stb. kapcsolatokkal rendelkeznek. A gyakorlatban az általánosításokat implementáló legegyszerűbb modulban egy-egy általánosítás nem más, mint **jegyek** (primitívnek tekintett entitások) egy halmaza, az általánossági reláció pedig a jegyhalmazok közötti **részhalmaz** reláció. Természetesen ez a modul tetszőlegesen bonyolult más modulra lecserélhető. Nem különböztetem meg a memóriamodellben tárolt általánosításokat a rendszer működése során beérkező inputoktól, vagyis minden olyan tapasztalatot, amelyet a rendszer beépít, illetve amelyre reagál, szintén egy-egy általánosítás (a legegyszerűbb esetben egy-egy jegyhalmaz) képvisel.

A rendszernek egyetlen üzemmódja van: amikor egy tapasztalat bekerül a rendszerbe, akkor a rendszer először is **felismeri,** ami azt jelenti, hogy a neki megfelelő csomópont **aktiválódik.** Ha az illető tapasztalatot már korábban beépítette a rendszer (és még nem felejtette el), akkor már létező csomópont aktiválódik; ha nem, akkor új, aktív csomópont keletkezik. A rendszer egyik paramétere, hogy a beérkező inger hatására mekkora ennek a **kezdeti aktivációnak** a mértéke (egy 0 és 1 közötti szám, alapértelmezése 1).

A következőkben a rendszerben néhány cikluson keresztül (az is a rendszer egyik paramétere, hogy hány ilyen ciklus van) terjed a csomópontok aktivációja. A frissen kapott tapasztalatnak megfelelő csomópont minden ciklus elején újra aktiválódik (az előbb említett kezdeti aktivációs szintre).

A szokásos szemantikai hálókkal szemben a memóriamodellben a csomópontok közötti kapcsolatokat a rendszer nem tanulja, és ezek nem is változnak a rendszer működése során: minden X általánosításnak megfelelő csomópont a nála közvetlenül általánosabb X^+ és a nála közvetlenül kevésbé általános X^- általánosításoknak megfelelő csomópontokkal van összeköttetésben, ezek felől és

ezek felé terjedhet az aktiváció a rendszerben. Ezeken a kapcsolatokon kívül csak (szintén kétirányú) gátló kapcsolatok léteznek, az egymást kizáró általánosítások között.

Mint az általánosításokkal kapcsolatban a fenti 3.2. pontban említettem, minden tapasztalat automatikusan aktiválja a nála általánosabbakat. Ez a gyakorlatban azt jelenti, hogy az aktiváció terjedése az egyes csomópontoktól a náluk közvetlenül általánosabbakig szinte akadálytalan (ezt is egy 0 és 1 közötti paraméter határozza meg, kezdeti értéke közel 1). Ugyanott azt is említettem, hogy az általánosabbaktól a konkrétabbak felé az asszociációk gyengébbek és kiszámíthatatlanabbak – ezt egy másik 0 és 1 közötti paraméter paraméter szabályozza, amelynek kezdeti értéke közelebb van a 0-hoz, mint az 1-hez.

A rendszer működésének középpontjában tehát az aktiváció terjedése áll, ami nem jelent mást, mint hogy minden ciklusban újraszámoljuk minden egyes csomópont aktivációs szintjét, annak alapján, hogy a vele összeköttetésben álló csomópontoknak mekkora az aktivációs szintjük. A számítás sok tényezőt vesz figyelembe, amelyek között a legfontosabb a **gyakoriság** (l. 3.1. pont): minél nagyobb egy csomópont (relatív) gyakorisága, annál nagyobb intenzitással terjed róla tovább az aktiváció minden irányban. A gyakoriság számolása úgy történik, hogy minden ciklus végén eggyel növekszik azon csomópontok abszolút gyakorisága, amelyeknek az aktivációja elér egy bizonyos szintet (ez a küszöb is 0 és 1 közötti paraméter), és egy-egy csomópont relatív gyakorisága az abszolút gyakoriságának és az abszolút gyakoriságok összegének a hányadosa. (Az abszolút gyakoriságok összegének valamilyen hatványát is használhatjuk.) A gyakoriság számításánál tehát az aktivációs szintet vesszük alapul, akár külső inger, akár az aktiváció terjedése (endogén aktiváció) okozza azt.

Az új aktivációs szint kiszámolása úgy történik, hogy a szomszédos (általánosabb és specifikusabb képzeteknek megfelelő) csomópontok aktivációját megszorozzuk a relatív gyakoriságukkal (vagy annak valamilyen hatványával), valamint annak a két konstansnak valamelyikével, amelyik kifejezi, hogy általánosabb vagy specifikusabb szomszédról van-e szó, majd az így kiszámított értékeket összegezzük. Ugyanígy összegezzük a csomóponttal gátló kapcsolatban álló csomópontokra hasonlóan adódó értékeket is; ezek negatív irányban fogják befolyásolni a vizsgált csomópont aktivációját. Az összegzést a jól ismert „valószínűségi vagy" (*probabilistic or,* POR) művelettel végzi a rendszer, az $a - b$ alakú kivonásokat pedig hasonlóan a „valószínűségi" értelemben vett $a \wedge \neg b = a \cdot (1 - b)$ művelettel.

A ciklus végén minden csomópontra érvényesítenünk kell a **spontán deaktiválás** és a **spontán felejtés** követelményét. Ez azt jelenti, hogy azoknak a csomópontoknak, amelyek nem váltak aktívabbá a ciklus során, egy bizonyos

tényezővel (ez is egy 0 és 1 közötti paraméter) csökkentjük az aktivációját, valamint hogy azokat a csomópontokat, amelyeknek a relatív gyakorisága nem ér el egy bizonyos értéket (szintén egy 0 és 1 közötti paraméter), egyszerűen töröljük a memóriából. Mivel a gyakoriság növekedése az aktivációs szinttől függ, és mivel a nem aktivált csomópontok aktivációja spontán módon csökken, a felejtés azokat a csomópontokat fogja érinteni, amelyek nem aktiválódtak elég gyakran, és nem érinti azokat, amelyek (korábbi gyakoriságuktól függetlenül) a közelmúltban aktívak voltak.

4.2. A rendszer működése

A program jelenlegi implementációjában igen lassan és kevéssé hatékonyan működik, ezért csak néhány kísérletet tudtam végezni rajta. Az eddigiek alapján megállapítható, hogy a paraméterek ésszerű korlátok között viszonylag szabadon változtathatóak, a rendszer működését nem befolyásolják lényegesen. Néhány (4–5) ciklus után a memória általában olyan állapotba kerül, amely később már nem változik jelentősen. Ha a végállapotban megvizsgáljuk, hogy mely csomópontok a legaktívabbak, a hozzájuk tartozó általánosítások megadják, hogy mik a memóriában tárolt ismeretek alapján a legvalószínűbb **kiegészítései** az ingerként megadott általánosításoknak. Például ha a memóriát úgy építettük fel, hogy *-a* végű nem kicsinyítő alakokat és ugyanilyen értelmű, de kicsinyítő *-u* végű alakokat ismert meg a korábbiakban, akkor a 'pater + kicsinyítő' jegyhalmazhoz (az *apa* 'pater + nem kicsinyítő' jegyhalmaz ismeretében) az *apu* jegyeinek megfelelő csomópontok aktiválódnak, még akkor is, ha más tövek más módon alkotott kicsinyítő alakjait is megadtuk. Persze, mint említettem, kívülről kell meghatározni, hogy a 'kicsinyítő' és a 'nem kicsinyítő' jegyek kizárják egymást, valamint az '*-a* végű' és a hasonló hangtani leírásokat is egyértékű jegyek formájában, vagyis igen mesterséges alakban kell megadni.

Valamivel látványosabb kísérletet is végeztünk, az MTA Nyelvtudományi Intézetében készített igei bővítménykeret-szótár segítségével (Gábor–Héja 2007). Ez 17511 ige bővítménykeretét tartalmazza. A bővítménykereteket jól tudtuk egyértékű jegyek formájában kódolni (egy-egy igét 5–10 jeggyel jellemeztünk), és a bővítménytípusoknak megfelelő jegyek együttes előfordulásaiból építettük fel az asszociatív memóriát. A program implementációs hiányosságai miatt ez csak sok lépésben, apránként volt lehetséges. A tanulási folyamat után a rendszerben 1501 jegyhalmaznak megfelelő csomópont volt. A memória tesztelését úgy végeztük, hogy egy-két jegyből álló bővítménymegjelöléseket közöltünk ingerként a rendszerrel, majd 3–4 ciklus elteltével megvizsgáltuk a legaktívabb csomópontokat.

Az eredmény a vártnak megfelelő volt, például a 'célhatározó szerepű infinitívusz' jegy aktiválása után nem sokkal az 'szándékosan cselekvő alany', a 'valahonnan valahova' és a 'mozgásige' jegyeket tartalmazó kombinációk voltak a legaktívabbak. Vagyis ha azt tudjuk, hogy egy igének van célhatározói értelmű infinitívuszi bővítménye, akkor a legvalószínűbb az, hogy az ige valahonnan valahova irányuló mozgást jelöl.

Ezeknél sokkal bonyolultabb feladatokat egyelőre csak elméletben tudunk elképzelni. Például tételezzük fel, hogy a feladat a 'híd' jelentésű szó szuperesszívusz (SUE) alakjának megalkotása. Feltételezhetjük, hogy a 'híd' jelentéshez a *híd* szótő különböző alakjai vannak társítva, a SUE funkcióhoz a legkülönbözőbb tövek szuperesszívuszi alakjai (és ezek általánosításai), valamint különböző szópárok, így azonos tövek nominatívusz–szuperesszívusz párjainak általánosítása is. Egy ilyen feladatban két problémával kell a rendszernek megküzdenie. Az egyik, hogy a *híd* tőhöz, bár csak elölképzett magánhangzót tartalmaz, hátulképzett toldalékok járulnak – ez a tudás a *híd* alakjainak ismeretében található meg; a SUE alaknak ezekhez az alakokhoz kell hasonlítania, többek közt hátulképzettségben, így a *hídon* és a **hídan* alakok maradnak meg (mivel a SUE alakok mind -V*n* végűek). És éppen ez a másik probléma, hogy a **hídan* alaknak megfelelő általánosítások ne legyenek aktívak, annak ellenére, hogy a *híd* tőnek **hído-* kezdetű alakjai egyáltalán nincsenek a SUE alakon kívül, ilyenekről tehát nem tartalmazhat általánosítást a memória. Viszont ugyanígy nincsenek emlékeink *-an* végződésű SUE alakokról sem, ilyenek ugyanis nem léteznek. Mivel az *-on* végződésű SUE alakokra vonatkozó általánosítások sokkal erősebbek, mint a *hída-* (*hida-*) kezdetű alakokra vonatkozóak, reményeink szerint a *hídon* alakra jellemző általánosítások lesznek a legaktívabbak.

5. Következtetések és tervek

Nagy eredménynek tartom, hogy – ha laboratóriumi méretekben is – sikerült egy olyan memóriaalapú modell alapjait lefektetni, amely hosszabb távon képes lehet megragadni a nyelvi viselkedés legáltalánosabb mechanizmusaira jellemző folyamatokat, és így alapjául szolgálhat bonyolultabb és minden eddiginél hatékonyabb nyelvi szimulációknak. A modell a vizsgált nyelvtől független mechanizmusokat használ, olyanokat, amelyek mind a megértés, mind a produkció szempontjából alapvetőek, tehát mindezeknek a folyamatoknak, illetve ezek részeinek a modellálását lehetővé teszik.

Az a logikai probléma, amelynek közelítő megoldását ez a modell lehetővé teszi, nem más, mint az **abdukció** problémája. Ez azt jelenti, hogy egy formula-

halmazhoz, amit egy következtetés konklúziójának tekintünk, keresünk minél nagyobb konzisztens premisszahalmazt, amelyből következhet. (A konklúziót esetünkben az inger hatására aktiválódó csomópontok ábrázolják.) A premisszákat egy olyan adatbázisban próbáljuk megtalálni, amelyben különböző megbízhatósággal (valószínűséggel) igaznak tekinthető formulák vannak tárolva – ez az adatbázis esetünkben nem más, mint a korábbi tapasztalatokat tároló memória. Az abdukciós feladat megoldása tehát éppen olyan kiegészítési tevékenységet jelent, amilyet a memóriaalapú nyelvelméletek a nyelvi tevékenység lényegének tekintenek. Amikor egy funkció (jelentés) és esetleg néhány formai ismérv alapján, ezeket az információkat kiegészítve abdukció segítségével előállítunk egy nyelvi formát, akkor analógiás nyelvi produkciót valósítunk meg. Ha egy forma (hangalak) és esetleg némi kontextuális információ kiegészítésével egy részletesebb közlési szándékról alkotunk hipotézist, akkor analógiás megértést végzünk.

Ennek alapján talán némi alappal reménykedhetünk abban, hogy a fent felvázolt modell fokozatos gazdagításával, egyre hatékonyabb implementációival fokozatosan közelíthetünk ahhoz a kitűzött célhoz, hogy a legalapvetőbb nyelvi képességeket egyre finomabban és a gyakorlatban egyre inkább alkalmazható módon tudjuk modellálni.

Irodalom

Aamodt, Agnar – Enric Plaza 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. Artificial Intelligence Communications 7: 39–52.

Barsalou, Lawrence W. 1999. Perceptual symbol systems. Behavioral and Brain Sciences 22: 577–660.

Barsalou, Lawrence W. 2005. Situated conceptualization. In: Henri Cohen – Claire Lefebvre (szerk.): Handbook of categorization in cognitive science. New York: Elsevier. 619–650.

Blevins, James P. – Juliette Blevins 2009. Introduction: Analogy in grammar. In: James P. Blevins – Juliette Blevins (szerk.): Analogy in grammar: Form and acquisition. Oxford: Oxford University Press. 1–12.

Blevins, Juliette 2004. Evolutionary phonology: The emergence of sound patterns. Cambridge: Cambridge University Press.

Blevins, Juliette 2006. A theoretical synopsis of Evolutionary Phonology. Theoretical Linguistics 32: 117–165.

Blevins, Juliette – Andrew Garrett 2008. Analogical morphophonology. In: Kristin Hanson – Sharon Inkelas (szerk.): The nature of the word. Essays in honor of Paul Kiparsky. Cambridge MA: MIT Press. 527–546.

Bybee, Joan L. 2001. Phonology and language use (Cambridge Studies in Linguistics 94). Cambridge: Cambridge University Press.

Bybee, Joan L. 2005. Language change and universals. In: Ricardo Mairal – Juana Gil (szerk.): Linguistic universals. Cambridge: Cambridge University Press. 179–194.

Bybee, Joan L. 2006. Frequency of use and the organization of language. Oxford: Oxford University Press.

Bybee, Joan L. – Revere Perkins – William Pagliuca 1994. The evolution of grammar. Tense, aspect and modality in the languages of the world. Chicago & London: Chicago University Press.

Collins, Allan M. – M. Ross Quillian 1969. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior 8: 240–247.

Daelemans, Walter – Peter Berck – Steven Gillis 1997. Data mining as a method for linguistic analysis: Dutch diminutives. Folia Linguistica (Acta Societatis Linguistica Europaeae) 31: 57–75.

Daelemans, Walter – Antal van den Bosch 2005. Memory-based language processing. Cambridge: Cambridge University Press.

Esper, Erwin Allen 1973. Analogy and association in linguistics and psychology. Athens GA: University of Georgia Press.

Fahlman, Scott E. 1979. NETL: A system for representing and using real-world knowledge. Cambridge MA: MIT Press.

Fillmore, Charles J. 1988. The mechanisms of 'Construction Grammar'. BLS 14: 35–55.

Gábor, Kata – Enikő Héja 2007. Clustering Hungarian verbs on the basis of complementation patterns. In: John A. Carrol – Antal van den Bosch – Annie Zaenen (szerk.): Proceedings of the ACL'07 conference, Prague. Prague: Association for Computational Linguistics. 91–96.

Goldberg, Adele 1997. Construction grammar. In: Keith E. Brown – Jim E. Miller (szerk.): Concise encyclopedia of syntactic theories. New York: Elsevier.

Kay, Paul 1995. Construction grammar. In: Jef Verschueren – Jan-Ola Östman – Jan Blommaert (szerk.): Handbook of pragmatics: Manual. Amsterdam & Philadelphia: John Benjamins. 171–177.

Kay, Paul – Charles J. Fillmore 1999. Grammatical constructions and linguistic generalizations: The *what's x doing y?* construction. Language 75: 1–33.

Kolodner, Janet 1983. Reconstructive memory: A computer model. Cognitive Science 7: 281–328.

Kolodner, Janet L. 1993. Case-based reasoning. San Mateo CA: Morgan Kaufmann.

Langacker, Ronald W. 1987. Foundations of cognitive grammar, Vol. 1: Theoretical prerequisites. Stanford: Stanford University Press.

Langacker, Ronald W. 1990. Subjectification. Cognitive Linguistics 1: 5–38.

Langacker, Ronald W. 1991. Foundations of cognitive grammar, Vol. 2: Descriptive application. Stanford: Stanford University Press.

Langacker, Ronald W. 2008. Cognitive grammar: A basic introduction. Oxford: Oxford University Press.

Lebowitz, Michael 1983. Memory-based parsing. Artificial Intelligence 21: 363–404.

Locke, John 1700. An essay concerning human understanding. Negyedik kiadás. London: Awnsham and John Churchil. Első kiadás: 1689.

Paul, Hermann 1880. Prinzipien der Sprachgeschichte. Tübingen: Niemeyer.

Quillian, M. Ross 1966. Semantic memory. Doctoral dissertation, Carnegie-Mellon University.

Russell, Stuart J. – Peter Norvig 1995. Artificial intelligence: A modern approach. Upper Saddle River NJ: Prentice Hall.

Sag, Ivan 1997. English relative clause constructions. Journal of Linguistics 33: 431–484.

Sag, Ivan 2001. Aspects of a theory of grammatical construction. Paper presented at the First International Conference on Construction Grammar, Berkeley, CA.

Saussure, Ferdinand de 1916. Cours de linguistique générale. Paris: Payot.

Schank, Roger 1982. Dynamic memory: A theory of learning in computers and people. Cambridge: Cambridge University Press.

Schank, Roger C. (szerk.) 1975. Conceptual information processing. Amsterdam: North-Holland.

Schank, Roger C. – Alex Kass – Christopher K. Riesbeck 1994. Inside case-based explanation. Hillsdale NJ: Erlbaum.

Skousen, Royal 1989. Analogical modeling of language. Dordrecht: Kluwer.

Skousen, Royal 1992. Analogy and structure. Dordrecht: Kluwer.

Skousen, Royal – Daryle Lonsdale – Dilworth B. Parkinson (szerk.) 2002. Analogical modeling: An exemplar-based approach to language. Amsterdam & Philadelphia: John Benjamins.

Sowa, John F. (szerk.) 1991. Principles of semantic networks: Explorations in the representation of knowledge. San Mateo CA: Morgan Kaufmann.

Taylor, John 2002. Cognitive grammar. Oxford: Oxford University Press.

Analogical learning using an associative memory model

Abstract: The paper comprises three larger sections. The first one summarizes that part of the history of linguistics which proposes or presupposes models of language different from the generativist one, i.e., which does not conceive of language as "the class of well-formed sentences". In the second section, I formulate a concrete proposal on what type of models emerge from those heterogeneous non-generativist approaches. I call this alternative type of models "memory-based", because it is based on a view of language as a system of memory traces and the associations and inhibitions between them. Finally, in the third section, I sketch the experimental computational system that I have implemented in order to simulate most aspects of the working of a memory-based model. I also present the experimental results that have been achieved using the implementation.

Keywords: computer simulation, memory-based models, non-generative linguistics, analogy, models of language

A mondatoktól a hatóköri relációkig – és vissza*

Alberti Gábor – Károly Márton – Kleiber Judit

Pécsi Tudományegyetem, Nyelvtudományi Tanszék, ReALIS Kutatócsoport, Pécs
alberti.gabor@pte.hu; harczymarczy@gmail.com; kleiber.judit@pte.hu

Kutatási célunk a minőségi gépi fordítás és megbízható információkinyerés megvalósítása. Ennek érdekében indítottunk egy alprojektet, melyben a referencialitást és az információstruktúrát tárjuk fel a (magyar) kijelentő mondatban. Az információkinyerés legfontosabb része az a folyamat, melynek bemenete egy mondat, kimenete egy információstruktúra, ami a gyakorlatban nem más, mint operátorok lehetséges hatóköri sorrendjei (elfogadásnál). A gépi fordítás első részében is hasonló folyamat megy végbe: a forrásnyelvi mondat információstruktúrájára van szükségünk. Ez után egy ellentétes irányú folyamat következik (generálás), melynek inputja egy információstruktúra, outputja egy intonált szósor, vagyis egy célnyelvi mondat. Az elfogadás menetét a generálás folyamatára alapozzuk. És mivel megközelítésünk „totálisan lexikalista", az igék lexikai leírása felel a generált mondat szórendjéért és intonációjáért.

Kulcsszavak: lexikalista nyelvtan, operátori hatókörök (fókusz, kvantor, topik, kontrasztív topik), intonáció, szórend, referencialitás

1. Bevezetés: magyar mondatok generálása és elfogadása

Mivel kutatási célunk a minőségi gépi fordítás (Alberti–Kleiber 2004; 2010) és megbízható információkinyerés (Alberti–Kleiber 2003) megvalósítása (l. még Alberti et al. 2010), indítottunk egy alprojektet, melyben a **referencialitást** és az **információstruktúrát** tárjuk fel a kijelentő mondatban.

Elsődlegesen magyar adatokkal dolgozunk, kihasználva, hogy e nyelv mondataira egyrészt nagyon gazdag és explicit módon megjelenő információs struktúra jellemző: különféle topikok, kvantorok és fókuszok rendszere (Kiefer 1992; Szabolcsi 1997; É. Kiss 2002; Alberti–Medve 2000); másrészt egy szintén viszonylag explicit (négyfokozatú) referencia-rendszer (Alberti 1997), melynek legkésőbb feltárt eleme a **specifikus határozatlan** fokozat (de Jong–Verkuyl 1984; É. Kiss 1995; Kálmán 1995). Az általunk tekintett bemeneti adatok egyik fajtája:

* A ReALIS Elméleti és Számítógépes Nyelvészeti Kutatócsoport működésének és e cikk első verziója megírásának a lehetőségét az OTKA (60595, 2006-2011) támogatásának köszönhetjük, a végső verziót pedig már a TÁMOP-4.2.1.B-10/2/KONV/2010/ KONV-2010-0002 (A Dél-dunántúli régió egyetemi versenyképességének fejlesztése) segítségével írtuk meg, a két anonim lektor tanácsai alapján.

hangsúlyjelekkel ellátott (magyar) szavak rendezett halmaza.[1] Ekkor programunk feladata annak meghatározásában áll, hogy egyáltalán jól formált mondattal állunk-e szemben, megfelelő referencia-fokozatú argumentumokkal és lehetséges információstruktúrával; elvárás továbbá e szemantikai adatok megadása, beleértve ebbe a topikok, kvantorok és fókuszok lehetséges hatóköri sorrendjeinek közlését. Ezt a feladatot **elfogadásként** (**elfogadási irányként**) fogjuk említeni. Hangsúlyjelek nélkül érkező szósorok „elfogadására" is vállalkozunk; ennek első lépése: felruházni őket minden lehetséges hangsúlymintázattal. Az ellentétes irányú feladat, amit elvégzünk, **generálásnak** nevezhető, és ennek kimenete: egy intonált mondat. A generálás az időjeles igék gazdag lexikai leírásán alapul, amely az argumentumok ellenőrzését és mondatbeli elrendezését irányítja. Összhangban grammatikafelfogásunk „totálisan lexikalista" filozófiájával (Alberti et al. 2004), a mondatok szövevényes preverbális operátorzónájának kialakításáért lexikonon belüli speciális **generátorszabályok** felelősek.

2. A referencialitási követelmények és az információs struktúra

A magyar nyelvben, hasonlóan ebben a tekintetben az angolhoz, egy határozatlan (*egy/a(n)*) és egy határozott (*a(z)/the*) névelő áll rendelkezésre a **referencialitási fokozatok** elkülönítésére.[2] Bár ez a tény önmagában két referencialitási fokozat létét sugallja, a nyelvi adatok komplex vizsgálata révén arra juthatunk (nemcsak az angol és a magyar (Alberti 1997) adatok alapján, hanem még a névelőket nélkülöző finn nyelv adatai alapján is), hogy (legalább) három referencialitási fokozatnak kell lennie az Univerzális Grammatika mögött álló szemantikai háttérben (amint az (1–5) példasor szemlélteti) – túl a negyedik fokozatnak tekinthető referenciahiányon, ami a magyarban megszámlálható főnevek esetén is előfordulhat (ezt majd a (7) példa szemlélteti):

[1] Négyféle hangsúlyt veszünk figyelembe: „hangsúlytalan"/„ALAPHANGSÚLYOS"/„**FÓKUSZHANGSÚLYOS**"/„↑KONTRASZTÍV HANGSÚLYOS↓"). A magyar mondat releváns intonációs mintázatainak ennél kifinomultabb figyelembe vételére egyelőre nem vállalkozunk, mint ahogy az ige mögötti mondatzóna kevéssé feltárt tényezőit sem vonjuk még be a tárgyalásba (ezekről l. Hunyadi 2002).

[2] Lényegében elfogadjuk ugyan Szabolcsi (1992) érvelését a mellett, hogy az *egy* szóban mindig számnevet lássunk, és ne neki tulajdonítsuk a határozatlanság forrását a magyar főnévi szerkezetben, a gyakorlatban azonban mégis e szócskán keresztül érhetjük tetten a határozatlansági referencialitási fokozatot.

nem referenciális	referenciális		
nem specifikus		specifikus	
határozatlan			határozott
∅ (csupasz egyes sz.)	*egy*	*egy*	*a(z)*

1. táblázat. A négy referencialitási fokozat (és kifejezése a magyarban)

Vessük ugyanis össze a határozatlan névelő jelentéshozadékát az (1b) példában szemléltetett *there*-konstrukcióban és az (1e) mondattípusban: az utóbbi esetben **specifikus** jelentéstartalomról beszélhetünk, ugyanis a kérdéses főnévi szerkezethez tartozó referens, bár konkrétan nem ismert az aktuális diskurzustartományban, részhalmaza egy ismert referensnek.[3]

(1) Referencialitási fokozatok az angolban: három (pozitív) fokozat

a. *There is cock in the kitchen.
ott van kakas -bAn a konyha

b. There is a cock in the kitchen. ⟨+ref, –spec⟩
ott van egy kakas -bAn a konyha
'Van egy kakas a konyhában.'

c. *There is the cock in the kitchen.
ott van a kakas -bAn a konyha

d. *Cock is in the kitchen.
kakas van -bAn a konyha

e. (?)A cock is in the kitchen. ⟨+spec, –def⟩
egy kakas van -bAn a konyha
'Egy kakas benn van a konyhában.'

f. The cock is in the kitchen. ⟨+def⟩
a kakas van -bAn a konyha
'A kakas benn van a konyhában.'

A következő finn példasor azt szemlélteti, hogy még névelők nélkül is képes lehet egy nyelv különbséget tenni a tárgyalt három pozitív referencialitási fokozat között; ami segítségünkre siet, az egyrészt a szórend (⟨–spec⟩: (2a) ~ ⟨+spec⟩: (2b–c)), másrészt a számbeli egyeztetés (⟨–def⟩: (2b) ~ ⟨+def⟩: (2c)):

(2) Referencialitási fokozatok a finnben: szintén három (pozitív) fokozat

a. Tul-i kaksi suomalais-ta tyttö-ä. ⟨+ref, –spec⟩
jön-múlt-3sg két finn-part lány-part
'Érkezett két finn lány.'

[3] „Its referent is a subset of a set of referents already in the domain of discourse" (Enç 1991).

b. Kaksi suomalais-ta tyttö-ä tul-i. ⟨+spec, –def⟩
két finn-part lány-part jön-múlt-3sg
'Két finn lány megérkezett *(mondjuk a várható négy közül)*.'

c. Kaksi suomalais-ta tyttö-ä tul-i-**vat**. ⟨+def⟩
két finn-part lány-part jön-múlt-3pl
'A két finn lány megérkezett.'

A fentiekhez hasonló konstrukciók a magyarban is kiváltják a **nemspecifikussági hatást** (3b), akárcsak a **specifikussági hatást** (3e):

(3) Referencialitási fokozatok a magyarban. I. Létezés

a. *Van kakas a konyhá-ban.

b. Van egy kakas a konyhá-ban. ⟨+ref, –spec⟩

c. *Van a kakas a konyhá-ban.

d. *Kakas benn van a konyhá-ban.

e. (?)Egy kakas benn van a konyhá-ban. ⟨+spec, –def⟩

f. A kakas benn van a konyhá-ban. ⟨+def⟩

A rendszer a nyelvi adatokban az, hogy a létezést (3a–c), létrejövést (4a), illetve létrehozást (5a) jelentő igék páciensi argumentuma **nemspecifikussági hatást** mutat, míg ezeknek az igéknek van egy olyan párja, amelynek a páciensi argumentuma éppen az ellentétes **specifikussági hatást** mutatja, a fenti (3e) példa, illetve az alábbi (4b) és (5b) példa tanúsága szerint.

(4) Referencialitási fokozatok a magyarban. II. Létrejövés

a. Érkezett [*∅ / egy / *a] mexikói a konferenciára.

b. Megérkezett [*∅ / (?)egy / a] mexikói a konferenciára.

(5) Referencialitási fokozatok a magyarban. III. Létrehozás

a. A gyerekek alakítottak [*∅ / egy / *az] énekkart a műsorra.

b. A gyerekek megalakított-ak/ák [*∅ / (?)egy / az] énekkart a műsorra.

Egyes argumentumpozíciókra tehát pozitív vagy negatív specifikussági követelmény vonatkozik. Hasonló módon vizsgálhatjuk a referencialitást is. Azt vehetjük alapul, követve Alberti (1997) hipotézisét, hogy az argumentumoknak referenciálisnak kell lenniük, legalábbis a semleges magyar mondat ige utáni szakaszában. Foglaljuk össze eddigi tapasztalatainkat:

(6) A referencialitási fokozatra vonatkozó pozitív és negatív követelmények a magyar mondat ige utáni szakaszában

a. +ref: (3a), (3d), (4), (5)

b. +spec: (3d-f), (4b), (5b)

c. –spec: (3a-c), (4a), (5a)

Tovább bonyolítja a helyzetet, hogy az imént számba vett követelmények mindegyike semlegesítődhet a magyar mondat ige előtti operátorzónájában, amint láthatjuk is majd a (7)–(9) példasorokban. A ⟨+ref⟩ követelmény semlegesítése olyan (az adott mondatszerkezeti pozícióban) jól formált főnévi kifejezéseket eredményez, amelyek semmilyen névelőt nem tartalmaznak. Vegyük szemügyre a (7) adatsort! Ilyen nem referenciális („csupasz") főnévi kifejezés még semleges mondatokban is előfordulhat, az ige(tő) előtti különleges (igemódosítói) pozíciónak köszönhetően (amely magára vonja a más semleges mondatokban az igetőhöz társuló alaphangsúlyt; l. pl. (4a)). A (7a) mondatban tehát a páciens szerepű argumentum ezért lehet nem referenciális.

(7) A pozitív referencialitásifokozat-követelmények (6a–b) semlegesítésének néhány módja a magyar mondat ige előtti operátorszakaszában

a. Igemódosító (M): A GYErekek *énekkart* alakítottak a MŰsorra.

b. Fókusz (F): A GYErekek ***É**nekkart* alakítottak a műsorra.

c. Kvantor (Q): A GYErekek *énekkart is* Alakítottak a MŰsorra.

d. Kontrasztív topik (K): ↑*Énekkart*↓ Alakíthattok a MŰsorra!

Ezek szerint a semleges mondattartalom akár háromféle szórendi változatban is testet ölthet, ahogyan azt a (8a) példasorban számba vesszük. Melléknévi argumentum esetén persze szűkülnek a lehetőségek (8b), hiszen a referencialitási követelmény (6a) elől egyedül az igemódosítói pozíció jelenthet „menedéket", tekintve, hogy egy melléknévi kifejezést nem tudunk determinálni.

(8) A referencialitási követelmény (6a) semlegesítésének következményei

a. A GYErekek [*∅ / egy*] *énekkart* alakítottak a MŰsorra. ~~⟨+ref⟩~~, ⟨–spec⟩
A GYErekek Alakítottak *egy énekkart* a MŰsorra. ⟨+ref⟩, ⟨–spec⟩

b. *A GYErekek FEstették *ZÖLDre* a KErítést. ⟨+ref⟩, ⟨–ref⟩
A GYErekek *ZÖLDre* festették a KErítést. ~~⟨+ref⟩~~, ⟨–ref⟩

Lássunk végül néhány példát arra, hogy a negatív módon korlátozó referencialitásifokozat-követelmény hogyan semlegesítődhet a magyarban:

(9) A negatív referencialitásifokozat-követelmény (6c) semlegesítése a magyar mondat ige előtti operátorszakaszában, köszönhetően annak, hogy F vagy K operátorpozíciót foglal el egy kifejezés, amely különbözik a páciensi argumentumtól

a. A NAGYszobában van *a kakas*. vö. (3c)

b. TEGnap érkezett *a mexikói* a konferenciára. vö. (4a)

c. A GYErekek alakították *az énekkart* a műsorra. vö. (5a)

Az alábbi 2. táblázatban a +/– referencialitási követelmények eloszlását magyarázó hipotézisünket mutatjuk be. Egy prototipikus semleges mondatban (A. típus) a mondatéli topikzóna és az ige mögötti komplementumzóna a referensek **horgonyzásának** feladatát hivatott ellátni, ami ⟨+ref⟩ argumentumokat kíván, míg az időjeles ige képezi a mondat **új állítást** közlő centrumát, ami megadja az új információt a lehorgonyzott referensekről. Azok a páciensek viszont, amelyeknek a létezését állítja a mondat, nyilvánvalóan az új állítást közlő centrumszakaszhoz tartoznak (B. típus). Az igetövet közvetlenül megelőző pozíció is az új állítást közlő centrumszakaszhoz tartozik, így aztán „menedéket" képes nyújtani nem referenciális argumentumoknak (C. típus). A D. és az E. típusban az a közös, hogy valamilyen új állítás közlésére szolgáló operátor jelenik meg a mondatban (például fókusz), ami magához vonja az új állítást közlő centrum funkcióját a mondat más zónáitól. Ennek következményeképpen a pozitív referenciafokozat-követelmények semlegesítődnek az új centrumzónában (D.), míg a negatívak éppen a mondat egyéb zónáiban semlegesítődnek (E.).

3. Mondatok generálása, (intonált) szósorok elfogadása

Az általunk **információkinyerésként** említett feladat meghatározó része egy olyan eljárás – nevezzük **elfogadásnak** – amelynek bemenete egy mondat, kimenete pedig egy információstruktúra, ami lényegében a lehetséges operátorhatóköri sorrendeknek a megadása. Egy hasonló eljárás képezi a **gépi fordítási** feladat első felét is: a forrásnyelvi mondat információstruktúrájára van szükségünk, hogy aztán egy ellentétes irányú eljárás, a **generálás** megkapja bemeneti adatként az információstruktúrát, kimenetként pedig előállítson ebből egy intonált szósort, azaz egy célnyelvi mondatot. Tekintsük először ez utóbbi eljárást, mivel az előbbi eljárás erre fog épülni.

Minthogy nyelvészeti megközelítésünk – korábbi próbálkozásainkhoz hasonlóan (Alberti–Kleiber 2004) – „totálisan lexikalista" (Alberti et al. 2004), döntően az igék lexikai leírása felelős a generált mondat szavainak sorrendjéért és intonációjáért. Az alábbi (10a) pontban azt a – **maglexikonban** regiszt-

A. Prototipikus semleges mondat **horgonyzó** szerepű argumentumokkal és egy új állítást közlő igével:		
A GYErekek$_{+ref}$	MEGalakították	az Énekkart$_{+ref}$ a MŰsorra$_{+ref}$. (5b)
B. Semleges mondat egy létezést kifejező argumentummal, ami e jellegéből kifolyólag az **új állítást közlő** mondatszakaszhoz tartozik:		
A GYErekek$_{+ref}$	Alakítottak egy Énekkart$_{+ref,-spec}$	a MŰsorra$_{+ref}$. (5a)
C. Semleges mondat egy létezést kifejező argumentummal az igemódosítói pozícióban, ami e jellegéből kifolyólag **az új állítást közlő** mondatszakaszhoz tartozik:		
A GYErekek$_{+ref}$	∅/egy Énekkart$_{-spec}$ alakítottak	a MŰsorra$_{+ref}$. (8a)
D. Fókuszos mondat I.: az **új állítást közlő** mondatszakaszt elfoglalja egy amúgy referencialitási hatás hatálya alá eső argumentum, fókusz-státuszából adódóan (miközben az ige kikerül az új állítást közlő zónából, ugyancsak a fókusz konstrukcióból következően):		
A GYErekek$_{+ref}$	Énekkart$_{+ref}$	alakítottak a műsorra. (7b)
E. Fókuszos mondat II.: az új állítást közlő mondatszakaszt ezúttal egy fókuszált kifejezés foglalja el, miközben az ige és a nemspecifikussági hatás hatálya alá eső argumentum **kikerül az új állítást közlő zónából** (ismét a fókusz konstrukcióból következően):		
A GYErekek	alakították az énekkart$_{-spec}$ a műsorra. (9d)	

2. táblázat. Horgonyzó (szürke háttérrel megjelenített) és új állítást közlő (fehér hátterű) információdarabok különféle mondattípusokban

rált – követelményt szemléltetjük, miszerint az **alakít** ige alanyának (alaphangsúlyos) topikként kell megjelennie egy megfelelő mondat bal perifériáján (’⟨1,T⟩’), a tárgynak a (szintén alaphangsúlyos) igemódosítói pozíciót kell elfoglalnia (’⟨2,M⟩’), maga mögött egy hangsúlytalan igetővel, a *-rA* ragos argumentumnak pedig egy (alaphangsúlyos) posztverbális argumentumpozícióban kell maradnia (’⟨3,A⟩’). A sorszámok egy hatóköri sorrendet adnak, ami azonban egyelőre még, semleges mondat esetében, gyakorlatilag irreleváns. Ezt a lexikai szabályt a továbbiakban **generátornak** fogjuk nevezni.

A (10b) példában azt láthatjuk, hogy a maglexikonbeli default generátor az adott ige esetében a *-rA* ragos argumentum számára írja elő az igemódosítói pozíciót. A (10c) pont pedig azt szemlélteti, hogy a semleges magyar szórend kialakításának követelménye miatt olyan generátor szükségeltetik, amely az igekötőt „küldi” az igemódosítói pozícióba, a nem ágentív alanyi argumentum számára pedig egy posztverbális argumentumpozíciót ír elő.

(10) Default hatóköri sorrendek megadása a maglexikonban néhány magyar lexikai egységben

a. FORM($Arg_{\emptyset}$, $Arg_{\text{-t}}$, $Arg_{\text{-rA}}$)
⟨⟨1,T⟩, ⟨2,M⟩, ⟨3,A⟩⟩ (**default generátor** a maglexikonban)
A GYEREKEK énekkart alakítottak a MŰSORRA. (7a)

b. PAINT($Arg_{\emptyset}$, $Arg_{\text{-t}}$, $Arg_{\text{-rA}}$)
⟨⟨1,T⟩, ⟨3,A⟩, ⟨2,M⟩⟩
A GYEREKEK ZÖLDre festették a KERÍTÉST. (8b)

c. ARRIVE(Prefix, $Arg_{\emptyset}$, $Arg_{\text{-rA}}$)
⟨⟨1,M⟩, ⟨2,A⟩, ⟨3,A⟩⟩
MEGérkezett a MEXIKÓI a KONferenciára. (4b)

Két másik generátortípus igeváltozatokat állít elő; ezek egy **kiterjesztett** lexikontartományban működnek. Amit az alábbi (11a) pont első sorában szemléltetünk, az egy **kiterjesztő generátor**; feladata abban áll, hogy egy argumentumstruktúrába beilleszen egyes szabad határozókat afféle „hamisvonzatként".[4] Majd a (11a) második sorában egy **indukáló** generátort mutatunk be: ennek hatására kerül a két hamisvonzat a mondatéli topiktartományba, az alany egy kvantorpozícióba, a tárgy pedig a fókuszba, mindegyikük megfelelő hangsúllyal (ezért is az indukáló generátor felelős). A (11b-c) pontokban egy kiterjesztő generátort és két indukáló generátort szemléltetünk. Az '⟨1,K⟩' formularészlet például azért felel, hogy a szabad helyhatározóból lett hamisvonzat (*a klubban*) a mondatéli operátortartomány kontrasztívtopik-pozíciójába jusson el a generált mondatban.

(11) Lexikai szabályok, amelyek kiterjesztik az argumentumstruktúrát, illetve nem semleges mondatváltozatokat indukálnak a módosított ige körül

a. FORM($Arg_{\emptyset}$, $Arg_{\text{-t}}$, $Arg_{\text{-rA}}$) ^ ⟨Arg_{Time}, Arg_{Place}⟩ (**kiterjesztő generátor**)
⟨⟨3,Q⟩, ⟨4,F⟩, ⟨5,A⟩, ⟨1,T⟩, ⟨2,T⟩⟩ (**indukáló generátor**)
TEGnap a KLUBban A GYEREKEK is **É**nekkart alakítottak a műsorra.

b. FORM($Arg_{\emptyset}$, $Arg_{\text{-t}}$, $Arg_{\text{-rA}}$) ^ ⟨Arg_{Place}⟩
⟨⟨2,F⟩, ⟨3,M⟩, ⟨4,A⟩, ⟨1,K⟩⟩
A ↑KLUB-ban↓ a **GYE**rekek alakítottak énekkart a műsorra.

c. PAINT($Arg_{\emptyset}$, $Arg_{\text{-t}}$, $Arg_{\text{-rA}}$)
⟨⟨1,T⟩, ⟨2,Q⟩, ⟨3,F⟩⟩
A GYEREKEK MINdegyik KERÍTÉST **ZÖLD**re festették.

[4] Így kezeljük azt a jelenséget, hogy a preverbális operátorzónába éppúgy beléphetnek *egyes* szabad határozók, mint az igevonzatok, érdemesnek látszik tehát egy listába helyezni őket, ahogyan azt egyes HPSG-elemzésekben is láthatjuk (Szécsényi 2009): például *TEGnap is (Q) a **KLUB**ban (F) léptek fel a gyerekek.* Más szabad határozók azonban nem lépnek be hamisvonzatként az igei argumentumszerkezetekbe: pl. **SAJ**nos / **NA**gyon (F) álmosodtam el.

Amit egy generátor előállít, az általánosságban mondatok egy **halmaza**, amelyben tipikusan preferencia szerinti sorrendbe rendezett szórendi permutációk állnak. A magyarban különösen a kvantor felelős azért, hogy egy adott operátorhatóköri sorrendhez többféle szórend is társulhat: ugyanis egy kvantorkifejezésnek „jogában áll" választani a hatóköri sorrendnek megfelelő preverbális operátorhely elfoglalása (σ_1, σ_{10}, σ_{11}, σ_{20}, σ_{30}) és az ige mögötti mondatszakaszban való helyben maradás között (σ_2, σ_3, σ_{20}, σ_{30}, σ_{40}, σ_{50}).

(12) Intonált mondatok generálása: $v \rightarrow \langle \sigma_1, \sigma_2, \ldots, \sigma_K \rangle$

a. PAINT($Arg_{\emptyset}$, $Arg_{\text{-t}}$, $Arg_{\text{-rA}}$)
v: $\langle\langle 3,A\rangle, \langle 1,Q\rangle, \langle 2,M\rangle\rangle \rightarrow \langle \sigma_1, \sigma_2, \sigma_3 \rangle$
σ_1: MINdegyik KERÍTÉST ZÖLDre festették a GYErekek.
σ_2: ZÖLDre festették a GYErekek MINdegyik KERÍTÉST.
σ_3: ZÖLDre festették MINdegyik KERÍTÉST a GYErekek.

b. PAINT($Arg_{\emptyset}$, $Arg_{\text{-t}}$, $Arg_{\text{-rA}}$)
v: $\langle\langle 1,Q\rangle, \langle 2,Q\rangle, \langle 3,M\rangle\rangle \rightarrow \langle \sigma_{10}, \sigma_{20}, \sigma_{30}, \sigma_{40}, \sigma_{50} \rangle$
σ_{10}: A GYErekek is MINdegyik KERÍTÉST ZÖLDre festették.
σ_{20}: A GYErekek is ZÖLDre festették MINdegyik KERÍTÉST.
σ_{30}: MINdegyik KERÍTÉST ZÖLDre festették a GYErekek is.
σ_{40}: ZÖLDre festették a GYErekek is MINdegyik KERÍTÉST.
σ_{50}: ZÖLDre festették MINdegyik KERÍTÉST a GYErekek is.

c. PAINT($Arg_{\emptyset}$, $Arg_{\text{-t}}$, $Arg_{\text{-rA}}$)
v: $\langle\langle 2,Q\rangle, \langle 1,Q\rangle, \langle 3,M\rangle\rangle \rightarrow \langle \sigma_{11}, \sigma_{30}, \sigma_{20}, \sigma_{50}, \sigma_{40} \rangle$
σ_{11}: MINdegyik KERÍTÉST a GYErekek is ZÖLDre festették.

A generált mondathalmaz üres is lehet (λ=$\langle\ \rangle$). Az alábbi (13a) példa azt szemlélteti, hogy egy melléknévi argumentum (ami inherensen $\langle$−ref$\rangle$) nem foglalhatja el a megadott indukáló generátor javasolta argumentumpozíciót (5b). A (13b) pontbeli indukáló generátor kétszeresen is megsérti a referencialitási követelményt (6a), amit a (nem kontrasztív) topik pozíció nem semlegesít (7). Végül a (13c) intonált szósor a magyar nyelvben lehetséges operátorhatóköri sorrendet (13d) sérti meg.[5]

(13) Intonált mondatok üres halmazának generálása: $v \rightarrow \langle \sigma_1, \sigma_2, \ldots, \sigma_K \rangle = \lambda$

PAINT($Arg_{\emptyset}$, $Arg_{\text{-t}}$, $Arg_{\text{-rA}}$)

a. v_1: $\langle\langle 1,T\rangle, \langle 3,A\rangle, \langle \boldsymbol{2,A} \rangle\rangle \rightarrow \lambda$
*A GYErekek FESTETTÉK *ZÖLDre* a KERÍTÉST.

[5] A mondat élén topikfélék állhatnak (egy vagy több vagy egy sem, ezt szimbolizálja a Kleene-féle * jel), amit kvantorok és fókuszok követhetnek, elvileg akár keveredve is. A hatóköri sorrendet a szórend némileg elhomályosíthatja, az ige ugyanis a legnagyobb hatókörű fókuszhoz társul: pl. *A **KLUB**ban (F_1) léptek TEGnap is (Q_2) csak a **GYE**rekek (F_3) fel.*

b. v_2: $\langle\langle \boldsymbol{1,T}\rangle, \langle \boldsymbol{3,A}\rangle, \langle 2,\mathrm{M}\rangle\rangle \rightarrow \lambda$
*GYErekek ZÖLDre festettek KERítést.

c. v_3: $\langle\langle \boldsymbol{1,Q}\rangle, \langle \boldsymbol{2,T}\rangle, \langle 3,\mathrm{M}\rangle\rangle \rightarrow \lambda$
*A GYErekek is a KERítést ZÖLDre festették.

d. $\{\mathrm{T}, \mathrm{K}\}^* \wedge \{\mathrm{Q}, \mathrm{F}\}^* \wedge (\mathrm{M})\ \mathrm{A}^*$

Egy intonált szósor **elfogadása** éppen a generálás ellentéte. Az elfogadási eljárás ráépíthető a generálásra: számba kell vennünk a lehetséges maglexikonbeli igéket, majd össze kell ezeket társítani az összes szóba jövő generátorral, végül pedig az esélyes kombinációkat ki kell értékelni valamilyen hatékony módon. Az alábbi (14a) pont ennek szemléltetésére szolgál. Amennyiben pedig azzal nehezedik a feladat, hogy a bemeneti szósor intonáció nélkül értékelendő ki, az elfogadási eljárásnak azzal kell kezdődnie, hogy a szósort felruházzuk valamennyi esélyes hangsúlymintázattal – nyilván ezúttal is a hatékonyságra törekedve, amit nyelvészeti tapasztalatainkra alapított heurisztikákkal segíthetünk (15).

(14) Hatóköri sorrendek elfogadása (bemenet: intonált mondat): $\sigma \rightarrow \langle v_1, v_2, \ldots, v_\mathrm{K}\rangle$

a. σ_{10}: A GYErekek is MINdegyik KERítést ZÖLDre festették.
$\sigma_{10} \rightarrow \langle v_1\rangle$ (l. a fenti (12) pontot)
PAINT($\mathrm{Arg}_{\emptyset}$, $\mathrm{Arg}_{\text{-t}}$, $\mathrm{Arg}_{\text{-rA}}$)
v_1: $\langle\langle 1,\mathrm{Q}\rangle, \langle 2,\mathrm{Q}\rangle, \langle 3,\mathrm{M}\rangle\rangle$

b. σ_{11}: MINdegyik KERítést a GYErekek is ZÖLDre festették.
$\sigma_{11} \rightarrow \langle v_2\rangle$
v_2: $\langle\langle 2,\mathrm{Q}\rangle, \langle 1,\mathrm{Q}\rangle, \langle 3,\mathrm{M}\rangle\rangle$

c. σ_{20}: A GYErekek is ZÖLDre festették MINdegyik KERítést.
$\sigma_{20} \rightarrow \langle v_1, v_2\rangle$

d. σ_{30}: MINdegyik KERítést ZÖLDre festették a GYErekek is.
$\sigma_{30} \rightarrow \langle v_2, v_1\rangle$

e. σ': *GYErekek ***ZÖLD***re* MINdegyik KERítést festették.
$\sigma' \rightarrow \emptyset$

(15) Hatóköri sorrendek elfogadása (ezúttal a bemenet: intonálatlan szósor):
$\kappa \rightarrow \langle \sigma_1, \sigma_2, \ldots, \sigma_\mathrm{N}\rangle$, ahol
$\sigma_1 \rightarrow \langle v_{1,1}, v_{1,2}, \ldots, v_{1,\mathrm{K}1}\rangle, \ldots, \sigma_\mathrm{N} \rightarrow \langle v_{\mathrm{N},1}, v_{\mathrm{N},2}, \ldots, v_{\mathrm{N,KN}}\rangle$

κ: a gyerekek is zöldre festették mindegyik kerítést
$\kappa \rightarrow \langle \sigma_{20}, \sigma_{22}, \sigma_{23}, \ldots\rangle$; PAINT($\mathrm{Arg}_{\emptyset}$, $\mathrm{Arg}_{\text{-t}}$, $\mathrm{Arg}_{\text{-rA}}$)

a. σ_{20}: A GYErekek is ZÖLDre festették MINdegyik KERítést.
$\sigma_{20} \rightarrow \langle v_1, v_2\rangle$; v_1: $\langle\langle 1,\mathrm{Q}\rangle, \langle 2,\mathrm{Q}\rangle, \langle 3,\mathrm{M}\rangle\rangle$; v_2: $\langle\langle 2,\mathrm{Q}\rangle, \langle 1,\mathrm{Q}\rangle, \langle 3,\mathrm{M}\rangle\rangle$

b. σ_{22}: A GYErekek is **ZÖLD**re festették MINdegyik KERítést.
$\sigma_{22} \rightarrow \langle v_{12}, v_{22}\rangle$; v_{12}: $\langle\langle 1,\mathrm{Q}\rangle, \langle 2,\mathrm{Q}\rangle, \langle 3,\mathrm{F}\rangle\rangle$; v_{22}: $\langle\langle 2,\mathrm{Q}\rangle, \langle 1,\mathrm{Q}\rangle, \langle 3,\mathrm{F}\rangle\rangle$; $^?v_{23}$: $\langle\langle 1,\mathrm{Q}\rangle, \langle 3,\mathrm{Q}\rangle, \langle 2,\mathrm{F}\rangle\rangle$

c. σ_{23}: $^{??}\uparrow$A GYErekek is$\downarrow$ **ZÖLD**re festették MINdegyik KERítést.
$\sigma_{23} \rightarrow \langle \nu_{13}, \nu_{23} \rangle$; $^{?}\nu_{13}$: $\langle\langle 1,K\rangle, \langle 2,Q\rangle, \langle 3,F\rangle\rangle$; $^{??}\nu_{23}$:
$\langle\langle 1,K\rangle, \langle 3,Q\rangle, \langle 2,F\rangle\rangle$

d. σ_{44}: *A **GYE**rekek is ZÖLDre festették MINdegyik KERítést.
$\sigma_{44} \rightarrow \emptyset$

E szakasz utolsó példája a magyar és angol közötti igényes fordítás egyik kulcslépését érinti. Minthogy az angol szórend rendkívül kötött, ami megfelel a magyarban egy hatóköri rendezést indukáló generátornak, ez nem egyszerűen ugyanaz az indukáló generátor (mert nem feltétlenül kompatibilis vele), hanem annak kombinációja egy olyan szabállyal, amit szintén kiterjesztett lexikonbeli generátor segítségével kívánunk megragadni. Olyan generátorokról van szó, amelyek a passzivizálási vagy a **dative shift** képzés argumentumstruktúra-módosító hatásáért felelősek.[6]

(16) Magyar operátorok ~ angol argumentumstruktúra-változatok[7]

a.	GIVE ($\text{Arg}_{\emptyset}$, $\text{Arg}_{\text{-nAk}}$ $\text{Arg}_{\text{-t}}$) ν_2: $\langle\langle 1,T\rangle, \langle 3,A\rangle, \langle 2,F\rangle\rangle$ Péter egy **KÖNY**vet adott Marinak.	GIVE($\text{Arg}_{\emptyset}$, $\text{Arg}_{\text{Obj}_1}$, $\text{Arg}_{\text{Obj}_2}$) ν_2: $\langle\langle 1,T\rangle, \langle 3,A\rangle, \langle 2,F\rangle\rangle$ Peter gave Mary a **BOOK**. Peter adott Mary egy könyv
b.	GIVE ($\text{Arg}_{\emptyset}$, $\text{Arg}_{\text{-nAk}}$ $\text{Arg}_{\text{-t}}$) ν_3: $\langle\langle 1,T\rangle, \langle 2,F\rangle, \langle 3,A\rangle\rangle$ Péter **MA**rinak adott egy könyvet.	GIVE($\text{Arg}_{\emptyset}$, $\text{Arg}_{\text{Obj}_1}$, $\text{Arg}_{\text{Obj}_2}$) $\langle \text{Arg}_{\emptyset}, \text{Arg}_{\text{to}}, \text{Arg}_{\text{Obj}} \rangle + \nu_3$ Peter gave a book to **MA**ry. Peter adott egy könyv -nAk Mary
c.	GIVE ($\text{Arg}_{\emptyset}$, $\text{Arg}_{\text{-nAk}}$ $\text{Arg}_{\text{-t}}$) ν_4: $\langle\langle 2,F\rangle, \langle 1,T\rangle, \langle 3,A\rangle\rangle$ MArinak **PÉ**ter adott egy könyvet.	GIVE($\text{Arg}_{\emptyset}$, $\text{Arg}_{\text{Obj}_1}$, $\text{Arg}_{\text{Obj}_2}$) $\langle \text{Arg}_{\text{by}}, \text{Arg}_{\emptyset}, \text{Arg}_{\text{Obj}} \rangle + \nu_4$ Mary was give-n a book by **PE**ter. Mary volt ad-va egy könyv által Peter
d.	GIVE ($\text{Arg}_{\emptyset}$, $\text{Arg}_{\text{-nAk}}$ $\text{Arg}_{\text{-t}}$) ν_5: $\langle\langle 2,F\rangle, \langle 3,A\rangle, \langle 1,T\rangle\rangle$ A KÖNYvet **PÉ**ter adta Marinak.	GIVE($\text{Arg}_{\emptyset}$, $\text{Arg}_{\text{Obj}_1}$, $\text{Arg}_{\text{Obj}_2}$) $\langle \text{Arg}_{\text{by}}, \text{Arg}_{\text{to}}, \text{Arg}_{\emptyset} \rangle + \nu_5$ The book was give-n to Mary by **PE**ter. a könyv volt ad-va -nAk Mary által Peter

[6] Megközelítésünket Croft (2001, 8. fejezet) nyelvtipológiai elemzései is alátámasztják: a világ nyelveiben felbukkanó számtalan átmenet az aktív igealak és az angol típusú sztenderd passzív forma között *(voice continuum)* leginkább a különféle argumentumokra irányuló topikalizálási igények felől értelmezhető.

[7] Ezúton szeretnénk köszönetet mondani Laczkó Tibornak a fordítások megbeszéléséért.

4. Implementáció

Csak néhány publikáció foglalkozik az operátorok hatóköri viszonyainak és a referencialitásnak a szórendből és intonációból kiinduló, mélyelemzés révén történő pontos meghatározásával. Egy (kiváló) példa Traat és Bos (2004) cikke, de hasonló rendszert magyar nyelvre, amely nyelv releváns és a mi szempontunkból előnyös tulajdonságait már a 2. pontban megvitattuk, még nem építettek.

A mondatokat mindenekelőtt fonológiailag és morfológiailag elemezzük. Megközelítésünk totálisan lexikalista – a totális lexikalizmuson alapuló nyelvtanok nem építenek frázisstruktúrát. A szórendet e helyett **rangparaméterekkel** kezeljük. Általában egész számokat használunk 1-től 7-ig úgy, hogy az 1-es rang a legerősebb; alapesetben közvetlen szomszédosságot takar. A lexikonunk morfémákat tartalmaz szavak helyett, amelyek bármely más morfémát kereshetnek azon a szón belül vagy kívül, amelynek részét képezik.

A mi megközelítésünkben az egyetlen különbség morfológia és szintaxis között az, hogy a szintaktikai alrendszerben a program **más** szóban keresi az egy bizonyos grammatikai relációban lévő morfémákat. De ugyanezért a morfológiai és szintaktikai rangparamétereket külön kezeljük.

A lexikon többféleképpen bővíthető. A további szavak és morfémák hozzávételén kívül új jegyeket vehetünk fel az adatszerkezetbe, amelyek az elemzés pontosságát javítják. A harmadik lehetőség az, hogy a **maglexikont** generáló szabályok révén kibővítjük, létrehozva ezzel a **kiterjesztett lexikont**. A maglexikon egy morféma minden alaptulajdonságát tartalmazza, az alapértelmezett viselkedést is beleértve (pl. egy ige vonzatstruktúráját). A generálást elsődlegesen az intonációra alkalmazzuk: a maglexikon tartalmazza a 'hangsúly' jegyét (egyes szavak egyáltalán nem hangsúlyozhatók, míg mások egy semleges mondatban hangsúlyt kapnak) – de ez a kiterjesztett lexikonban felülíródhat, ha a mondat nem semleges. Például a 'fókuszhangsúlyos' jegyértékkel ellátott entitást automatikusan generáljuk, és betesszük a kiterjesztett lexikonba.

Tudjuk, nem túl hatékony az a módszer, ha egy hosszú mondaton – generálás révén – kipróbáljuk az összes lehetséges intonációs sémát. Két dolgot mindazonáltal figyelembe kell vennünk: a hangsúlyt vagy fókuszhangsúlyt soha nem kapó morfémák (pl. névelők) esetén túl az ige előtti és utáni vonzatok hangsúlyozása is csak bizonyos korlátok között lehetséges.

Alapesetben a magyar vonzat az ige **mögött** foglal helyet. Vannak azonban olyan argumentumok, amelyek az igemódosítói pozícióba vagy topikba szeretnek kerülni. Ezeket a preferenciákat a maglexikonban kell tárolni az argumentumokkal együtt (10). Ha a mondat nem illeszthető a preferált sémába, a legjobb, ha heurisztikákat alkalmazunk a megfelelő generátor létrehozására (11)–(12).

A Prolog (egy logikai programozási nyelv), ideértve a Visual Prolog 7-et (amely talán a Prolog legkidolgozottabb verziója – mi ezt használjuk) lehetővé teszi, hogy „visszafelé" is hívható predikátumokat írjunk (**generálás** (12) és **elfogadás** (14)). Bár a VP7-ben ez nem minden esetben engedélyezett (a predikátumok argumentumait inputra, outputra vagy mindkettőre definiálhatjuk – ily módon sok beépített predikátum használata korlátozva van), kiindulhatunk abból, hogy a predikátumok kiértékelése megfordítható. Ezt az elvet alapul véve a jövőbeli gépi fordító szimmetrikus lehet. Az `anyflow` kulcsszó révén a predikátum minden argumentumát be- és kimenetre is használhatjuk, lehetővé téve akár egész programok „visszafelé" történő végrehajtását is. A `procedure, determ, multi, nondeterm` stb. kulcsszavakkal korlátozhatjuk a predikátum meghiúsulását (a `procedure` mindig sikeres kell hogy legyen), valamint a visszalépési pontok számát (`nondeterm` és `determ`).

Maga az elemzés (**elfogadás**) a következő fázisokból áll:

0. fázis. Mielőtt figyelembe vennénk az intonációt, minden szót elemzünk fonológiailag és morfológiailag. Ezáltal a szófajokat egyértelműen meghatározhatjuk. Gyakorlatilag az utolsó szófajváltó morféma (képző) tartalmazza a releváns 'szófaj' kimenő jegyet. A bemenő és a kimenő szófajokat a képzőhöz tartozó maglexikon-elemben tároljuk.

1. fázis. A tulajdonképpeni szintaktikai elemzés során, más keresésekkel együtt, az ige vonzatait 7-es rangban keressük, ami jelenleg a leggyengébb. Ez kétirányú keresés: minden nem predikatív főnévi kifejezés keresi a predikátumot is, és az eredményt eltároljuk a számítógép memóriájában. A szabad határozói bővítmények is 7-es ranggal keresik az igét, de ha megtalálják, a **kiterjesztő** generátort meg kell hívni, hogy az ige alapértelmezett vonzatstruktúráját módosíthassuk, és az új formát betegyük a kiterjesztett lexikonba (11a–b).

2. fázis. Az **indukáló** generátor megkísérli meghatározni a vonzat diskurzusfunkcióját – T, Q, F, M, A, l. (7) – úgy, hogy létrehozza az összes lehetséges mintát és megkísérli alkalmazni őket a mondatra. Ha az intonáció jelen van az inputban, itt figyelembe vehető, így az elemzés gyorsul. Azonban ha a magyar operátorsorrend sérül, az elemzés meghiúsul és a Prolog visszalépéses mechanizmusa aktiválódik, még mielőtt elérjük a mondat végét (13c).

3. fázis. A referencialitási fokozatok (6) közül kettő az 1. fázisban kezelhető. Először is feltesszük, hogy alapértelmezés szerint minden főnév nem-referenciális. Mivel a névelő keresi a főnevet, a főnév pozitív és negatív referencialitási fokozatait (mint jegyeket) is a sikeres keresés után a kiterjesztett lexikonban létrejövő új formák tartalmazhatják.

Az egyetlen megmaradó probléma a (nem-)specifikusság. Tekintsük a (3) példát. Ha jelen van egy határozatlan determináns (határozatlan névelő, szám-

név stb.), ezt csak a 2. fázis közben vagy után dönthetjük el. Egyéb korlátozások hiányában, amennyiben a semleges mondatban jelen van egy igemódosító, az argumentum specifikus. Ha két példányt generálunk az *egy mexikói*-ból, egyet 'specifikus' és 'határozatlan', egy másikat 'nem–specifikus' és 'referenciális' jegyekkel, a (3b) elemzésének ez utóbbival meg kell hiúsulnia. Természetesen a generált példányokat be kell tennünk a kiterjesztett lexikonba. Ha az igének van módosítója, annak vonzatstruktúrája **eltér ugyanazon** igének módosító nélküli változatától – a spcifikussági kritérium miatt. Egyszerűbb, de lassúbb módszer, ha a maglexikonba az **egy** határozatlan névelőből kettőt teszünk be, egy specifikusat és egy nem specifikusat. Ha a maglexikonban van két morféma ugyanazon testtel, mindkettőt figyelembe veszi a rendszer **kezdettől** fogva (0. fázis), ami lassítja az elemzést, mert ha az 1. vagy 2. fázisban hiba történik, az elemzés nagy része, beleértve részben a 0. fázist is, újraindul. Ilyenkor ugyanis az **ugyanazon maglexikonbéli** elemhez tartozó **összes** unifikációt – a legalapvetőbbeket is – fel kell oldanunk.

5. Összefoglalás

Nagyon kevés számítógépes rendszer létezik, melynek célja a hatóköri viszonyok és a referencialitás meghatározása a szórend és az intonáció alapján, miközben ez a folyamat a megbízható információkinyerésben és a minőségi gépi fordításban kulcsfontosságú. Célunk egy ilyen rendszer építése – először – a magyarra alapozva, melynek kapcsán ismert, hogy a hatóköri relációkat (az ige előtti operátorzónában) explicit módon kifejezi, ahogy a referencialitást is viszonylag magától értetődő módon. Szintén elkezdtük a megközelítésünket más nyelvekre (pl. angol) is kiterjeszteni, hangsúlyozva a hasonlóságokat, különbségeket és megfeleléseket például a magyar szórendet és az angol argumentumstruktúrát változtató operációk között.

Hogy elérhessük ezt a célt a mi totálisan lexikalista megközelítésünkben, az implementációban szükség van egy kétrétegű lexikonra. Ez áll egy maglexikonból (amely a morfémák alapértelmezett viselkedését tartalmazza – pl. az igékét) és egy kiterjesztett lexikonból, amelynek elemeit egy általános lexikaiszabálymodullal generáljuk. A maglexikon elemei felelnek a semleges mondatok elemzéséért, míg a kiterjesztett lexikonban kezeljük a különböző topik-, fókusz- és kvantorkonstrukciókat, amelyekben a referencialitási követelmények is gyakran eltérnek a semleges mondatoktól.

Így elsődlegesen semleges és nem-semleges mondatokat generálunk a kétrétegű lexikon elemeiből, és – erre a generálásra alapozva, a Prolog-környezet keretei között – létrehozzuk az (intonált és nem intonált) mondatok információstruktúráját.

Irodalom

Alberti, Gábor 1997. Restrictions on the degree of referentiality of arguments in Hungarian sentences. Acta Linguistica Hungarica 44: 341–362.

Alberti, Gábor – Márton Károly – Judit Kleiber 2010. The ReALIS model of human interpreters and its application in computational linguistics. In: José Cordeiro – Maria Virvou – Boris Shiskov (szerk.): Proceedings of ICSOFT 2010, 5th International Conference on Software and Data Technologies, Athens, Greece, vol. 2. Portugal: SciTePress. 468–74.

Alberti, Gábor – Judit Kleiber 2003. Extraction of discourse-semantic information from Hungarian sentences by means of a totally lexicalist grammar. In: Hamish Cunningham – Elena Paskaleva – Kalina Bontcheva – Galia Angelova (szerk.): Information extraction for Slavonic and other Central and Eastern European languages. Borovets, Bulgaria: RANLP. 63–69.

Alberti, Gábor – Judit Kleiber 2004. The GeLexi MT Project. In: John Hutchins (szerk.): Proceedings of EAMT 2004 Workshop (Malta). Valletta: Univ. of Malta. 1–10.

Alberti, Gábor – Judit Kleiber 2010. The grammar of ReALIS and the implementation of its dynamic interpretation. Informatica 34: 103–110.

Alberti, Gábor – Judit Kleiber – Anita Viszket 2004. GeLexi project: Sentence parsing based on a GEnerative LEXIcon. Acta Cybernetica 16: 587–600.

Alberti, Gábor – Anna Medve 2000. Focus constructions and the "scope–inversion puzzle" in Hungarian. In: Gábor Alberti – István Kenesei (szerk.): Approaches to Hungarian 7: Papers from the Pécs conference. Szeged: JATEPress. 93–118.

Croft, William 2001. Radical construction grammar: Syntactic theory in typological perspective. Oxford: Oxford University Press.

É. Kiss, Katalin 1995. The definiteness effect revisited. In: Kenesei (1995, 63–88).

É. Kiss, Katalin 2002. The syntax of Hungarian. Cambridge: Cambridge University Press.

Enç, Mürvet 1991. The semantics of specificity. Linguistic Inquiry 22: 1–25.

Hunyadi, László 2002. Hungarian sentence prosody and Universal Grammar. Frankfurt am Main: Peter Lang.

Jong, Franciska de – Henk J. Verkuyl 1984. Generalized quantifiers: The properness of their strength. In: Johan van Benthem – Alice ter Meulen (szerk.): GRASS 4. Dordrecht: Foris. 21–45.

Kálmán, László 1995. Definiteness effect verbs in Hungarian. In: Kenesei (1995, 221–242).

Kenesei, István (szerk.) 1995. Approaches to Hungarian 5: Levels and structures. Szeged: JATEPress.

Kiefer Ferenc (szerk.) 1992. Strukturális magyar nyelvtan 1. Mondattan. Budapest: Akadémiai Kiadó.

Szabolcsi Anna 1992. A birtokos szerkezet és az egzisztenciális mondat. Budapest: Akadémiai Kiadó.

Szabolcsi, Anna 1997. Strategies for scope taking. In: Anna Szabolcsi (szerk.): Ways of scope taking (SLAP 65). Dordrecht: Kluwer. 109–154.

Szécsényi Tibor 2009. Lokalitás és argumentumöröklés. A magyar infinitívuszi szerkezetek leírása HPSG keretben. Dotori értekezés, Szegedi Tudományegyetem, Szeged.

Traat, Maarika – Johan Bos 2004. Unificational Combinatory Categorial Grammar: Combining information structure and discourse representations. In: Proceedings of the 20th International Conference on Computational Linguistics. Genova.

The implementation of the Hungarian scope hierarchy and degree of referentiality

Abstract: As our team of ReALIS strives for sophisticated machine translation and reliable information extraction, we have launched a subproject aiming at the revelation of reference and information structure in Hungarian declarative sentences. The crucial part of information extraction is a procedure with a sentence as its input and an information structure as its output, which is practically a set of possible operator scope orders (acceptance). A similar procedure forms the first half of machine translation, too: first the information structure of the source-language sentence should be calculated; and then an opposite procedure should take place (generation), whose input is an information structure, and whose output is an intoned word sequence, that is, a sentence in the target language. The procedure of acceptance thus is based upon that of generation. And as our approach to grammar is "totally lexicalist", the lexical description of verbs is responsible for the order and intonation of words in the generated sentence.

Keywords: lexicalist grammar, operator scopes (focus, quantifier, topic, contrastive topic), intonation, word order, reference

Tudásalapú koreferencia- és birtokviszony-feloldás magyar szövegekben

Mihálty Márton

Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Budapest
mmihaltz@gmail.com

A főnévi csoportok közötti koreferenciaviszonyok azonosítása egy szövegben levő, ugyanarra a való világbeli entitásra hivatkozó kifejezések azonosítását jelenti. A cikkben bemutatott számítógépes rendszer a következő nyelvi jelenségek kezelésére tesz kísérletet: koreferencia kifejezése ismétléssel, tulajdonnév-változatokkal, szinonimákkal és hiperonimákkal/hiponimákkal, személyes névmások és zéró névmások. A birtokviszony-feloldás a koreferencia azonosításához hasonló feladat, amelynek célja az elvált, akár más mondatrészek közbeékelődésével egymástól távolra került birtokos–birtok párok azonosítása. A bemutatott, szabályalapú rendszer többféle tudásra támaszkodik: a MetaMorpho gépi fordítórendszer magyar mély nyelvi elemzőjének kimenetében található morfológiai, szintaktikai és szemantikai információkra; a kormányzás és kötés elméletének a magyar szintaxiselméletben megfogalmazott változatára, továbbá a magyar mondatmegértés pszicholingvisztikájában elért kutatási eredményekre támaszkodó szabályokra; a Magyar WordNetben található szemantikai tudásra; valamint karakteralapú heurisztikákra.

Kulcsszavak: anaforafeloldás, koreferenciafeloldás, birtokviszony-feloldás, tudásalapú, szabályalapú

1. Bevezetés

Természetes nyelvű szövegek gépi feldolgozásában fontos segítség lehet a szövegbeli entitások közötti kapcsolatok – koreferenciaviszonyok, birtokviszonyok – automatikus felismerése. A feladat megoldása célszerű lehet olyan nyelvtechnológiai alkalmazások számára, mint a gépi fordítás, az információ-kivonatolás, a szöveg-összefoglalás, a véleményanalízis és egyéb szövegfeldolgozó alkalmazások (Mitkov 1999).

NP-koreferenciák feloldásán egy adott dokumentumban eltérő pontokon megjelenő, de azonos entitásra referáló főnévi csoportok (NP-k) közötti viszonyok azonosítását értjük.[1] A birtokviszony-feloldás a szöveg különböző pontja-

[1] A magyar formális nyelvészeti szakirodalomban a DP (determinánsi csoport) megnevezés használatos, mi itt mégis a nemzetközi koreferenciakutatás (számítógépes nyelvészeti) irodalmában használatos NP terminust használjuk.

in megjelenő, egymással birtokviszonyban lévő NP-k – elvált birtokos szerkezet birtokosa és birtoka – felismerését és párosítását jelenti (erről bővebben a 3. részben lesz szó.)

A következő részben bemutatjuk szabályalapú megközelítést alkalmazó koreferenciafeloldó rendszerünket, majd a kiértékelésére kialakított, annotált korpuszokat használó környezetet. A 3. részben külön tárgyaljuk ezzel rokon működésű, de a távoli birtokosok és birtokok közötti viszonyokat azonosító megoldásunkat és annak kiértékelését. Az utolsó részben részletesen ismertetjük a további lehetséges fejlesztések irányát.

2. Koreferenciafeloldás

A bemutatott rendszerben az alábbi koreferenciajelenségek kezelésére – a visszautaló elem (anafora) és a szövegben korábban előforduló, vele koreferens NP (antecedens) közötti kapcsolat azonosítására – tettünk kísérletet magyar nyelvű szövegekben (Pléh 1998 alapján):

Típus	Példa
Ismétlés	A kislány meghúzta **a macska** farkát. **A macska** keservesen nyávogott.
Tulajdonnév-variáns	**Kovács Jakab** tegnap sajtótájékoztatót tartott. Az eseményen **Kovács úr** bejelentette az új termékeket.
Szinonima	Kázmér kapott **egy biciklit**. Én is láttam **a kerékpárt**.
Hipero-/hiponima	Találtam **egy kutyát**. **Az állat** elvesztette a gazdáját.
Névmás	Beszéltem **Julival**. Megadtam **neki** a számodat.
Zéró névmás	**Helga** ismeri **Hubát**, de (**ő**) nem kedveli (**őt**) túlságosan.

1. táblázat. A vizsgált koreferenciatípusok, példákkal (az egymással koreferens NP-k félkövérrel kiemelve)

Nem foglalkoztunk a személyes névmáson és bizonyos mutató névmásokon kívüli egyéb névmástípusok (visszaható és kölcsönös névmások, vonatkozó névmások stb.) feloldásával. Főnévi csoportokon a mondatban előforduló maximális NP-ket értjük, amelyek jellemzően a mondat főigéjének vonzatai, illetve főnévi eredetű szabad határozói. Nem foglalkoztunk a komplex, hierarchikus szerkezetű főnévi csoportok (koordinált NP-k, appozíciós szerkezetek stb.) összetevőivel. Szintén nem foglalkoztunk az anaforát a szövegben követő antecedensű koreferenciatípus (katafora), valamint az epithetonnak nevezett jelenség *(**Balázs** nem találta a kulcsát. **A szerencsétlen** nem tudott bejutni a lakásba.)* kezelésével.

2.1. Módszerek

Az anaforafeloldás korai megoldásai tudásalapú megközelítést alkalmaztak: szabályalapú, algoritmikus megoldásokat, amelyek az anaforikus jelenségekkel kapcsolatban megfigyelt heurisztikákat implementálták. Az egyik legelső ilyen kísérlet, Hobbs (1977) naiv algoritmusa csupán (az inputnak nyelvtani elemzővel történő feldolgozása révén rendelkezésre álló) szintaktikai információkat használt. Az antecedensjelölteket az elemzési fákban kereste meg, és a szám, nyelvtani nem egyezését, valamint a kötéselmélet (l. később) megszorításait vizsgálta. A diskurzusalapú módszerek ezzel szemben elsősorban a centering-elmélet (Grosz et al. 1995) eredményeit használták fel, amely a figyelem középpontjában álló entitások követését igyekszik modellezni. Brennan, Friedman és Pollard munkája (1987) a legismertebb ezek közül. Lejtovicz és Kardkovács (2006) magyar nyelvű szövegekben történő anaforafeloldáshoz használta a BFP-algoritmust.

Lappin és Leass (1994) munkája már a több tudásforrás (szintaxis, szemantika, morfológia, diskurzus) integrációján alapuló megközelítések közé tartozik (Tejaswini 2004). Algoritmusuk az antecedensjelölteket egyrészt a szám- és nembeli egyeztetés, valamint a kötéselméleti szabályok alapján szűrte, másrészt többféle szempont (a távolságra, illetve különböző strukturális felállásokra alapozó, súlyozott heurisztikák) alapján rangsorolta.

Az újabb keletű, felügyelt gépi tanulásra támaszkodó, adatalapú módszerek a tudásalapú megközelítéssel szemben „tudásszegények", csupán megfelelő számú annotált tanítópéldára van szükségük egy osztályozó betanításához (Ng 2005). Ebben a paradigmában a koreferenciafeloldás során a bináris osztályozó minden egyes visszautaló elem esetében megvizsgálja az összes, azt megelőző, szóba jöhető antecedenst, és egyenként hoz döntést arról, hogy van-e közöttük koreferenciaviszony vagy sem. Soon et al. (2001) mutatott be elsőként ilyen rendszert, amelynek teljesítménye összevethető volt a korábbi tudásalapú megközelítésekével. Az osztályozó a C5.0 döntésifa-tanuló algoritmust használta, amelyet a tanítókorpuszból előállított, 12 különböző jegyet tartalmazó jegyvektorokkal tanítottak (pl. az NP típusa, távolság az anaforikus elem és a jelölt között, számbeli, nembeli és szemantikai kategóriabeli egyezés, tulajdonnevek és appozíciós szerkezetek stb.) Egy újabb munkában Uryupina (2006) már 351 különféle, különböző tudásforrásokon (karakteralapú távolság, szintaktikai, szemantikai és diskurzusszintű jelenségek) alapuló, nyelvészetileg is motivált jegyet használ, 5 különböző gépi tanuló algoritmus vizsgálatával.

A munka kezdetekor nem állt rendelkezésünkre az adatalapú megközelítésekhez elengedhetetlen, jellemzően több ezer, kézzel annotált példából álló tanítókorpusz a magyar nyelvre, így tudásalapú megközelítést alkalmaztunk.

Rendszerünk többféle tudásra támaszkodik. A legfontosabb inputot a MetaMorpho fordítóprogram-projektben fejlesztett magyar mondatelemző (Prószéky et al. 2004) elemzéseiben kapott morfológiai, szintaktikai és szemantikai jegyek, nyelvtani szerepek, mélyszerkezeti elemzési struktúrák stb. jelentik. Ezekre támaszkodnak a kötéselmélet (Kenesei 1992) és a magyar mondatmegértés kutatásainak (Pléh–Radics 1976; Pléh 1998) eredményeire támaszkodó szabályaink. További, világismereti tudáson alapuló szabályok forrásaként a Magyar WordNet ontológiát (Miháltz et al. 2008; Prószéky–Miháltz 2008) használjuk. Végül a tulajdonnevek közötti referenciaazonosság felismeréséhez karakteralapú megközelítéseket alkalmazunk (Uryupina 2004).

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

A koreferenciafeloldás a teljes input dokumentum nyelvi elemzésével kezdődik. A bekezdésekre tagolt szöveg mondatainak mindegyikéhez a MetaMorpho elemzővel előállított szintaktikai fák egyszerűsített változatát rendeljük, amelyek a gyökércsomópont alatt csak a (fő, alá- és mellérendelt) tagmondatoknak, a maximális igei frázisoknak (VP) és a főnévi csoportoknak (NP) megfelelő csomópontokat tartalmazzák. A szintaktikai elemző gyakran (főként hosszabb, összetett mondatok esetében) nem képes teljes, a mondat minden szavát lefedő elemzési fát előállítani, ilyenkor a rendelkezésre álló részelemzéseket használjuk fel (VP-k, NP-k, illetve főnévi eredetű határozói csoportok (AdvP-k)). Az azonosított főnévi csoportokban 25 jegy reprezentálja a MetaMorpho segítségével meghatározott lexikai, morfológiai, szintaktikai és szemantikai tulajdonságokat.

A nyelvi előfeldolgozást követi a koreferenciaviszonyok feldolgozása, amely az antecedensjelölteket szűrő megszorítások és a fennmaradó jelöltek közül választó preferenciák módszerén alapul (Mitkov 1999). A módszer minden egyes lépése a feloldandó anaforikus elem típusától (tulajdonnév, határozott névelős köznév vagy (zéró) névmás) függő szabályokat tartalmaz. Az általános algoritmus a következő:

1. *Előszűrés*: az anaforikusnak feltételezett, tovább feldolgozandó NP-ket azonosítjuk. A jelenleg nem kezelt visszautaló elemek mellett próbáljuk felismerni és kizárni azokat a formailag visszautaló, azonban valójában a szövegből kiutaló, tehát szövegbeli előzménnyel nem rendelkező NP-ket is, amelyeknek a további feldolgozása zajként jelentkezne (Varasdi 2005). Ebbe a lépésbe beépítettünk öt olyan heurisztikát is, amelyeknek a célja a nyelvi elemző ál-

tal nagy valószínűséggel hibásan elemzett, így a koreferenciafeloldásban is szükségképpen hibát okozó NP-k felismerése és kizárása.

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

3. *A jelöltek szűrése:* ebben a lépésben az antecedensjelöltek közül megpróbálunk minél többet kizárni (a konkrét módszer az anafora típusától függ, l. később), illetve a jelöltekre is alkalmazzuk az 1. lépésben ismertetett, elemzési hibákat felismerő heurisztikákat.

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

Az alábbiakban ismertetjük az algoritmus konkrét lépéseit a különböző anaforatípusok esetében.

2.1.1. Tulajdonnevek

Az antecedensjelöltek listázásának hatóköre a teljes megelőző dokumentum: az összes tulajdonnévi NP-t hozzáadjuk a listához az anaforát tartalmazó VP kezdetéig. Ezek között nem alkalmazunk előszűrést (minden, a szövegben előforduló, MetaMorpho által azonosított tulajdonnevet feldolgozunk).

Az anafora és az antecedensjelölt normalizálása (a kezdő determinánsok elhagyása, a fej tövesítése) után kiszámítjuk közöttük a Levenshtein-távolságot (Uryupina 2004), amelyet a hosszabbik string hosszával normalizálunk. Az algoritmusnak nem kötelező az antecedensjelöltek közül választania, így az anaforához legjobban hasonlító (a legkisebb Levenshtein-távolságot mutató) antecedensjelöltet csak azok közül a jelöltek közül választjuk ki, amelyek egy paraméterben meghatározott küszöbérték alatti hasonlóságot mutatnak (amennyiben a lista nem üres.)

2.1.2. Határozott névelős köznevek

Előszűrésként a „szemantikus NP"-knek (Varasdi 2005) nevezett, közös világismeretből azonosítható egyedi objektumokra referáló, tehát a szövegben antecedenssel nem rendelkező határozott névelős közneveket próbáljuk meg felismerni és kizárni a feloldás alól (pl. *az amerikai elnök*). Ehhez egy külön, előre összeállított listát használunk. Az antecedensjelöltek a tulajdonnevek és köznevek (a determináns típusától függetlenül) az anaforát megelőző teljes bekezdésben, az anafora VP-jéig.

Az antecedens kiválasztása úgy történik, hogy a jelöltek közül meghatározzuk az anaforához legközelebb eső, vele azonos fejű NP-t (ismétlés), vagy szinonimát, vagy hipo-/hiperonimát.

A szinonimitás vizsgálatához mind az anafora, mind az antecedensjelölt lehetséges jelentéseit kikeressük a Magyar WordNetben, és ha van olyan synset, ami mindkettőt tartalmazza, szinonimáknak tekintjük őket. Mivel nincs jelentésegyértelműsítés, a módszer nyilvánvalóan nem lesz minden esetben helyes.

Az anafora és az antecedensjelölt közötti hiperonimaviszony meghatározására a Leacock–Chodorow szemantikai hasonlósági képletet alkalmazzuk (Leacock–Chodorow 1998), amely a visszautaló és a jelölt összes WordNet-beli megfelelőit összekötő, hiperonimareláció szerinti útvonalak közül a legrövidebb alapján számít ki egy, az útvonal hosszától függő pontértéket. Hiperonima/hiponima jelölteket csak az anaforát megelőző mondatban fogadunk el akkor, ha a Leacock–Chodorow hasonlósági függvény értéke meghaladja egy előre beállított küszöb értékét, és csak akkor, ha nem találunk azonos fejű vagy szinonim antecedenst a bekezdésben. Jelentésegyértelműsítés hiányában a lexikális többértelműségek nyilvánvalóan itt is fognak hibákat okozni.

2.1.3. Névmások

Csak (a mondatelemző által azonosított, formailag NP-ként reprezentált) zéró névmásokkal, személyes névmásokkal, valamint az *az* mutató névmással foglalkozunk – feltéve, hogy utóbbi a VP-jében alanyi szerepben áll, és nem egy alárendelt tagmondatra utal. Nem foglalkozunk az első, illetve második személyű, ún. deiktikus névmásokkal és zéró névmásokkal (Pléh–Radics 1976).

Antecedensjelölteknek az anafora mondata előtti második mondattól kezdve (ha létezik ilyen a bekezdésben) választjuk ki az összes NP-t, az anaforát tartalmazó tagmondat határáig. Ezeket aztán megszűrjük úgy, hogy az anafora és az antecedensjelölt számának, személyének és két szemantikai jegyének (+/– élő, +/– ember) egyezését vizsgáljuk. Utóbbiak értéke lehet alulspecifikált (zéró névmások, illetve az elemző szótárában többértelmű főnevek esetében), ezek minden lehetséges értékkel kompatibilisek.

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

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

A (tag)mondatukban alanyi szerepű névmási vagy zéró névmási visszautalók antecedensének meghatározásában Pléh Csaba és munkatársainak a magyar mondatmegértés pszicholingvisztikája körében végzett kutatási eredményeire támaszkodtunk (Pléh 1998). A heurisztika a szerkezeti párhuzamosság feltételezéséből indul ki, amely szerint az alanyi helyzetű anafora az előzménymondat alanyára utal vissza. Ezt felülbírálhatja az alanyi szerepben álló *az* mutató névmás, amely alanyváltást jelöl:

(1) a. Hugó$_j$ felhívta Amáliát$_k$. (Ő$_j$) elmondta nekik a történetet.

b. Hugó$_j$ felhívta Amáliát$_k$. Az$_k$ elmondta neki$_j$ a történetet.

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

Az formájú alany esetén, ha az előzménymondatban több, nem alanyi szerepű antecedensjelölt NP is található, az alábbi szabályok alapján választunk:

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

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

Nem alanyi pozícióban álló névmások, zéró névmások esetén több, az alannyal nem koreferens antecedensjelölt közül szintén a fenti két szabály alkalmazásával választunk.

A koreferenciafeloldást először minden mondatban a tulajdonnevekre, majd a határozott névelős köznevekre végezzük el, ez után következik a mondat névmási, zéró névmási anaforáinak feldolgozása. Reményeink szerint ezzel további segítséget adunk a névmási anaforák feloldásához a szűrési feltételekben leírt szabály alkalmazásával.

2.2. Kiértékelés

A koreferenciafeloldó rendszer pontosságának kiértékeléséhez létrehoztunk egy kézzel annotált kiértékelő korpuszt, amely 10 darab, általános iskolai történelemkönyvekből kiemelt szövegrészletet tartalmaz (2. táblázat). A szövegekben a

MetaMorpho segítségével azonosítottuk a maximális NP-ket, majd manuálisan annotáltuk közöttük a koreferenciaviszonyokat. Az automatikus annotációhoz hasonlóan a koreferencialáncokban mindig az anaforához legközelebbi antecedenseket jelöltük be. A munkát egyetlen annotátor végezte.

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

Szövegek száma	10
Bekezdések száma	31
Mondatok száma	99
NP-k száma	488
Antecedenssel annotált NP-k száma	111

2. táblázat. A kiértékelő korpusz jellemzői

A korpuszban 14 különböző fajta koreferenciajelenséget annotáltunk manuálisan, ez összesen 111 különböző NP-t érintett. Az automatikus koreferenciaannotáló rendszer ebből 5 különböző fajta koreferenciajelenséget kezel, ez 81 különböző NP-t jelentett. A 3. táblázatban bemutatjuk a különböző NP-koreferenciatípusok eloszlását a korpuszban.

Koreferenciatípus	Előfordulások száma
Személyes névmás	**47**
Ismételt NP	**15**
Tulajdonnév-változat	**14**
Mutató névmás	8
Frame	7
hogy-os mellékmondat	6
Hiperonima	**3**
Vonatkozó névmás	5
Szinonima	**2**
Appozíció	1
Kopula	1
Hiponima	**1**
Meronima	1
Holonima	0
Összesen:	111

3. táblázat. A kiértékelő korpuszban annotált koreferenciatípusok (félkövérrel a bemutatott rendszer által felismert típusok)

A koreferenciafeloldó algoritmust ezután lefuttattuk a korpusz szövegein, majd összevetettük az automatikus annotáció eredményeit a kéziével. Azokat az esete-

ket is helyesnek fogadtuk el, amikor a rendszer által jelölt antecedens nem egyezett meg a kézzel azonosítottal, de a kettő ugyanabba a koreferencialáncba tartozott (koreferens volt). A 4. táblázat bemutatja a különböző visszautalási típusok felismerésének pontosságát külön-külön is.

Visszautalási típus	Manuálisan annotált NP	Automatikusan annotált NP, összesen	Automatikusan, helyesen annotált NP	Pontosság (%)
Tulajdonnév	14	15	12	80,00%
Névmás	46	35	25	71,43%
Ismétlés	15	18	13	72,22%
Szinonima	2	4	1	25,00%
Hiperonima	4	2	0	0,00%
Összesen/Átlag:	81	74	45	68,92%

4. táblázat. A különböző koreferenciafeloldó módszerek pontossága

Az eredmények első ránézésre biztatónak tűnnek, hiszen a rendszer által kezelt leggyakoribb anaforatípusokra (tulajdonnevek, névmások, ismétlés) a pontosság 71–80% közötti értéket mutat. A szinonima- és hiperonima-heurisztikák teljesítménye ugyanakkor nagyon gyenge, de mivel a kiértékelő korpusz csak nagyon kevés ilyen típusú példát tartalmazott, nem biztos, hogy a kiértékelésük megbízható eredményt adott.

Kíváncsiak voltunk arra is, hogy mennyiben befolyásolja a nyelvi elemző a koreferenciafeloldás teljesítményét, ezért részletesen megvizsgáltuk a rendszer által elkövetett hibákat. Az automatikusan azonosított antecedensek mindegyikét az alábbi címkék egyikével láttuk el (5. táblázat):

(a) *OK:* a rendszer által megjelölt antecedens megegyezik a manuális annotációval

(b) *OK_equ*: a rendszer által megjelölt antecedens nem egyezik meg a manuális annotációval, de ugyanarra az entitásra referál (a koreferencialánc egy másik eleme), tehát helyesnek tekinthető

(c) *KO_parser:* a rendszer által megjelölt antecedens nem egyezik meg a manuális annotációval, és a hiba a helytelen nyelvi elemzés következménye (ha az elemző helyes eredményt adott volna, az automatikusan azonosított antecedens is helyes lenne)

(d) *KO_cr:* az anaforának volt a szövegben antecedense, és a nyelvi elemzés is helyes volt, de a koreferenciafeloldó algoritmus helytelenül (vagy nem) azonosította az antecedenst.

A táblázatból látható, hogy az automatikus annotáció hibáinak körülbelül a fele a nyelvi elemző hibáinak (hibás NP-határok, hibásan felismert zéró névmások

Koreferencia típusa	OK	OK_equ	KO_parser	KO_cr
Névmás	19	6	7	3
Ismétlés	13	0	4	1
Tulajdonnév	12	0	0	3
Hiperonima	0	0	0	2
Szinonima	1	0	0	3
Hiponima	0	0	0	0
Összesen:	45	6	11	12

5. táblázat. A koreferenciafeloldás hibáinak besorolása

stb.) következménye. Hibák nélküli szintaktikai elemzésre támaszkodva a koreferenciafeloldás átlagos pontossága a jelenlegi algoritmussal 75%-os értéket érne el, a névmások/zéró névmások feloldási pontossága 91%-ot.

A kiértékelő halmazban kézzel annotálttal nem teljesen megegyező, de referenciálisan ekvivalens antecedenseket csak a névmási anaforák feloldásában találhatunk. Ha a koreferencialáncokat végigkövetnénk a legelső tagjukig, hogy minden esetben a legelső említett koreferáló entitás legyen az antecedens, akkor a pontosság csökkenne (a koreferenciafeloldás és a parser hibái miatt.)

3. Birtokviszony-feloldás

Ebben az esetben a rendszer feladata az összetartozó birtokosoknak és birtokoknak megfelelő kifejezések közötti viszonyok azonosítása a szövegben. Ezen belül olyan esetekre koncentrálunk, ahol a birtokosnak és a birtoknak megfelelő NP-k közé egyéb mondatrészek, vagy akár mondathatárok kerültek.

3.1. Módszerek

Háromféle birtokos szerkezetről beszélhetünk, melyekben a birtokos és a birtok elválhatnak egymástól (a birtokosnak és a birtoknak megfelelő NP-k a példákban kövérrel kiemelve):

(2) a. *Birtoklásmondat:*
Jánosnak van egy nagy, sárga **esőkabátja**.

b. *Elvált datívuszos birtokos (topikalizáció stb.):*
Jánosnak ellopták a **könyvét.**

c. *Zéró névmási birtokos:*
János tegnap itt hagyta az (**ő**) **esernyőjét**.

A (2a) típusú mondatokban a létige egy speciális predikátum, amely birtokviszonyt fejez ki az argumentumai között. A (2b) típusú mondatokban a birtokos szerkezet egy komplex NP, melyben egyéb mondatrészek kerülhetnek a birtokos és a birtoka közé. A (2c) típusú mondatok inkább diskurzusszintű jelenségek, amikor is a hallgatónak a korábban bevezetett entitások közül kell kiválasztania a birtokost.

A bemutatott koreferencia- és birtokosazonosító rendszer parser komponense, a MetaMorpho mély szintaktikai elemző nyelvtana képes – amennyiben az elemzés teljes – kezelni a (2a) és (2b) típusú jelenségeket, így ilyenkor a szintaktikai elemzésben rendelkezésre álló birtokviszony-pointereket használjuk a rendszer kimenetében.

A (2c) típusú mondatokban, az alanyesetű, elvált birtokos azonosítására alkalmazott módszer rokon a névmási anafora feloldására alkalmazott módszerünkkel. Feltételezzük, hogy

(i) a birtokot dominálló VP alanyi vonzata az alapértelmezett birtokos,

(ii) a birtokos számban és személyben egyezik a birtokon azonosított birtokos jel számával és személyével.

A második feltételezés felülbírálhatja az elsőt, vagyis ha a birtokos VP-jének alanya nem egyezik számban és személyben, akkor az előző (tag)mondat alanya töltheti be a birtokos szerepét, feltéve, hogy ugyanabban a diskurzusszegmentumban (bekezdésben) van:

(3) **János** elutazott nyaralni. Én vigyázok a **lakására**.

A rendszernek természetesen figyelembe kell vennie, hogy a magyarban egyes szám harmadik személyű birtokos jellel rendelkező birtok is tartozhat többes számú birtokoshoz, pl. *Ádám almája, a lányok almája*.

A fentiek fényében a birtokviszony-feloldás algoritmusa a (2c) típusú birtokos szerkezetek esetében a következő:

1. Azonosítjuk a birtok NP-t tartalmazó (tag)mondat előtti (tag)mondatokban található, alanyi szerepű NP-ket, legfeljebb két megelőző (tag)mondatnyi távolságban, de nem lépve messzebb az aktuális bekezdés legelső mondatánál.

2. A birtokon azonosított birtokos jel számával és személyével egyező számú és személyű, fent kiválasztott NP-k közül azt választjuk ki birtokosnak, amelyik a birtok NP-hez legközelebb esik (a „legjobboldalibbat".)

3. Amennyiben nem áll rendelkezésünkre teljes szintaktikai elemzés, vagyis információ az NP-k nyelvtani szerepéről a MetaMorpho elemző kimenetében, akkor a birtok előtt álló, számban és személyben egyező, alanyesetű NP-k közül választjuk ki a legjobboldalibbat.

3.2. Kiértékelés

A távoli birtokosokat azonosító algoritmus kiértékeléséhez ugyanazt a korpuszt vettük alapul, amit a koreferenciafeloldás kiértékeléséhez is használtunk. A 10 szövegrészletben található 488 NP közül 38 volt elvált birtokos szerkezet. Ezeken a birtok NP-ken manuális annotációval jelöltük a birtokosaiknak megfelelő NP-k azonosítóit.

Ezen a korpuszon lefuttattuk a rendszert, a kimenetét összehasonlítottuk az annotált referenciaértékekkel, és megállapítottuk a következő értékeket:

- valódi pozitívok: az NP tartalmaz mind manuálisan, mind gépileg meghatározott birtokosazonosítót, és ezek értéke megegyezik
- hibák: az NP tartalmaz mind manuálisan, mind gépileg meghatározott birtokosazonosítót, de ezek értéke eltér
- hamis pozitívok: az NP csak automatikusan meghatározott birtokosazonosítót tartalmaz
- hamis negatívok: az NP csak manuálisan meghatározott birtokosazonosítót tartalmaz

Ezek alapján a következőképpen határoztuk meg a pontosságot és a fedést:

(4) pontosság = |valódi pozitívok| / (|valódi pozitívok| + |hibák| + |hamis pozitívok|)
fedés = |valódi pozitívok| / (|valódi pozitívok| + |hibák| + |hamis negatívok|)

Pontosságot, fedést és F-mértéket (utóbbit a szokásos módon, a pontosság és a fedés harmonikus közepeként számítva) a rendszer átlagos teljesítményére határoztunk meg, valamint külön pontosságértékeket az egyes birtokviszony-feloldó módszerekhez (fedést a módszerekhez külön-külön a korpuszban hiányzó ilyen adatok miatt nem lehetett meghatározni). Az eredmények összefoglalása a 6. táblázatban látható.

Ahogy a 6. táblázatból látszik, a kiértékelő korpusz nem tartalmazott egyetlen olyan (2a) típusú példát sem, amit a birtokviszony-azonosító rendszer megpróbált volna feldolgozni. A (2b) típusú elvált birtokos szerkezetek mindegyikét – a nyelvtani elemző kimenetének felhasználásával – helyesen oldotta fel a rendszer (100% pontosság). A (2c) típusú példák feldolgozásának pontossága, a fent vázolt algoritmus segítségével 71,43% volt, a teljes rendszer teljesítménye így 76,47%-os pontosságot ért el, 68,42%-os fedés mellett.

Az elvált birtokosok azonosítása – a koreferencia azonosításához hasonlóan – nagymértékben a nyelvtani elemző eredményére támaszkodik, így érzékeny annak hibáira. Vannak azonban olyan esetek is, amikor a fenti algoritmus hibázik, mint az alábbi, iskolai történelemkönyvi szövegből származó, könnyűlovas magyar harcosok harci taktikáját bemutató szövegrészletben:

	(2a) típus (parser)	(2b) típus (parser)	(2c) típus (szabályok)	Átlag
Valódi pozitívok	0	6	20	26
Hibák	0	0	7	7
Hamis pozitívok	0	0	1	1
Hamis negatívok	—	—	—	5
Pontosság	0	100,00%	71,43%	76,47%
Fedés	—	—	—	68,42%
F-mérték	—	—	—	72,22%

6. táblázat. A birtokviszony-feloldás kiértékelése

(5) […] az első, ellenséggel való összecsapás után menekülést színlelve megfordultak, és futásnak eredtek. **Az ellenfél** ekkor üldözőbe vette őket. Ez lett a **vesztük**. A harcosok a vágtató ló hátán „kengyelbe állva", hátrafordulva lenyilazták üldözőiket.

A kérdéses (félkövérrel kiemelt) entitások közötti birtokviszony helyes értelmezéséhez szükségünk van kiegészítő világismeretre. Mivel a birtokos száma (egyes szám) nem egyezik a birtokon található birtokos jel számával (többes szám), szükségünk van arra a tudásra, hogy a birtokosnak megfelelő egyes számú NP jelölete ebben a kontextusban – csatában az ellenfél – valójában egy több személyből álló csoport, tehát lehet rá utalni többes számú alakkal. A probléma egy megoldása lehet, ha az ilyen fajta tudás is bekerül a rendszer munkáját segítő ontológiába.

4. Összegzés, további lehetséges munka

A továbbiakban érdemes lenne egy baseline megoldásnak megfelelő algoritmust implementálni, amelyhez képest meghatározható a rendszerünk teljesítménye. Ehhez a centering-elméleten alapuló, a szakirodalomban jól ismert BFP-algoritmust (Brennan et al. 1987) szeretnénk kipróbálni, amelyet vizsgáltak már magyar szövegekkel is (Lejtovicz–Kardkovács 2006).

Szeretnénk létrehozni egy olyan, koreferenciával annotált kiértékelő korpuszt, amely mások számára is hozzáférhető, így a rendszerünk teljesítménye más hasonló rendszerekkel is összevethető lesz. Ehhez legalkalmasabbnak a frázisannotációkat tartalmazó Szeged Treebank 2.0 változata tűnik (Csendes et al. 2005). A Szeged Treebank használatával nyelvi elemzőtől független, nagy pontosságú szintaktikai elemzésekre lehetne koreferenciafeloldó algoritmusokat építeni

(ugyanakkor bizonyos jegyek, amelyek a MetaMorpho kimenetében azonosíthatók, nem lesznek elérhetők).

Az anaforafeloldás fedésének növelésére további főnévi anaforikus jelenségek kezelésére lesz szükség: visszaható és kölcsönös névmások, vonatkozó névmások, birtokos névmások, valamint a komplex NP-k részegységeinek elemzésére és koreferenciakapcsolataik feltárására. Mutató névmások kezelésénél külön problémát jelent az entitásokra és az eseményekre utaló esetek felismerése (pl. *Tegnap láttam Ákost Noémivel.* ***Ezen*** *nagyon meglepődtem.*)

A pontosság növelése érdekében a tulajdonnevek felismerésére további karakterhasonlóságon alapuló módszereket és normalizációs eljárásokat mutat be Uryupina (2004). A karakteres és a szemantikai hasonlóságképletek számára a küszöbértékeket empirikus úton, korpuszpéldák segítségével lenne célszerű optimalizálni.

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

Határozott névelős közneveknél felmerül a kérdés, hogy az azonos fejű, számban egyező, de eltérő módosítókat tartalmazó párokat mikor tekintsük koreferensnek és mikor nem, pl. *a katonák – az út szélén elrejtőzött katonák,* vö. *az első jelentkező – a második jelentkező*.

A névmási anaforák kezelésénél további heurisztikák alapja lehet a centering-elmélet, a diskurzustopik változásának figyelése (Brennan et al. 1987), illetve a Pléh (1998) által leírt egyéb jelenségek modellezése (pl. a predikátum által preferált vonzatok korpuszstatisztikai vizsgálata.)

További lehetőség a zajt okozó, feloldást nem igénylő NP-k azonosítása további módszerekkel. Az egyik a szükségszerű/valószínű rész viszony, pl. *Tegnap szerelőhöz vittem a* ***biciklim****, mert eltört a* ***pedál***. Ha rendelkezésre állna megfelelő adatbázis, az esetkeretből levezethető entitásokat is fel lehetne ismerni, pl. ***A konferencia*** *véget ért.* ***A résztvevők*** *elégedetten távoztak.* Ehhez hasonlóan, megfelelő lexikonnal az idiomatikus igevonzatként funkcionáló NP-ket is azonosítani lehetne és kizárni a feloldásból, pl. ***feleségül*** *vesz, ad* ***egy esélyt****, megkéri* ***az árát*** stb. Továbbá hasonló felsorolás szükséges az anaforikus értelemben használt egyértelműsítő kifejezésekről: *az előbbi/utóbbi* stb.

Irodalom

Brennan, Susan E. – Marilyn W. Friedman – Carl J. Pollard 1987. A centering approach to pronouns. In: Candy L. Sidner (szerk.): Proceedings of the 25th meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics. 155–162.

Csendes Dóra – Alexin Zoltán – Csirik János – Kocsor András 2005. A Szeged Korpusz és Treebank verzióinak története. In: Alexin Zoltán – Csendes Dóra (szerk.): A III. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 409–412.

Grosz, Barbara – Aravind Joshi – Scott Weinstein 1995. Centering: A framework for modelling the local coherence of discourse. Computational Linguistics 21: 203–226.

Hobbs, Jerry 1977. Resolving pronoun references. In: Barbara Grosz – Karen S. Jones – Bonnie Webber (szerk.): Readings in natural language processing. Los Altos CA: Morgan Kaufman. 339–352.

Kenesei István 1992. Az alárendelt mondatok szerkezete. In: Kiefer Ferenc (szerk.): Strukturális magyar nyelvtan 1. Mondattan. Budapest: Akadémiai Kiadó. 79–176.

Lappin, Shalom – Herbert Leass 1994. An algorithm for pronominal anaphora resolution. Computational Linguistics 20: 535–562.

Leacock, Claudia – Martin Chodorow 1998. Combining local context and WordNet similarity for word sense identification. In: Christiane Fellbaum (szerk.): WordNet: An electronic lexical database. Cambridge MA: MIT Press. 265–285.

Lejtovicz Katalin – Kardkovács Zsolt 2006. Anaforafeloldás magyar nyelvű szövegekben. In: Alexin Zoltán – Csendes Dóra (szerk.): A IV. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 362–364.

Miháltz, Márton – Csaba Hatvani – Judit Kuti – György Szarvas – János Csirik – Gábor Prószéky – Tamás Váradi 2008. Methods and results of the Hungarian WordNet project. In: Attila Tanács – Dóra Csendes – Veronika Vincze – Christiane Fellbaum – Piek Vossen (szerk.): Proceedings of the Fourth Global WordNet Conference. Szeged: University of Szeged. 311–321.

Mitkov, Ruslan 1999. Anaphora resolution: The state of the art. Ms. University of Wolverhampton.

Ng, Vincent 2005. Machine learning for coreference resolution: From local classification to global ranking. In: Kevin Knight – Hwee Ton Ng – Kemal Oflazer (szerk.): Proceedings of the 43rd Annual Meeting of the Associaction for Computational Linguistics. Ann Arbor, MI: Association for Computational Linguistics. 157–164.

Pléh Csaba 1998. Mondatközi viszonyok feldolgozása: az anafora megértése a magyarban. In: Pléh Csaba: A mondatmegértés a magyar nyelvben. Budapest: Osiris Kiadó. 164–195.

Pléh Csaba – Radics Katalin 1976. „Hiányos mondat", pronominalizáció és a szöveg. Általános Nyelvészeti Tanulmányok 9: 261–277.

Prószéky Gábor – Miháltz Márton 2008. Magyar WordNet: az első magyar lexikális szemantikai adatbázis. Magyar Terminológia 1: 43–57.

Prószéky, Gábor – László Tihanyi – Gábor Ugray 2004. Moose: A robust high-performance parser and generator. In: John Hutchins – Michael Rosner (szerk.): Proceedings of the 9th Workshop of the European Association for Machine Translation. La Valletta: University of Malta. 138–142.

Soon, Wee Meng – Hwee Tou Ng – Daniel Chung Yong Lim 2001. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27: 521–544.

Tejaswini, Deoskar 2004. Techniques for anaphora resolution: A survey. Kézirat. Cornell University. (http://tinyurl.com/c9j6w7c)

Uryupina, Olga 2004. Evaluating name-matching for coreference resolution. In: Maria Teresa Lino – Maria Francisca Xavier – Fátima Ferreira – Rute Costa – Raquel Silva (szerk.): Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon: European Language Resource Association. 1339–1342.

Uryupina, Olga 2006. Coreference resolution with and without linguistic knowledge. In: Proceedings of the 6th International Conference on Language Resources and Evaluation. Genoa: European Language Resource Association. 893–898.

Varasdi Károly 2005. Koreferenciák feloldása. Kézirat. MTA Nyelvtudományi Intézet, Budapest.

Knowledge-based coreference resolution and possessor identification in Hungarian texts

Abstract: The task of NP-coreference resolution involves the identification of expressions referring to the same real-world entities. In this paper, we present an automatic system aimed at treating phenomena like personal pronouns and zero pronouns, synonyms, hypernyms and hyponyms, name variants and name repetitions. Possessor identification is a similar task, where the system has to match detached phrases – often separated by other constituents – corresponding to possessors and their possessions. The rule-based system presented here relies on several knowledge sources: deep parsing information from the MetaMorpho parser, rules implementing Government and Binding Theory, results from psycholinguistic research, semantic information from Hungarian WordNet and character-based heuristic methods.

Keywords: anaphora resolution, coreference resolution, possessor identification, knowledge-based, rule-based

Igék lexikai reprezentációja és a nyelvtechnológia

Héja Enikő[1] – Gábor Kata[2]

[1,2]*Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Budapest*
[2]*LASELDI, Université de Franche-Comté, Besançon*
heja.eniko@nytud.mta.hu; gabor.kata@nytud.mta.hu

Tanulmányunkban azt vizsgáljuk, hogy milyen elvárásoknak kell megfelelnie egy igei lexikonnak. Két feltételt fogalmaztunk meg: a koherenciát és az explicitséget. A lexikon koherenciáját formai-szintaktikai alapú vonzattesztek biztosítják, az explicitséghez pedig az igei tételek megfelelő reprezentációjára van szükség. Állításunk szerint a magyarban nem született még általános érvényű, a koherenciát biztosítani képes vonzatteszt, így javasoljuk, hogy a hagyományos elemzési sorrendet megfordítva először írjuk le a produktív szerkezetek körét, majd tekintsük azokat az összetevőket vonzatnak, amelyekhez nem találtunk produktív szabályokat. Célunk tehát a produktivitás fogalmának pontosabb meghatározása. A vizsgálódás segítségünkre van abban, hogy általánosításokat tegyünk az igék bővítménykeretére vonatkozóan, így növelve a rendelkezésre álló adatbázis koherenciáját és explicitását.

Kulcsszavak: igei lexikon, lexikai reprezentáció, vonzatteszt, kompozicionalitás, igeosztályok

1. Bevezetés

Az idioszinkratikus információk tárhelye a generatív nyelvészeti elméletekben a lexikon, ahol – többek közt – az elemi, nem megjósolható forma–jelentés párokat tároljuk. A lexikon létrehozásán túl az elméletek további feladata, hogy megfelelő szabályok felállításával számot adjanak a nyelvben előforduló produktív jelenségekről.

A számítógépes nyelvészetben és a számítógépes nyelvfeldolgozásban a lexikon szerepe hasonló: a lexikai adatbázisok a lexikai egységekhez tartozó nem megjósolható információt tartalmazzák. Fontos különbség, hogy a számítógépes lexikonok célja az adatok lehető legnagyobb lefedettségének biztosítása. Ezzel szemben az elméleti kutatásokban rendszerint csak a nyelvi adatok egy szűkebb halmazára fókuszálnak.

Milyen elvárásoknak kell megfelelnie a lexikonnak, legyen az akár számítógépes, akár nyelvészeti? Először is, a lexikonnak **koherensnek** kell lennie. Ez egyfelől azt jelenti, hogy elvárjuk, hogy ugyanolyan típusú dolgok ugyanúgy le-

gyenek reprezentálva. Ehhez szükséges egy olyan nyelvészeti elmélet, amely lefedi a lexikonban szerepeltetendő tételek teljes körét. Másfelől arra is szükség van, hogy a használt nyelvészeti elmélet olyan teszteket biztosítson, amelyek az emberi intuíció helyett a szerepeltetendő tételek megfigyelhető viselkedésére helyezik a hangsúlyt. Ez biztosítja, hogy ugyanazt a nyelvi jelenséget különböző kódolók ugyanúgy kódolják. A koherencián túl egy lexikai adatbázisnak **explicitnek** is kell lennie, vagyis a lexikon nem támaszkodhat a felhasználó intuíciójára.

A fentiekkel összhangban egy igei adatbázis építése során az igei szerkezethez tartozó nem megjósolható elemeket kell felsorolni a lexikonban. A koherencia szükséges feltétele, hogy legyenek olyan formai-szintaktikai alapú vonzatteszt ek, amelyek alapján a vonzatok elkülöníthetőek az egyéb grammatikai funkcióval rendelkező konstituensektől. Az explicitség ebben az esetben azt jelenti, hogy az igei tételek reprezentációjánál sem támaszkodhatunk a felhasználók intuíciójára.

Mint látni fogjuk, a magyarra eddig nem született általános érvénnyel alkalmazható, pusztán formai alapon működő vonzatteszt. Mivel véleményünk szerint ez kizárólag felszíni szintaktikai alapon nem is lehetséges, azt javasoljuk, hogy a hagyományos elemzési sorrendet megfordítva első lépésben írjuk le, hogy pontosan mik azok a szerkezetek, amelyek produktívan használhatók a magyarban, és második lépésben azokat az összetevőket tekintsük vonzatnak, amelyekhez nem találtunk produktív szabályokat.

Jelen cikkben azzal foglalkozunk, hogy pontosan mit jelent a produktivitás, illetve hogy hogyan határozhatjuk meg a produktív bővítmények körét. Ez a vizsgálódás segítségünkre van abban, hogy a rendelkezésünkre álló vonzatkeret-adatbázist (Gábor et al. 2008) olyan lexikai-szemantikai információval bővítsük ki, amely segítségével hasznos általánosításokat tehetünk az igék bővítménykeretére vonatkozóan, így növelve az adatbázis koherenciáját és explicitását.

Cikkünk az alábbi szerkezetet követi: a 2. szakaszban ismertetjük, hogy milyen elméletek születtek az igék lexikai reprezentációjának leírására, illetve bemutatjuk a legfontosabb igei adatbázisokat. A 3. szakaszban azt tárgyaljuk, hogy véleményünk szerint miért nem lehetséges pusztán formai-szintaktikai alapú vonzatteszt re támaszkodni a magyarban. A 4. szakaszban az általunk javasolt vonzatdefiníciót fejtjük ki. Az 5. szakaszban két olyan kísérletről számolunk be, amelyeknek a célja az általunk javasolt lexikai-szemantikai reprezentáció kidolgozása, illetve a kapcsolódó nehézségeket is tárgyaljuk. A 6. szakaszban pedig az adatbázis egy alkalmazására teszünk javaslatot.

2. Igei lexikai adatbázisok létrehozása

2.1. Az igék lexikai reprezentációja

A szintaktikai kutatás és a szintaxis–szemantika interfész területe azzal foglalkozik, hogy milyen elvek szerint kombinálódhatnak az egyszerű és az összetett nyelvi egységek szintaktikai és szemantikai szinten. Ezen belül a lexikon szerepe, hogy leírja, milyen egyedi, lexikai egységekre vonatkozó megszorítások érvényesülnek ezeken a területeken. A lexikonnak tulajdonított szerep jelentősége a 80-as évek vége óta folyamatosan nő: ekkor kezdtek megjelenni a hierarchikus lexikon-modellek (pl. Bresnan–Zaenen 1990; Copestake 1993; Koenig–Davis 2003; Pinker 1989; Jackendoff 1990; Levin 1993; Pustejovsky 1995; Levin–Rappaport Hovav 2005). Az igék lexikai reprezentációja kibővült, az igei vonzatkeret szintaktikai jellemzőinek egyszerű felsorolásán túl helyet kapott benne az ige által megkívánt argumentumok szemantikai szerepének jellemzése (többnyire thematikus szerepek formájában), az argumentumokra tett szelekciós megkötések, a szemantikai argumentumok és a szintaktikai vonzatkeret közti leképeződés, az esetleges beágyazott argumentumok kontrolljára vonatkozó információ, valamint az esetleges szintaktikai alternációk. Ezzel szemben az adjunktumokról az idézett elméletek feltételezik, hogy nem a lexikonban, hanem a nyelvtan szintaktikai moduljában, produktív szabályokkal kezelendők.

Megállapítást nyert, hogy nemcsak a nyelvelmélet/nyelvleírás, hanem a számítógépes nyelvészet szempontjából is alapvető fontosságú az elemzéshez rendelkezésre álló lexikai információ mennyisége és minősége. A lexikalizált (az-az lexikai információra is támaszkodó) szintaktikai elemzők kimutathatóan jobb eredményeket érnek el; a kizárólag általános szintaktikai szabályokra támaszkodó elemzőkhöz képest például eredményesebben oldják meg a szerkezeti homonímiák feloldását (Schabes–Waters 1993; Abeillé–Candito 2000). Briscoe és Carroll (1993) becslése szerint a szintaktikai elemzés hibáinak körülbelül fele az igei vonzatkeret hibás felismeréséből vagy fel nem ismeréséből fakad, az igei lexikon tehát fontos szerepet játszik az automatikus szintaktikai elemzésben is.

Az igék lexikai reprezentációjának kérdésköre a szintaxis–szemantika interfészhez tartozik, mivel összefüggés figyelhető meg a szintaktikai viselkedés és az igék szemantikai tulajdonságai között. Harris (1954) „disztribúciós hipotézise" fogalmazta meg azt a megfigyelést, amely szerint a szemantikailag hasonló szavak hasonló környezetben fordulnak elő. Az igei vonzatkeret szempontjából ez azt jelenti, hogy hasonló szemantikai tulajdonságokkal rendelkező igék hasonló szubkategorizációs tulajdonságokat mutatnak. A Semantic Basis Hypothesis (Koenig–Davis 2006) a következőképpen fogalmazza meg a feltételezést: egy ige

jelentéséből levezethető az ige szubkategorizációs tulajdonságainak jelentős része, sőt, akár a teljes szubkategorizációja is, ha az ige nem rendhagyó.

Az argumentumrealizációs elméletek *(mapping/linking theories)* ebből a megfigyelésből kiindulva keresik a választ arra, hogy milyen szemantikai információnak van helye az igék lexikai tételeiben, amelyből nyelvészetileg megalapozott formában megjósolhatjuk az argumentumok szintaktikai megjelenését. A generatív nyelvtanok jelenlegi argumentumrealizációs elméletei (pl. Reinhardt 2002 (minimalizmus); Koenig–Davis 2003 (HPSG); Copestake 1993 (HPSG); Bresnan–Zaenen 1990 (LFG)) erősen támaszkodnak a thematikus szerep (Fillmore 1968) fogalmára. Feltételezik, hogy létezik a szerepeknek egy véges, univerzális halmaza, amellyel az argumentumrealizáció tetszőleges természetes nyelvben leírható. Az idők során kialakult néhány általánosan elfogadott elv a theta-szerepek kiosztásáról (théta-kritérium – Chomsky 1981; Uniformity of Theta Assignment – Baker 1988, thematikus szerepek hierarchiája – Grimshaw 1990; Transparency Principle – Lightfoot 1979, illetve az argumentumrealizációs modellekhez igazítva: Koenig–Davis 2003). Az argumentumrealizációs elméletek a thematikus szerepek hierarchiájából a fenti elveket elfogadva próbálják természetes osztályokba sorolni a predikátumokat, amely osztályok tagjai hasonló módon jelenítik meg a lexikonban tárolt szemantikai argumentumaikat.

Bár a thematikus szerepek használata valóban hasznos általánosításokat tesz lehetővé, az adatokkal szembesítve több problémát vet fel. Miközben a szerepek halmazát univerzálisnak tételezzük, nem lehetünk benne biztosak, hogy minden nyelvben ugyanazok a szemantikai komponensek határozzák meg az argumentumrealizációt. Nem véletlen, hogy az elmúlt negyven évben nem alakult ki konszenzus a thematikus szerepeknek egy jól definiált, végleges készletéről. Többek közt Koenig–Davis (2003) is megjegyzi, hogy a nyelvészek gyakran élnek csak szintaktikailag motivált, ad hoc reprezentációkkal, melyek nem feleltethetők meg semmilyen szemantikailag motivált természetes osztálynak – pusztán azért van szükség rájuk, hogy az adott argumentumrealizációs elmélet, illetve a fent idézett megszorítások követelményei teljesüljenek.

A Levin (1993) és Pustejovsky (1995) által képviselt argumentumrealizációs irányzatok közös eleme, hogy szintén a disztribúciós elvből indulnak ki, de nem előre definiált és univerzálisnak gondolt szemantikai reprezentációval dolgoznak. Éppen ellenkezőleg, a disztribúcióból kiindulva keresik a szintaktikai megjelenítést magyarázó lexikális szemantikai tulajdonságokat az adott nyelvben. Levin (1993) és Levin–Rappaport Hovav (2005) az angol igék vonzatkeret-alternációinak vizsgálatából indul ki. Feltételezik, hogy azok az igék, melyek ugyanazokban az alternációkban vesznek részt, bizonyos szemantikai jelentéskomponensekben osztoznak, amelyek felelősek az alternáció megjelenéséért.

A jelentéskomponensekhez (metapredikátumokhoz) tehát a szintaktikai alternációk vizsgálatán keresztül juthatunk el. Ebből következik, hogy Levin csak az adott nyelvben szintaktikai váltakozást kiváltó metapredikátumokat kívánja felvenni az ige lexikai szemantikai reprezentációjába. Az általunk javasolt megközelítés leginkább Levin munkájához hasonlít, noha a vizsgált jelenségek köre nem esik teljesen egybe: ahogy a későbbiekben látni fogjuk, mi szintén az adatok felől kiindulva kívánjuk levezetni a szemantikai reprezentációt.

2.2. Számítógépes igei lexikai adatbázisok

Az alábbiakban áttekintést adunk a különböző, elektronikus formában elérhető, a számítógépes nyelvészeti használat céljainak megfelelő igei adatbázisokról. A bemutatott lexikonok az igei szintaxis (vonzatkeret, szubkategorizáció és szemantikai szelekció), az igei jelentés és az igék közti szemantikai relációk (pl. szinonímia, hiperonímia), illetve szemantikai–szintaktikai predikátumosztályok leírását tűzik ki célul.

2.2.1. Kézzel épített adatbázisok

A nyelvészek által, emberi munkával létrehozott lexikonok a közvélekedés szerint pontosabbak, mint az automatikus, többnyire annotált korpuszokból kinyert adatbázisok. Ahhoz, hogy ez az állítás teljesüljön, olyan koherens szerkesztési elveket kell megfogalmazni, amelyek lehetővé teszik, hogy az adatbázison párhuzamosan dolgozó kódolók munkájának eredménye egységes alapelveknek feleljen meg.

Maurice Gross igei reprezentációs módszere, a lexique-grammaire (Gross 1975) szintaxis- és adatvezérelt. Amint a neve is mutatja, a lexikai és a grammatikai (itt elsősorban szintaktikai) információ különböző szinten való kezelése ellen foglal állást. A lexique-grammaire adatbázisban az igék osztályokba vannak sorolva a rájuk jellemző szintaktikai alapkonstrukciók alapján. Az alapkonstrukció az egyszerű mondatban az igével **kötelezően** előforduló bővítmények összességét tartalmazza. A predikátumosztályra jellemző további konstrukciók Gross szerint az alapkonstrukcióból hozhatók létre transzformációk útján. Ezek azonban már nem feltétlenül jellemzőek minden igére, így egyedileg van kódolva minden lexikai tételnél, hogy elfogadja-e az adott szerkezetet vagy nem.

A Proton szubkategorizációs lexikon (Eynde–Mertens 2003) készítői olyan disztribúciós teszteket használtak a vonzatkeretek megállapításához, amelyek a névmási behelyettesíthetőségen alapulnak, és végső soron szintén előfeltételezik

a vonzatok kötelezőségét. Feltételezésük szerint a névmási vonzatokat tartalmazó mondatok grammatikalitásának megítélése könnyebb és egyértelműbb.

A Prague Dependency Treebank (Hajič et al. 2001) annotálásával párhuzamosan készült a Vallex cseh szubkategorizációs lexikon (Lopatková 2003). A vonzatkeretbe tartozó elemek meghatározásánál szempont volt a kötelezőség, valamint az ige és bővítménye közti szemantikai viszony. A lexikon egy harmadik, átmeneti kategóriát, az ún. kvázi-komplementumokét is megkülönbözteti. Ezt azok a bővítmények alkotják, amelyek egyes igecsoportok mellett produktívan feltűnhetnek, általában opcionálisak, az igecsoport tagjai mellett ugyanazt a szemantikai szerepet kódolják ugyanabban a szintaktikai formában, ám ezt a szintaktikai formát (Lopatková 2003 szerint) az egyedi ige lexikai tétele írja elő.

A FrameNet lexikon (Baker et al. 1998) Charles Fillmore szemantikai reprezentációs elméletén (Fillmore 1982) alapul. A predikátumok jelentésük és argumentumaik szemantikai szerepe szerinti csoportokba vannak sorolva. Az egy csoporthoz tartozó igék jelentésének közös részét a bővítményeknek kiosztott, a csoportra specifikus szemantikai szerepek összessége (frame) definiálja. A „kommunikációs igék" csoportjához tartozó frame például tartalmazza a „közlő", „üzenet", „címzett" szerepeket. Fontos tehát, hogy a szemantikai szerepek csoportról csoportra változnak, így sokkal nagyobb a számuk és kevésbé általánosak, mint például a thematikus szerepek.

Jackendoff (1983; 1990) predikátum-dekompozíción alapuló szemantikai reprezentációt javasol. A predikátumok szemantikai dekompozíciója lexikai konceptuális struktúrák (Lexical Conceptual Structures, LCS) által történik. A konceptuális struktúrák lehetnek konceptuális primitívumok (elemi jelentésegységek, pl. *okoz, megy*), szemantikai mezők, amelyeknek a segítségével szelekciós megszorításokat definiálhatunk (pl. *idő, hely*), és konceptuális konstituensek, amelyek a predikátumok felső szintű ontológiai besorolását adják meg (pl. *esemény, állapot*). A thematikus szerepeket Jackendoff úgy fogalmazza át, hogy az adott szerepet azzal jelemzi, hogy milyen konceptuális konstituens szemantikai argumentumaként reprezentálható. Ezen konceptuális struktúrák segítségével olyan lexikai reprezentáció építhető, amelyből megjósolhatjuk az igei argumentumok szintaktikai realizációját.

Jackendoff nyomdokain halad Levin (1993) és Kipper et al. (2000), a VerbNet kidolgozója is. A VerbNet adatbázis szemantikai és szintaktikai tulajdonságokban osztozó igecsoportokat definiál Levin elvei szerint, az ő osztályozásából kiindulva. Kifejezetten számítógépes nyelvészeti felhasználásra készült, így alkotója nagy hangsúlyt fektetett arra, hogy a benne tárolt információ explicit legyen. A gépi tanulás céljainak megfelelően igyekeztek továbbá az igeosztályokat kompakt módon reprezentálni, a lehetséges általánosításokat minél magasabb szinten

megtenni. Az igei argumentumok szemantikai jellemzése a VerbNetben thematikus szerepek által történik.

A WordNet (Miller 1995; Fellbaum 1998) jelenleg a legszélesebb körben használt lexikai adatbázis. Az angol változat szerkezetét megőrizve számos európai nyelvre készült WordNet, és minden ismert hibája ellenére széles körben használják különböző nyelvtechnológiai kutatásokban és alkalmazásokban. Elsősorban szemantikai információt kódol. Az adatbázis alapegysége a *synset* (*synonym set*, szinonimahalmaz). A szinonímia WordNetben használt definíciója a kölcsönös implikáció. A synset-ek közti relációkat hierarchikus formában, elsősorban a hipo- és hiperonímia (alá-fölérendeltségi) viszonyok szerint kódolja. A WordNet szemantikai adatbázis, így nagyon kevés explicit szintaktikai információt tartalmaz. Ugyanakkor a magyar WordNet synset-jei egyedi azonosítók segítségével össze vannak kapcsolva az igei vonzatkeret-adatbázissal (Kuti et al. 2007).

2.2.2. Automatikusan épített igei adatbázisok

Az emberi munkával előállított adatbázisokhoz képest kevésbé munkaigényes, de a számítógépes nyelvészet mai állása szerint kevésbé pontos eredményt ad a lexikai információ gépi kinyerése *(lexical acquisition),* elsősorban korpuszokból. Ilyen jellegű feladatok megoldásához először is meg kell fogalmaznunk feltételezésünket arról, hogy a szóhoz tartozó keresett lexikai információt milyen, a korpuszbeli előfordulások alapján számszerűsíthető tulajdonságok reprezentálják. A vonzatkeret-kinyerés esetében (Brent 1991; 1993; Briscoe–Carroll 1997) arra a feltételezésre építhetünk, hogy a vonzatot specifikussága különbözteti meg a szabad határozótól, azaz gyakrabban fog előfordulni az őt szubkategorizáló igék mellett, mint általában egy tetszőleges ige mellett. A szabad határozó ezzel szemben minden ige mellett éppen ugyanakkora eséllyel fordul elő. Ezért a legtöbb módszer az ige és a vonzatjelölt együttes előfordulási gyakoriságából indul ki, és különböző statisztikai tesztekkel (pl. binomiális teszt, t-teszt) vizsgálja, hogy megállapítható-e, hogy az adott predikátum és a vonzatjelölt nem függetlenek egymástól (statisztikai értelemben). Nem teljesen igaz azonban, hogy az automatikusan kinyert vonzatkeret-információ csak arra a feltevésre épít, hogy minden, az adott predikátum mellett gyakran előforduló nyelvi elem vonzat. A módszerek mindig tartalmaznak több-kevesebb prekoncepciót arra vonatkozóan, hogy nézhet ki egy vonzatkeret. Ez a prekoncepció vagy magából a korpuszból jön (az annotáció meghatározza, hogy milyen vonzatkeret-jelöltek nyerhetők ki), vagy az algoritmusból: pl. milyen szinten általánosítja a korpuszban talált előfordulásokat, vagyis milyen reprezentációt alkot; másrészt utólagos szűrőket is gyakran

tartalmaznak, ami szintén építhet arra, hogy mennyire valószínű, hogy a jelölt vonzatkeret lehet az adott nyelvben.

A tisztán szintaktikai vonzatkeret-információ kinyerésén túl egyre nagyobb teret kap a szemantikai tulajdonságok automatikus felismerése is. Ide sorolhatjuk az igék szemantikai szelekciós tulajdonságainak kinyerését (Resnik 1993; Riloff–Schmelzenbach 1998), a szemantikai igecsoportokra jellemző alternációk detektálását (Lapata 1999; McCarthy 2001) valamint a korpuszban megfigylhető disztribúciós mintázataik alapján végzett automatikus szemantikai klasszifikációs/klaszterezési kísérleteket is (pl. Merlo–Stevenson 2001; Joanis–Stevenson 2003; Schulte im Walde–Brew 2002; Korhonen et al. 2003). A szemantikai igecsoportok gépi tanulását célzó kísérletek célja jellemzően a Levin-típusú igeosztályok kinyerése, az osztályzás automatikus kibővítése (Merlo–Stevenson 2001; Korhonen–Briscoe 2004), illetve hasonló osztályzás létrehozása más nyelvekre (Schulte im Walde–Brew 2002; Gábor–Héja 2007; Sass 2007).

Míg az automatikusan kinyert szubkategorizációs adatbázisok minősége napjainkban már vetekszik az emberi munkával létrehozott lexikonokéval (a nagy lefedettség miatt gyakran hasznosabbak is, mint az utóbbiak), addig az igék szemantikai klasszifikációja még nem jutott el a közvetlen felhasználhatóság szintjére. Ugyanakkor az eredmények megerősíteni látszanak a Semantic Basis Hypothesis-t, így az igecsoportok automatikus kinyerése továbbra is élénken kutatott terület maradt.

2.2.3. Konklúzió

Amint az áttekintésünkből kiderül, a nyelvészeti elméletek és a számítógépes lexikonok egyaránt nagy mértékben támaszkodnak arra a feltételezésre, hogy a vonzatok többnyire kötelezőek és specifikusak az adott régensre (kivéve a FrameNet lexikont, amely nem különbözteti meg a vonzatokat és a szabad határozókat). A számítógépes lexikonokról emellett elmondható, hogy gyakrabban alapoznak többé-kevésbé önkényes szerkesztési elvekre[1] – ez azonban a lexikon koherenciájának rovására mehet, így véleményünk szerint mindenképp kerülendő. Amint a következő részben a magyar vonzattesztek áttekintése során látni fogjuk, ezek nem használhatók következetesen, illetve nem biztosítanak kellő lefedettséget a

[1] L. például a szemantikai szelekciós jegyek leírását Gross (1975)-nál, aki felsorolással definiálja az általa szintaktikailag relevánsnak vélt szemantikai tulajdonságokat, illetve az argumentumszerkezet részének tekintett bővítmények thematikus szerepeit Lopatková (2003)-nál, szintén felsorolással meghatározva. A Vallex lexikon esetében a köztes kategória felvétele is mutatja, hogy a bővítmények osztályzásakor használt tulajdonságok nem korrelálnak a bővítmények státuszára vonatkozó egyéb feltételezésekkel.

magyar nyelvi adatokra. Olyan megoldást keresünk tehát, amely egyszerre nyújt fogódzót a vonzatok azonosításához, és teszi lehetővé, hogy – az argumentum-realizációs elméletek célkitűzéseihez hasonlóan – olyan lexikai reprezentációt alkossunk, amelyben természetes osztályokra általánosíthatóvá válnak az igei szintaxis megjósolható aspektusai.

3. Vonzattesztek az X′-elméleteken belül

Az általunk ismert vonzattesztek vagy szemantikai, vagy formai-szintaktikai alapon érvelnek egy bővítmény vonzat státusza mellett. A szemantikai vonzattesztek egy típusa abból az előfeltevésből indul ki, hogy az ige vonzatai csak az igei predikátum argumentumainak szintaktikai realizációi lehetnek (l. pl. Butt 2006-ot az ige mellett opcionálisan megjelenő, nem argumentum szerepű mennyiségjelölők szintaktikai funkciójáról).[2] Ezek a vonzattesztek számunkra nem alkalmazhatóak, hiszen „valaminek-a-szemantikai-argumentumának-lenni" nem tesztelhető tulajdonság, így a nyelvész vagy kódoló a saját intuíciójára van utalva, amikor meghatározza egy igei predikátum szemantikai argumentumainak körét. Az alábbi példamondatok azt szemléltetik, hogy sok esetben nem egyértelmű a döntés arra nézve, hogy valami a predikátum szemantikai argumentuma-e.

(1) Az emberek felriadtak a sziréna hangjára.

(2) Az emberek felnevettek a váratlan poénra.

(3) Az emberek felpillantottak az újságjaikból a furcsa zajra.

(4) Az emberek megrohamozták a bankokat a hírre.

Intuíciónk szerint a szublatívuszi bővítmény a FELRIAD predikátum mellett az (1) mondatban áll legközelebb a tipikus szemantikai argumentum kategóriájához, és a (4) felé haladva mondatról mondatra egyre kevésbé jellemezhető szemantikai argumentumként, ám nincs határozott intuíciónk arról, hogy pontosan hol húzhatnánk meg a határt. Ugyanakkor a szublatívuszi bővítmények első ránézésre hasonló szemantikai szerepe nem támasztja alá a szigorúan szemantikai szempontok szerint történő elkülönítést.

A szemantikai vonzattesztek egy másik csoportja megfogalmazható formai-szintaktikai alapon (pl. Levin–Rappaport Hovav 2005). Mint máshol már említ-

[2] Butt érve azért érdekes számunkra, mert éppen ellentétes eredményre jut, mint Komlósy (1992) és Levin–Rappaport Hovav (2005).

tettük, a formai alapú vonzattesztek egy része nem alkalmazható a magyarra (pl. passziválhatóság, felszíni sorrend; Héja–Gábor 2008), más részük pedig végső soron a vonzatok és a szabad határozók eltérő disztribúciójára vezethető vissza.

Ebben a részben azt kívánjuk megmutatni, hogy a magyar nyelv sajátosságai miatt a formai-szintaktikai kritériumokon alapuló vonzattesztek sem alkalmazhatóak a magyarra egy az X′-elméletet tételező kereten belül (pl. GB, LFG, GPSG).

Ehhez első lépésben felidézünk két bloomfieldi fogalmat:

> *Endocentrikus konstrukció*: „A szubordinatív endocentrikus konstrukciókban az eredményül kapott frázis ugyanahhoz a forma-osztályhoz tartozik, mint a konstrukcióban található konstituensek egyike, ez a fej." (Bloomfield 1933, 195)

Vagyis, ha egy szerkezetben az egyik konstituens rendszeresen behelyettesíthető az egész szerkezet helyére, úgy, hogy az eredményül kapott szerkezet jól formált marad, akkor a szerkezet endocentrikus, és a szóban forgó konstituens a szerkezet feje. Pl. a *tarka macskák* frázis feje a *macskák* konstituens, amely minden környezetben behelyettesíthető az eredeti frázis helyére.

> *Exocentrikus konstrukció*: „Annak ellenére, hogy egy exocentrikus frázis funkciója különbözik a frázis bármelyik részének funkciójától, általában van egy konstituens a frázison belül, amely jellemző a frázis egészére és amely meghatározza az eredményül kapott frázist, így például az angolban, a finit igék […] általában exocentrikus konstrukciókban jelennek meg, […] és elégségesek a konstrukciók meghatározására." (*op.cit.* : 194–195)

Vagyis ha van olyan környezet, amelyben a konstrukció egyik elemét sem helyettesíthetjük be a szerkezet egésze helyére, exocentrikus konstrukcióról beszélünk. Tipikus exocentrikus szerkezetek a kötelező vonzattal rendelkező igék.

Az alábbiakban az esetragos főnévi frázisokra az XP kategóriacímkével hivatkozunk. Egyfelől azért, mert nem kívánunk a vizsgált frázisok között kategoriális szempontból különbséget tenni, másfelől nem szeretnénk állást foglalni abban a kérdésben, hogy a releváns igei bővítmények mely kategória(k)ba tartoznak. Ezért a DP, NP és PP kategóriacímkék helyett egységesen az XP kategóriacímkét fogjuk használni. Az XP jelölés tehát esetragos NP-re vagy DP-re referál, függetlenül attól, hogy ezekhez milyen esetrag tartozik.

Az X′-elméletben a vonzat–adjunktum különbséget újraíró szabályok formájában így foglalhatjuk össze:

(5) a. Vonzat: $V' \rightarrow V + XP$
b. Vonzat: $VP \rightarrow V' + XP$
c. Adjunktum: $V' \rightarrow V' + XP$
d. Adjunktum: $VP \rightarrow VP + XP$

Mint látjuk, a legfontosabb különbség az adjunktumokat tartalmazó újraíró szabályok és a vonzatokat tartalmazó újraíró szabályok között az, hogy az előbbiek rekurzívak, azaz az adjunktumszabály mindkét változatában a bal és jobb oldalon ugyanaz a kategóriacímke áll. Ebből következik, hogy minden szabad határozóval az ige vagy igei szerkezet endocentrikus konstrukciót alkot, amelyben az ige a fej.

Ezzel szemben a vonzatszabályok bal és jobb oldalán szereplő eltérő kategóriacímkék azt reprezentálják, hogy a vonzat megváltoztathatja az eredeti ige módosítási lehetőségeit, így az ige + vonzat szerkezet lehet exocentrikus. Ebből következik, hogy az igével exocentrikus konstrukciót alkotó bővítmények vonzatok.[3]

Az alábbiakban azt mutatjuk meg, hogy az általunk ismert magyarra alkalmazható formai vagy formai alapú szemantikai vonzattesztek mind ezen az előfeltevésen alapulnak. Az alábbiakban Radford (1988) pronominalizációs és ellipszis-tesztjét tárgyaljuk.

Pronominalizációs teszt: A *do so* szerkezet V′ kategóriájú összetevőt helyettesíthet. A *do so*-val helyettesített V′ opcionálisan magában foglalhatja az adjuntumokat (6), de az adjunktumok ki is tehetők a *do so* szerkezet mellé (7), míg a komplementum kötelezően benne foglaltatik a V′-ban (8), nem hagyható el (9).

(6) John will [buy the book on Tuesday] and Paul will do so as well.

(7) John will [buy the book] on Tuesday and Paul will do so on Thursday.

(8) John will [put the book on the table] and Paul will do so as well.

(9) *John will [put the book] on the table and Paul will do so on the chair.

A pronominalizációs vonzatteszt azon alapszik, hogy a *do so* rendszeresen helyettesítheti az ige + vonzat$_1$ + vonzat$_2$ szerkezetet (8), de nem helyettesítheti rendszeresen az ige + vonzat$_1$ szerkezetet (9), hiszen az ige + vonzat$_1$ szerkezet nem helyettesíthető a *do so*-val, ha a vonzat$_2$ megjelenik a mondatban. Vagyis az ige + vonzat$_1$ + vonzat$_2$ és a ige + vonzat$_1$ szerkezetek disztribúciója különbözik, tehát vonzat$_2$ exocentrikus konstrukciót alkot az ige + vonzat$_1$ szerkezettel. A vonzatteszt azt is mutatja, hogy a *do so* nemcsak az ige + vonzat$_1$ szerkezetet helyettesítheti rendszeresen, hanem az ige + vonzat$_1$ + adjunktum egységet is, amiből következik, hogy az utóbbi az előbbinek az endocentrikus bővítése.

[3] Nem állítjuk, hogy minden vonzat exocentrikus konstrukcióban fordul elő: feltesszük, hogy egyes fakultatív vonzatok alkothatnak az igével endocentrikus konstrukciót.

Ellipszis-teszt: Az ellipszis-teszt két ponton tér el a pronominalizációs teszttől. Egyfelől a kérdéses összetevők üres elemmel való felcserélhetőségét vizsgáljuk a *do so*-val való felcserélhetőség helyett. Másfelől a példamondatok közül kimaradt a (8)-nak megfelelő. Az ige + vonzat$_1$ + adjunktum és az ige + vonzat$_1$ szerkezeteket egyaránt elhagyhatjuk a mondatból (10)–(11). Nem hagyhatjuk el viszont vonzat$_1$-et, ha vonzat$_2$ megjelenik (12). Mivel az ige + vonzat$_1$ + vonzat$_2$ szerkezet elhagyása, bár nem szerepel az eredeti mondatok között, jól formált mondatot eredményez (13), az ellipszis-teszt éppenúgy az exocentrikusságon alapszik, mint a pronominalizációs teszt.

(10) A: Who might be going to the cinema on Tuesday?
B: John might be _____.

(11) A: Who might be going to the cinema when?
B: John might be _____ on Tuesday.

(12) A: Who will put the book where?
B: *John will _____ on the table.

(13) A: Who will put the book on the table?
B: John will _____.

A magyarra Komlósy (1992) határozott meg vonzatpróbákat. Mint látni fogjuk, Komlósy vonzatpróbái szintén azt előfeltételezik, hogy az exocentrikus konstrukciókban megjelenő bővítmények vonzatok.

I. Vonzatpróba: Kötelezőség

> „Ha a mondatszerkezet valamely szintjén bővítménynek minősülő egység a mondatból nem hagyható el (ebbe beleértjük azt is, hogy elhagyása elliptikus vagy jelentésében idegen eredményt ad), akkor az vonzat." (Komlósy 1992, 316)

Egy bővítmény akkor kötelező, ha semmilyen kontextusban sem helyettesíthető be az ige az ige + bővítmény szerkezet helyére. Tehát ez a vonzatpróba feltételezi, hogy az ige + bővítmény konstrukció exocentrikus.

II. vonzatpróba: Típusváltoztató fakultatív vonzatok

> „Ha a kérdéses bővítmény (pl. *öt kilót*) kitétele lehetővé teszi egy újabb szabadon elhagyható bővítmény *(egy hét alatt)* megjelenését (és ez azzal jár, hogy az így kapott mondatból az előbbi bővítmény már nem hagyható el), akkor a kérdéses bővítmény *(öt kilót)* a régens fakultatív vonzatának tekintendő." (*ibid.*, 318)

Így például az *öt kilót* konstituens a *hízott* igének a fakultatív bővítménye, hiszen a *Mari hízott* és a *Mari hízott öt kilót* mondatok egyaránt grammatikusak. Ugyanakkor azt is látjuk, hogy a második fakultatív bővítmény csak az első fakultatív bővítmény mellett jelenhet meg: míg a *Mari öt kilót hízott egy hét alatt* mondat grammatikus, a **Mari hízott egy hét alatt* mondat agrammatikus. A II. vonzatpróba alapján ebből következik, hogy az *öt kilót* fakultatív bővítmény a *hízott* ige vonzata. Ez a vonzatpróba is az exocentrikusságon alapszik. Ha az első fakultatív bővítmény *(öt kilót)* bármikor szabadon elhagyható lenne, akkor a szerkezet endocentrikus lenne. A próba éppen azt mondja ki, hogy ha találunk olyan környezetet, amelyben az ige és az ige + bővítmény egység nem jelenhet meg egyszerre, akkor a kérdéses bővítmény vonzat.

III. vonzatpróba: Típust nem változtató fakultatív vonzatok

> „Amennyiben egy X szó és egy bővítmény viszonyát vizsgálva találunk olyan Y szót, amely
> (a) rendszeresen helyettesítheti az X + bővítmény egységet
> (b) a bővítményt nem tartalmazó mondatokban rendszeresen helyettesítheti X-et, de
> (c) a bővítményt is tartalmazó mondatokban nem válthatja fel X-et,
>
> úgy a kérdéses bővítményt X szó (típust nem változtató) vonzatának kell minősítenünk."
>
> (*ibid.*, 320)

Illusztrációként vegyük Komlósy példamondatait: *Mindannyian órákon át csodálkoztunk az eredményen* illetve *Mindannyian órákon át csodálkoztunk*. A III. vonzatpróba alapján az *eredményen*-t a *csodálkoztunk* vonzatának kell tartanunk, hiszen más olyan régensek, amelyek hasonló típusú tényállásokat jelölnek, nem engedélyezik – azonos jelentéssel – a szuperesszívuszi összetevő megjelenését a mondatban: *Mindannyian viháncoltunk/unatkoztunk *az eredményen.* Komlósy szerint ha az *eredményen* adjunktum lenne, akkor egy ilyen típusú helyettesítés nem ronthatná el a mondatot

A III. vonzatpróba hasonlít az angol *do so* pronominalizációs tesztre és az ellipszis-tesztre. Az (a) pontban kikötjük, hogy Y szónak rendszeresen helyettesítenie kell az X + bővítmény egységet. Ha az X + bővítmény szerkezet endocentrikus lenne, akkor X rendszeresen megjelenhetne az X + bővítmény szerkezet helyén, vagyis X rendszeresen megjelenhetne Y helyén. Ugyanakkor (c)-ből következik, hogy Y éppen az X + bővítmény szerkezetekben nem válthatja le X-et. Vagyis az X + bővítmény kontrukció exocentrikus.

Eddig azt mutattuk meg, hogy az általunk ismert pusztán formai vonzattesztek abból az előfeltevésből indulnak ki, hogy ha egy bővítmény exocentrikus szerkezetet alkot a fejjel, akkor az vonzat. Az alábbiakban azt fogjuk belátni, hogy a szemantikai vonzattesztek egy része mögött is hasonló megfontolás van.

Levin és Rappaport Hovav (2005) az eseménytípus-váltást előidéző bővítmények szintaktikai funkcióját vizsgálták. Akkor mondjuk, hogy egy bővítmény eseménytípus-váltó, ha megváltoztatja az ige + bővítmény(ek) által leírt esemény típusát. Az eseménytípust úgy tudjuk megadni, hogy felsoroljuk, milyen fajta szabad határozókkal módosítható a szóban forgó szerkezet.

Így például a rezultatív konstrukcióban megjelenő fakultatív bővítmények eseménytípus-váltóak:

(14) a. The blacksmith hammered the metal.
'A kovács kalapálta a vasat.'

b. The blacksmith hammered the metal flat.
'A kovács laposra kalapálta a vasat.'

A *hammered the metal* más típusú eseményt ír le, mint a *hammered the metal flat*, hiszen a két kifejezés eltérő időhatározókkal módosítható:

(15) a. *The blacksmith hammered the metal in two hours.
'A kovács két óra alatt kalapálta a vasat.'

b. The blacksmith hammered the metal flat in two hours.
'A kovács két óra alatt laposra kalapálta a vasat.'

Levin és Rappaport Hovav (2005) szemantikai alapon érvelnek az eseménytípus-váltást előidéző főnévi csoport vonzat státusza mellett. Szerintük a rezultatív konstrukcióban szereplő *flat* nem lehet adjunktum, hiszen „extra jelentéskomponenst" ad az ige jelentéséhez. Ez azt jelenti, hogy az eseménytípus-váltással az ige jelentése is megváltozik. A szabad határozók azonban nem képesek az ige jelentését megváltoztatni, hiszen csak az ige jelentésében foglalt esemény koordinátáit specifikálják. Ezzel szemben a szemantikai argumentumok az igei jelentés részei, tehát az általuk hordozott jelentéskomponens mindenképpen az ige jelentéséhez adódik hozzá. Ez a plusz jelentéskomponens okozza az idézett példában az eseménytípus-váltást.

Mint azt már korábban részletesen leírtuk (Héja–Gábor 2008), Komlósy II. vonzatpróbája Levin és Rappaport Hovav érvének szintaktikai megfelelője. Mivel az eseménytípust a kitehető adjunktumokkal definiáljuk, minden eseménytípus szabad határozókon értelmezett disztribúciója egyedi. Így az eseménytípus-változást előidéző bővítmény megváltoztatja a disztribúciót is. Tehát az eseménytípus-váltó bővítmények definíció szerint exocentrikus konstrukciót alkotnak az igével és az igéhez tartozó esetleges további bővítményekkel.

Gábor–Héja (2006) és Héja–Gábor (2008) alapján azt állítjuk, hogy a magyarban vannak olyan ige-fejű exocentrikus szerkezetek, amelyekben a bővítményt nem tekinthetjük az ige vonzatának.

Egyfelől, elvárásunk szerint a határozói igemódosítok endocentrikus szerkezetet alkotnak az igével:

(16) a. János olvasott.
b. János lassan olvasott.

Ugyanakkor az ilyen típusú szerkezetek exocentrikussá bővíthetők: vagyis találunk olyan környezetet, amelyben az ige csak a módosítóval együtt jelenhet meg.

(17) a. János lassan olvasott a fáradtságtól.
b. *János olvasott a fáradtságtól.

Mivel az X′-elméletből következik, hogy az ige-fejű exocentrikus bővítések vonzatok, a (17b) példából következne, hogy a *lassan* adverbium vonzata az igének. Ezt a következményt azonban nem tartjuk elfogadhatónak, így azt állítjuk, hogy a magyarban nem hivatkozhatunk az exocentrikusságra a vonzatok definíciója során.

Felmerülhet, hogy a *fáradtságtól* esetragos NP nem az igének, hanem a *lassan* adverbiumnak a bővítményeként szerepel a fenti mondatban. Ezt a megállapítást azonban nem támasztják alá az általunk vizsgált összetevőségi tesztek. A *fáradtságtól lassan*, illetve *lassan a fáradtságtól* potenciális összetevők nem tehetők fókuszba (oly módon, hogy csakis az összetevő első szava viseljen irtóhangsúlyt): **János ′LASSAN A FÁRADTSÁGTÓL olvasott*, illetve **János ′A FÁRADTSÁGTÓL LASSAN olvasta el a cikket.*

Természetesen a második mondat helyessé tehető, ha a *lassan* adverbium viseli az irtóhangsúlyt, vagyis egyedüli összetevőként van fókuszálva. A névmási helyettesíthetőségen alapuló magyar tesztek sem segítenek a kérdés eldöntésében. Véleményünk szerint nem szól erős érv a mellett, hogy a láthatóan külön mozgatható részekből álló szerkezetet egy összetevőnek tekintsük.

Fontos megjegyezni, hogy az idézett példa nem véletlenszerű, elszigetelt ellenpélda, hanem egy általános jelenség prototipikus esete. A példa megkonstruálása során abból a megfigyelésből indultunk ki, hogy az ablatívuszi bővítmények jelölhetik nem ágenses igék által leírt események természeti erő szerepű szereplőjét (l. pl. Komlósy 1992). Ezzel szemben, ágenses igék mellett az ablatívuszi bővítmény sosem jelenhet meg természeti erőként. Példánk azt szemlélteti, hogy bizonyos feltételek mellett, pédául ha az igét olyan adverbium módosítja, amely

az ige által leírt cselekvés egy nem szándékos aspektusát emeli ki, a természeti erő szerepű ablatívuszi bővítmény mégis megjelenhet a mondatban, azonban az adverbium ilyenkor kötelező.

Véleményünk szerint a *lassan* adverbium szemantikailag egy predikátumnak felel meg, amely így az igéből és esetleges bővítményeiből álló szerkezettel összekapcsolódva fölöttes predikátumot képez, lehetővé téve a természeti erő szemantikai szerepű konstituens megjelenését az igét és adjunktumot tartalmazó szerkezet mellett. A fenti példába más ágenses igét helyettesítve szintén grammatikus szerkezetet kapunk. Mivel az X′-elméletből következik, hogy az ige-fejű exocentrikus bővítések vonzatok, a (17b) példából következne, hogy a *lassan* adverbium vonzata az *olvas* igének, illetve a többi, hasonló bővítést megengedő ágenses igének is.

További vizsgálatot igényel, hogy a magyarra miért nem áll az az előfeltevés, hogy minden ige-fejű exocentrikus konstrukcióban a bővítmény vonzatszerepű. Feltételezésünk szerint ez a magyar diskurzus-funkcionális jellegével függ össze, amennyiben a fókusz vagy egyéb igemódosító pozícióban megjelenő összetevők (pl. adverbiumok, igekötők), mint ige fölötti predikátumok képesek megváltoztatni az ige által leírt esemény típusát és ezáltal az eredeti igei szerkezet módosítási lehetőségeit.

Nyitva hagyjuk azt a kérdést, hogy elképzelhető-e olyan egyéb formai-szintaktikai vonzatteszt, amely semmilyen formában nem támaszkodik az exocentrikusságra. A következő szakaszban azt vizsgáljuk meg, hogy milyen egyéb lehetőség van egy koherens és explicit lexikai adatbázis létrehozására.

4. Kompozicionalitás

Az előzőekben bemutatott érvekből levonhatjuk azt a következtetést, hogy a magyarban sem a VP exocentrikus jellege, sem felszíni szintaktikai tesztek, sem a kötelezőség alapján nem különíthetőek el egymástól a vonzatok és az adjunktumok. Ezért az alábbiakban a szintaxis és a szemantika egy szinten kezelése mellett fogunk érvelni. Megközelítésünkben a nyelvtanilag releváns szemantikai fogalmakat szintaktikai jelenségekhez horgonyozzuk le. Ez a gondolat a magyar vonatkozásában Alberti vonzatszerkezetekre vonatkozó Tau-modelljében is megtalálható (l. pl. Alberti–Kilián 2010).

Azonban a Tau-modellel és a korábban ismertetett megközelítésekkel szemben mi a szabad határozók leírását a vonzatokhoz képest elsődlegesnek tekintjük, amennyiben a vonzatok meghatározását a szabad határozókéból vezetjük le: azokat a bővítményeket tekintjük vonzatnak, amelyeknek a megjelenése egyáltalán

nem jósolható meg produktív szabályok alkalmazásával, illetve ezzel párhuzamosan, szemantikai szerepük nem vezethető le a szintaktikai megjelenítésükből.

Komlósy (1992) is említi, amikor az eseménytípust definiálja, hogy különböző típusba sorolható eseményeket kifejező igék különböző adjunktumokat engednek meg. Ez azt jelenti, hogy az igei jelentés bizonyos aspektusai befolyásolják az ige mellé kitehető adjunktumok körét, vagyis nem tartható az a nézet, amely szerint az adjunktumok mindegyike teljesen produktív lenne. A szigorú bináris különbségtétel a teljesen specifikus vonzatok és a teljesen produktív adjunktumok között tehát nem állja meg a helyét. E mellett érvel Kálmán (2006b) és Rákosi (2006) is.

Érdekes, hogy ennek ellenére az adjunktumok igék melletti eloszlása nem kapott nagy figyelmet, leszámítva az igei aspektus és az időhatározók közti összefüggés vizsgálatát (pl. Vendler 1957; Kiefer 1992a). Véleményünk szerint az adjunktumok disztribúcióját érdemes hasonló lexikális szemantikai keretben vizsgálni, mint az igei argumentumrealizációt szemantikai tulajdonságokra visszavezető elméletekben. Ez egybecseng azzal, hogy az adjunktumokat kompozicionálisan kapcsolódó szerkezetnek tekintjük, amelyeknek a megjelenése lexikális szemantikai jegyekkel definiált ige**osztályok** mellett megjósolható (Gábor–Héja 2006). A kompozicionalitás feltételezi, hogy az igeosztály predikátumai osztoznak legalább egy szintaktikailag is releváns jelentéskomponensben, amely szemantikailag kompatibilis az adjunktum szemantikai szerepével. Feltételezi továbbá, hogy az azonos morfoszintaktikai formában (névutó, esetrag) megjelenő adjunktumok szemantikai szerepe az igeosztályba tartozó valamennyi predikátum mellett állandó (tehát nem specifikus az ige lexikai tételére).

Az igék osztályokba sorolása egyfelől elméleti szempontból is érdekes általánosítások megfogalmazását teszi lehetővé, másfelől (véleményünk szerint) megkerülhetetlen részfeladat a produktív és a lexikális információ pontos elhatárolásában, hiszen biztosítja a lexikai adatbázis koherenciáját. A bővítmények széles körének visszavezetése az lexikai-szemantikai tulajdonságaira ugyanakkor nem mond ellent az explicitség követelményének, hiszen a részlegesen produktív adjunktumoknak a szintaktikai és szemantikai elemzéshez szükséges jellemzői a következő szakaszokban meghatározott lexikai reprezentációból levezethetők, illetve a lexikai ábrázolás ilyen szempontból explicitté tehető.

Az igei jelentést egy csak az igére jellemző, specifikus **magjelentés** és az igecsoport tagjaira jellemző általánosabb jelentéskomponens(ek) (**metapredikátumok**) összességeként fogjuk fel. Szemantikai szinten az adjunktumokat a metapredikátumok engedélyezik, míg a vonzatokat az ige saját magjelentése írja elő (kötelező vonzat esetében), illetve engedélyezi (opcionális vonzat estében).

A következő szakaszban bemutatott kísérletek célja a magyarban szintaktikailag releváns predikátumcsoportok, a mellettük megjelenő adjunktumtípusok és az utóbbiak szemantikai szerepének vizsgálata. Elsősorban Levin munkáiból indulunk ki, aki arra az eredményre jutott, hogy egyes szemantikai jelentéskomponensekben osztozó igék ugyanolyan vonzatkeret-alternációkat engednek meg. Vonzatkeret-alternáció helyett mi az adjunktumok szintaktikai disztribúcióját helyezzük a középpontba, ám itt érdemes megjegyezni, hogy a fenti definíció alapján az általunk adjunktumnak tekintett bővítmények köre szélesebb, például egy kompozicionális bővítményt pusztán a szintaktikai kötelezősége miatt nem sorolunk a vonzatok közé. Levin mintájára a szemantikai tulajdonságokat (az igecsoportra jellemző metapredikátumokat, illetve az adjunktumok szemantikai szerepének leírását) minden esetben szintaktikailag tesztelhető, disztribúciós jelenségekhez kötjük. Mindkét kísérlet szintaktikai jelenségek megfigyeléséből indul ki, szemantikailag is releváns általánosítások megtételére törekszik, és végső célja az igei adatbázis gazdagítása lexikai–szemantikai igecsoportok azonosítása révén.

5. Szintaktikailag releváns szemantikai igeosztályok azonosítása. Kísérletek

5.1. Adjunktumok szemantikai–szintaktikai funkcióinak leírása és automatikus annotálása

Az adjunktumok szemantikai és szintaktikai funkcióinak leírását egy korpuszalapú kísérlettel kezdtük (Gábor–Héja 2005). A mondatbeli legfelsőbb szintű NP-k esetragjaiból kiindulva tanulmányoztuk, hogy egy adott eset milyen predikátumok mellett jelenhet meg, és ott milyen szintaktikai funkciót testesít meg, illetve milyen szemantikai szerepet kódol. Amint az előző szakaszban leírtuk, a szintaktikai és a szemantikai szintet nem különítjük el egymástól. Egy szintaktikailag elemzett tesztkorpusz segítségével a *-val* esetrag valamennyi előfordulását megvizsgáltuk, és célul tűztük ki egy olyan szabályrendszer implementálását, amely tetszőleges mondatban funkciócímkét társít a *-val* esetragot viselő, az igével közvetlen dependenciaviszonyban lévő főnévi csoportokhoz.

A szabályok által kiosztott címkék szemantikai tartalommal rendelkeznek, de magukat a szabályokat szintaktikai műveletként fogjuk fel, amelyek lehetővé teszik az adott NP használatát az aktuális kontextusban. Cikkünkben ismertetjük a módszert, amelynek segítségével elkülönítjük az esetragok funkcióit, valamint leírjuk és az automatikus szintaktikai elemzésben megvalósítjuk az adjunk-

ciós szabályok rendszerét. Munkánk eredménye egy kritériumrendszer, melynek segítségével a vonzatok elkülöníthetők a produktívan használt, kompozicionális szerkezeteket alkotó adjunktumoktól.

Az esetragokról feltételezzük, hogy rendelkeznek saját szintaktikai és szemantikai tulajdonságokkal, amelyek szabályokkal leírhatók. Ezeket az általános szabályokat, amelyek az esetragok alapértelmezett funkcióit definiálják, default szabályoknak nevezzük. A default szabályok bemenete utalhat az őt tartalmazó főnévi csoport fejének szemantikai vagy morfoszintaktikai tulajdonságaira, hiszen egy esetrag többféle produktív adjunktumfunkciót is kódolhat. Mivel a szerepeket az NP-hez társító szabályokat szintaktikai (adjunkciós) szabálynak fogjuk fel, a szabályok kimenetében megjelenő szerepcímkék is szintaktikainak tekinthetők. Itt azonban fontos megjegyezni, hogy a szerepek erős szemantikai tartalommal bírnak, valamint a szabályok jellegéből is kiderül, hogy egyes adjunkciós műveletek szemantikailag megszorított bemeneten működnek.

Azok az [ige + NP + esetrag] szerkezetek, melyek nem írhatók le általános (nem egyetlen lexikai tételen működő) szabályokkal, [ige + vonzat] szerkezetként elemzendők. Azért nem rendelhető hozzájuk szabály, mert ezek a szerkezetek nem kompozicionálisak: az NP igéhez képesti szerepét nem lehet olyan szemantikai címkével ellátni, amely nem utal az ige jelentésére, azaz általánosítható.

Vannak azonban olyan esetragos szerkezetek is, amelyek köztes kategóriát képviselnek a produktív adjunkció és lexikális vonzat-fogalom között. Az esetragok ezen használatai csak a szintaktikailag releváns szemantikai igeosztályok mellett mondhatók produktívnak. A kísérlet fontos hozadéka a szintaktikai–szemantikai funkciót annotáló szabályok létrehozása mellett, hogy a *-val* esetrag eloszlásából kiindulva megpróbáltuk feltérképezni ezeket a szemantikai igeosztályokat.

Az esetrag egyes, részlegesen produktív, ám szemantikailag elkülönülő funkciói meghatározzák a számunkra érdekes igecsoport-jelölteket. Mivel a szemantikai szerep, illetve az igei jelentéskomponensek közvetlenül nem hozzáférhető, illetve nem tesztelhető fogalmak, nagyon fontos, hogy egyértelmű szintaktikai tesztekkel igazoljuk az általunk használt szemantikai jegyeket. Levinhez hasonlóan szintaktikai alternációkat kívánunk használni a csoportok szintaktikai relevanciájának igazolására. Ezen a ponton felmerül tehát a kérdés, milyen alternációk jelemzik a magyar predikátumosztályokat? Alternációnak tekinthető-e például a szintaktikai környezet igeképzőkhöz köthető megváltozása (pl. passziválás), vagy a szintaktikai környezet igekötőhöz köthető megváltozása (pl. *rá-/odaken vmit vmire* vs. *meg-/összeken vmit vmivel*)? Alternációnak tekinthető-e az olyan esetragváltozás egy adott összetevőn, amely a szóban forgó mondat igazságfeltételeit nem érinti (*meglepődött/kiakadt/megdöbbent a polgár-*

mesteren/polgármestertől)? Cikkünkben a példákban említett szintaktikai jelenségek alternációként való kezelése mellett foglaltunk állást. Ennek gyakorlati oka, hogy a szemantikai szerepek annotálása szempontjából releváns általánosításokat tesznek lehetővé, és a lexikont egyértelmű szintaktikai kritériumok alapján bővíthetjük általuk. Elméleti szempontból a bővítménykeret-alternációk megjelenése a magyar nyelvben kétségtelenül további vizsgálatokra szorul, ám a szabályainkban megfogalmazott általánosítások így is lehetővé teszik a szintaxis–szemantika interfész egy újabb aspektusának vizsgálatát. Hangsúlyozzuk, hogy kísérletünk célja a szintaktikailag releváns szemantikai tulajdonságok és az ezen alapuló igecsoportok feltérképezése, valamint a bővítményeként megjelenő főnévi csoportok szintaktikai/szemantikai funkciójának automatikus annotálása tetszőleges korpuszban. Ezzel szemben nem törekszünk az ige és bővítménye által alkotott szerkezet valamennyi szintaktikai tulajdonságának megjóslására (pl. a bővítmény kötelezősége, szemantikai szelekció kimerítő leírása, a szerkezetnek különböző szintaktikai konstrukciókban való részvétele/transzformációk alkalmazhatósága, kontrollviszonyok leírása stb.).

5.2. Szemantikai igecsoportok automatikus kinyerése

Mivel az esetragok funkcióinak kimerítő leírása rendkívül munkaigényes feladat, következő kísérletünk (Gábor–Héja 2007) a szintaktikailag releváns szemantikai igeosztályok automatikus kinyerésére irányult. A kutatás célja tehát azoknak a lexikai-szemantikai tulajdonságokkal meghatározható igeosztályoknak az azonosítása volt, amelyek relevánsak a magyar igei bővítménykeretek leírása szempontjából. Mivel nem állt rendelkezésünkre előzetes csoportosítás, nem felügyelt tanulási módszerhez folyamodtunk. A nem felügyelt csoportosítási, azaz klaszterezési eljárás alapját Schulte im Walde (2000) módszere képezte. A Szeged Treebank (Csendes et al. 2004) 150 leggyakoribb igéjét soroltuk csoportokba hierarchikus agglomeratív klaszterezési eljárással szintaktikai bővítménykereteik alapján. A kísérlet kettős célt szolgált: egyrészt magát a tanulási módszert akartuk tesztelni, vagyis arra kerestük a választ, hogy bővítménykeret-információ alapján kinyerhetők-e szemantikailag koherens osztályok a korpuszból. Amennyiben igen, úgy a kísérlet megerősíti a Semantic Basis Hypothesist, hiszen alátámasztja, hogy a szemantikailag hasonló igék szintaktikailag is hasonló viselkedést mutatnak. Másrészt azt akartuk megtudni, hogy melyek azok a szemantikai jelentéskomponensek, amelyek köré az alapvető igeosztályok szerveződnek. Előfeltevésünk szerint a leggyakoribb igékből előálló csoportok tükrözni fogják a legalapvetőbb igei jelentéskomponenseket.

Az igéket a Szeged Treebankből kinyert teljes bővítménykeretekkel jellemeztük. A bővítménykeretekben előforduló főnevek lemmáit nem, csak a bővítménykeretben előforduló frázisok fejének szófaját, névszói frázisok esetében a fej esetragját vagy névutóját vettük figyelembe. Az ige reprezentációja az ige és a különböző bővítménykeretek együttes előfordulásainak maximum likelihood becsléséből állt elő:

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

ahol *f(v)* az ige gyakorisága, *f(v,t)* pedig az ige és a keret együttes gyakorisága. Az értékeket a 150 igére és mind a 839 bővítménykeretre kiszámoltuk. Az igei előfordulásokat reprezentáló, bővítménykereteken számított valószínűségi eloszlások összehasonlításához távolsági mértékként a relatív entrópiát használtuk:

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

Ezután két hierarchikus agglomeratív klaszterezési eljárást alkalmaztunk az adatokra: az egyikben a klaszterek elemszámának, a másikban a klaszter elemei közti maximális távolságnak határoztuk meg a maximális értékét.

Természetesen az első módszer, vagyis a csoporton belüli igék számának korlátozása kevesebb klasztert eredményez és értékesebb eredményeket ad a kevésbé gyakori igék esetében. Ezzel szemben a második módszer, vagyis az egy csoportba sorolt igepárok közti maximális távolság korlátozása hatékonyabb az alapvető, nagy elemszámú igeosztályok meghatározására. Mivel az a célunk, hogy Levinéhez hasonló igeosztályokat találjunk a magyar nyelvben, a következő lépés annak megvizsgálása volt, hogy az igeosztályok koherensek-e szemantikailag, és ha igen, elemeik milyen jelentéskomponensekben osztoznak. Általánosságban elmondható az eredményül kapott igeosztályokról, hogy lehorgonyozhatók valamilyen szemantikai jelentéskomponenshez, vagy jól jellemezhetők valamely argumentumuk közös szemantikai szerepével. Például: állapotváltozást jelentő igék: *erősödik, gyengül, emelkedik*; beneficiens argumentummal rendelkező igék: *biztosít, ad, nyújt, készít*; képességet jelentő igék: *tud, lehet, sikerül*; mozgást jelentő igék: *indul, elindul, jön, megy, kimegy, elmegy*; létezést jelentő igék: *van, nincs, lesz, marad*.

Kevésbé egyértelmű az alábbi, az egyszerűség kedvéért modálisnak nevezett csoport igéinek közös szemantikai komponense: *próbál, megpróbál, szokik, szeret, akar, elkezd, fog, kíván, kell*.

Néhány csoportot a fentieknél specifikusabb metapredikátummal jellemezhetünk – például külön csoportot alkotnak a megjelenést, illetve az ítélkezést jelentő igék. Más esetekben viszont a szemantikai reláció sokkal kevésbé szoros,

mint például az „folytonos cselekvést jelentő igék" címkével leírható csoport esetében: *ül, áll, lakik, dolgozik*. A skála másik végén elhelyezkedő klaszterek olyan igéket tartalmaznak, amelyek szemlátomást nem osztoznak közös jelentéskomponensben, pusztán „véletlenül" ugyanolyan esetraggal járó vonzatuk miatt kerültek egy csoportba: *foglalkozik, találkozik, rendelkezik*.

Az osztályokat nemcsak szemantikai koherenciájuk szerint értékelhetjük, hanem vizsgálnunk kell azt is, hogy rendelhető-e hozzájuk az osztály igéire jellemző szintaktikai alternáció, valamint igaz-e rájuk az a korábbi feltételezésünk, hogy az azonos esetraggal megjelenő adjunktumok szemantikai szerepe az osztályon belül állandó. Kiindulásként megvizsgáltuk, milyen szemantikai szerepeket kódoló bővítmények megjelenését engedélyezik az egyes csoportok igéi. Ez azt jelenti, hogy az igeosztályokat mátrixokkal ábrázoltuk, amelyeknek az oszlopait a főnévi esetragok, sorait az igei lemmák töltik ki, a cellákban pedig az a szemantikai szerep áll, amelyet az adott esetraggal megjelenő bővítmény az ige mellett betölt. Az osztály akkor koherens, ha a hozzá tartozó mátrix megfelel az alábbi két követelménynek:

(18) a. A mátrix specifikus az adott igeosztályra.

b. Az egy oszlopba tartozó cellák ugyanazt a szerepet tartalmazzák.

Mivel nem volt előfeltételezésünk az eredményként várt igeosztályokról, és nem voltak az értékeléshez használható, kézzel kialakított csoportjaink sem, az eredmények kiértékeléséhez nyelvészeti elemzésre volt szükség. Mindazonáltal az első értékelés alapján azt mondhatjuk, hogy a kapott osztályok meglepően erős szemantikai koherenciát mutatnak. Figyelembe kell vennünk továbbá, hogy a biztató eredmények rendkívül kis korpusz használatával születtek, ami megerősíti, hogy a Semantic Basis Hypothesis jó eredménnyel használható szemantikai osztályok automatikus kinyeréséhez. Fontos feladatunk a jövőben az igeosztályok nyelvészeti elemzése, amely az eredmények kiértékelésével szorosan összefügg.

6. Egy alkalmazás: igék jelentésegyértelműsítése

Az alábbi szakaszban először azt mutatjuk be, hogy az igék jelentésegyértelműsítése milyen problémákat vet fel, majd pedig vázoljuk, hogy az általunk javasolt strukturált igei adatbázis milyen megoldást képes nyújtani.

A jelentésegyértelműsítés *(word sense disambiguation: WSD)* célja, hogy a szövegben található szóelőfordulásokhoz a megfelelő jelentést társítsa. A WSD két fő lépésre bontható. Először is szükség van a lehetséges jelentések egy listájára,

például egy WordNetre vagy egy megfelelő egynyelvű szótárra. Továbbá ki kell választani azt az algoritmust, amely a jelentéstárban szereplő jelentések közül a megfelelőt rendeli a célszóhoz.

A WSD-rendszerek kiértékelése során célszerű figyelembe venni, hogy a rendszer mennyire teljesíthet jól egyáltalán. Általában nem várjuk, hogy számítógépes eszközökkel pontosabban oldjuk meg ezt a feladatot, mint ahogyan az emberek megoldanák. Érdekes, hogy több kísérlet is azt mutatta, hogy még az emberi intelligenciára támaszkodva is nehéz feladat a megfelelő jelentés hozzárendelése egy adott kontextusban szereplő célszóhoz (pl. Véronis 2003; Héja et al. 2009).

A következő kísérlettel arra a kérdésre kerestük a választ, hogy miért nehéz feladat az igei WSD. A kísérletben (Héja 2008) a Princeton WordNet 3.0-t (PWN 3.0) (l. 2.2.1) használtuk jelentéstárként. A PWN 3.0 igei csomópontjaiban a synsetekhez tartozó lemmákon kívül a jelentés és a jelentés azonosítója van megadva példamondatokkal együtt. Az alábbiakban egy példát láthatunk PWN 3.0 egy csomópontjára:

> ***introduce, present, acquaint****(SYNSET_ID)(cause to come to know personally) "permit me to acquaint you with my son"; "introduce the new neighbors to the community"*
>
> A synset magyar megfelelője:
> bemutat, összeismertet (személyes ismeretséget hoz lére) „engedje meg, hogy összeismertessem a fiammal"; „mutasd be az új szomszédokat a közösségnek"

A PWN 3.0-ban 11529 igei lemmához 25047 jelentés tartozik[4] és összesen 10759 példamondat. A kísérlet során első lépésben a példamondatok mély szintaktikai elemzését végeztük el a Xerox Incremental Parser-ral (Aït-Mokhtar–Chanod 1997). Az elemzés eredményéből automatikusan jelentésegyértelműsítő szabályokat generáltunk. A jelentésegyértelműsítő szabályok a példamondathoz hasonló mondatokhoz rendelték a megfelelő synset azonosítóját.

Így például a *check* 'féken tart' synsetből a (19)-ben látható egyértelműsítő szabályt hoztuk létre:

> **Check** (1131473)(hold back, as of a danger or an enemy; check the expansion, influence of) "*Check the growth of communism in South East Asia*";
>
> A synset magyar megfelelője:
> Féken tart (1131473)(feltartóztatja a veszélyt, az ellenséget; feltartóztatja valaminek az elterjedését, a befolyását) „*Féken tartja Dél-kelet Ázsiában a kommunizmus növekedését*"

[4] Az egyjelentésű lemmákat figyelmen kívül hagyva az átlagos poliszémia 3.57.

(19) if((OBJ (check, growth]) & PREP[IN] (check, SOMEWHERE]))
⇒ SYNSET(check, 1131473).

Az egyértelműsítő szabály azt mondja ki, hogy ha a *check* igének a *growth* lemmájú főnév a tárgya és az *in* prepozícióval egy HELY szemantikai jegyű NP is szerepel a mondatban, akkor a *check* ezen mondatbeli előfordulása az egynyelvű jelentéstárnak az 1131473 azonosítójú igéjéhez rendelhető.

Harmadik lépésben egy tesztkorpuszt elemeztünk az automatikusan generált egyértelműsítő szabályokkal. Tesztkorpuszként a SEMCOR 2.1[5] egy részkorpuszát használtuk, amely 17308 mondatot és 41497 igét tartalmazott. Az elemzés eredményeképpen az egyértelműsítő szabályok a PWN 3.0 példamondatok alapján hozzárendelték a megfelelő synset azonosítót a tesztkorpuszban szereplő igékhez. Például a (20)-ban található példmondatra illeszkedett a (19)-es szabály.

(20) The White House is taking extraordinary steps to **check** the rapid **growth** of juvenile delinquency **in the United States**.
'A Fehér Ház különleges lépéseket tesz, hogy féken tartsa az Egyesült Államokban a fiatalkori bűnözés gyors növekedését.'

Így a szabályok a 1131473 synset azonosítót rendelték a (20)-ban szereplő *check*-hez.

Mivel az egyértelműsítő szabályok rendkívül specifikusak voltak, azaz a PWN 3.0-ban szereplő példamondatok teljes elemzésén kívül a bővítménykeretek fejének lemmáira is hivatkoztak, azt vártuk, hogy a jelentésegyértelműsítés eredményeképpen kevés, de jól egyértelműsített igét kapunk. Érdekes módon az eredmények sokkal rosszabbak lettek a vártnál. Baseline-ként egy olyan módszert használtunk, amely a leggyakoribb jelentést rendeli a célszóhoz (ez az első jelentés a PWN 3.0-ban). A baseline módszerrel az összes ige 35,6%-ához rendeltük a megfelelő jelentést. Ezzel szemben az általunk generált egyértelműsítő szabályok csak 296 igeelőforduláshoz rendeltek jelentést, amelyeknek mindössze 30,7%-a volt megfelelő. Így eredményeink azt mutatják, hogy az általunk generált szabályoknak nemcsak a lefedettsége rosszabb, mint a baseline módszer, de a pontossága is.

Az alacsony fedés összhangban van a kezdeti elvárásainkkal, de kérdésként merül fel, hogy mi lehet az alacsony pontosság oka.

Első lépésben ellenőriztük a szintaktikai elemző pontosságát. A PWN 3.0 példamondatokból generált szabályokkal a PWN 3.0 példmondataiban található igéket egyértelműsítve 98,7%-os pontosságot kaptunk. A PWN 3.0 mondatokon

[5] http://www.cse.unt.edu/~rada/downloads.html

elért eredmények azt sugallják, hogy a módszernek nagyobb pontosságú egyértelműsítést kellene eredményeznie.

A következő lépésben a szabályokat a tesztkorpuszon lefuttatva 100 találatot vizsgáltuk kézzel. Azt láttuk, hogy az alacsony pontosság legfőbb oka az, hogy a tesztkorpusz nincsen jól annotálva a PWN 3.0 jelentésekkel, így pl. a (20) tesztmondatban található *check*-hez más jelentést társított az annotátor.

Az alábbiakban röviden áttekintjük, hogy milyen nehézségekbe ütközik egy koherens jelentésannotációval rendelkező tesztkorpusz létrehozása. Feltételezzük, hogy egy tesztkorpusz jelentésannotációja akkor koherens, ha a kódolók mindig ugyanazt a jelentést rendelik ugyanahhoz a célszóhoz.[6] A jelentésegyértelműsítés azokban az esetekben könnyű feladat, amikor homonim jelentésekkel van dolgunk (l. pl. Manning–Schütze 1999). Vagyis a poliszém jelentések esetében nagyobb a bizonytalanság abban, hogy egy szóelőfordulást melyik jelentéshez rendeljünk. Végül abban az esetben, ha az egyes jelentések átfednek, az annotáció csak a véletlenen múlik. Például a *say* ige esetében a PWN 3.0 megkülönbözteti az alábbi két jelentést: (1) *express in words "He said that he wanted to marry her"* (szavakban kifejez „Azt mondta, hogy el akarja venni") (2) *report or maintain "He said it was too late to intervene in the war"* (jelent, kijelent „Kijelentette, hogy túl késő már, hogy beavatkozzanak a háborúba"). Ebben a példában már a példamondatokban sem egyértelmű, hogy az ige melyik jelentéssel szerepel. Ez felveti azt a kérdést, hogy szükséges-e egyáltalán megtenni egy olyan megkülönböztetést, amiről nem tudjuk egyértelműen eldönteni, hogy mikor kell használnunk.

A fentiekből következik, hogy egy jelentésegyértelműsítésre szolgáló adatbázisban csak olyan jelentéseket szabad szerepeltetni, amelyek közül minden célszó esetében egyértelműen kiválaszthatunk egyet. Ameddig ilyen adatbázis nem létezik, addig nem várhatjuk hatékonyan működő WSD alkalmazások megjelenését. Egy megfelelő adatbázistól azt várjuk, hogy a jelentéseket megfigyelhető felszíni adatokhoz horgonyozza le, és ennek megfelelően válassza szét a világismeretből, illetve egyéni intuícióból származó jelentéselemeket a nyelvi disztribúcióhoz köthető jelentéskomponensektől, és a jelentésmegkülönböztetés során csak ez utóbbit vegye figyelembe.

Állításunk szerint a szintaktikailag lehorgonyzott szemantikai igeosztályok alkalmasak arra, hogy igei jelentéstárként szolgáljanak. Az ige jelentésének szintaktikailag lehorgonyzott, alternációhoz köthető része a metapredikátum, a világismeretet, illetve egyéni intuíciót tartalmazó jelentéskomponens pedig a magjelentés. Így egy jelentésegyértelműsítésre szolgáló igei adatbázisba akkor kell

[6] Vagy legalábbis magas az annotátorok közötti egyetértés a célszó jelentése vonatkozásában.

egy igéhez több jelentést felvenni, ha az igealak több különböző alternációban is részt vesz.

Ezt a következő példával szemléltethetjük. Az 5.2. pontban leírt kísérletben előálló *akar kíván tud szeret* igecsoport koherens osztályt alkot, hiszen ezek az igék ugyanolyan bővítménykeretekkel jelenhetnek meg: a bővítményük lehet tárgy, infintívusz, valamint *hogy*-os mellékmondat is. Ezzel szemben csak akkuzatívuszi és datívuszi bővítménnyel csak a *kíván* jelenhet meg ebből az osztályból.

A fentiek alapján a *kíván*-nak két jelentését kell megkülönböztetni:

Kíván$_1$, amely az *akar, szeret, tud* igékkel közös metapredikátumokkal jellemezhető, és az alábbi példamondatokban fordulhat elő:

(21) a. A Világbank részletes adatokat *kívánt*.
b. Éva egyre nagyobb adagokat *kívánt*.

(22) a. A bank elnökével *kíván* beszélni számlanyitással kapcsolatban.
b. Lewis csak azt *kívánta*, hogy ő legyen az egyetlen az életemben.

Noha intuitíve úgy érezhetjük, hogy a (21a–b) példamondatban szereplő két *kíván* jelentése eltér, érvelésünk szerint csak abban az esetben különböztethetjük meg őket, ha találunk olyan további alternációkat, amelyek a jelentésmegkülönböztetést szintaktikailag indokolttá teszik. Ellenkező esetben a jelentésegyértelműsítésre szolgáló adatbázis koherenciája romlana, hiszen nem lehetünk benne biztosak, hogy a pusztán szemantikai definíció konzisztens jelentésmegkülönböztetés alapját képezheti.

A *kíván*$_2$ megkülönböztethető a *kíván*$_1$-től, hiszen a *vki kíván vkinek vmit* bővítménykeretben jelenik meg:

(23) Bartalis Saroltának örömteljes névnapot kíván nagytatája, Fogarasi Géza Csíkcsicsóból.

Ebben a részben egy jelentésegyértelműsítési kísérletet ismertettünk, amelynek eredménye azt mutatta, hogy a Princeton WordNet 3.0, illetve a hozzá hasonló, pusztán szemantikai információra támaszkodó jelentéstárak nem használhatók eredményesen sem emberi munkaerő általi, sem automatikus jelentésegyértelműsítésre. Ennek oka, hogy a kizárólag szemantikai információra alapozó jelentésmegkülönböztetés gyakran önkényes. Azt javasoltuk, hogy az igei jelentésegyértelműsítési feladathoz használt jelentéstárat szintaktikai alternációkon alapuló predikátumosztályok felhasználásával hozzuk létre. Az igei jelentés komponenseihez társított szintaktikai tulajdonságok a jelentés egyértelműbb azonosítását teszik lehetővé, így csökkentve az emberi intuíció szerepét az alapvetően szemantikai természetű nyelvészeti feladat megoldásában.

7. Összefoglalás

Cikkünkben az igei bővítménykeret lexikális reprezentációjának kérdéskörét vizsgáltuk több szempontból. Egyfelől az argumentumrealizációs elméletek eredményeit felhasználva megfogalmaztuk azt az igényt, hogy strukturált lexikai reprezentáció útján predikátumok természetes osztályaira hivatkozva fogalmazzunk meg általánosításokat, és különítsük el a produktív/megjósolható jelenségek körét az egyes igék idioszinkratikus tulajdonságaitól. A természetes osztályok leírásakor szintaktikai teszteket keresünk, melyekkel egy adott ige besorolása egyértelműen eldönthető. Másfelől a számítógépes lexikonok iránt megfogalmazódó követelményekből kiindulva is arra jutottunk, hogy az explicitás és a koherencia kritériumának akkor felelhet meg egy lexikai adatbázis, ha tartalmazza a releváns általánosításokat, és ugyanakkor az általánosítások hatóköre és következményei egyértelműen levezethetőek és szintaktikai tesztekkel igazolhatók. Harmadrészt, a magyar nyelvre vonatkozó (illetve vonatkoztatható) vonzat-definíciók és tesztek tárgyalása során beláttuk, hogy ezek nem teszik lehetővé, hogy koherens lexikai adatbázist készítsünk a segítségükkel. A vonzatok azon tulajdonsága, hogy egy bizonyos ige lexikai tételéhez köthetők, ismét elvezet a produktivitás kérdéséhez, így kimondhatjuk, hogy a produktív szintaktikai–szemantikai műveletek és a hozzájuk kapcsolódó általánosítások meghatározása előfeltétele egy valóban koherens lexikon megalkotásának. Azt javasoltuk tehát, hogy a Nyelvtudományi Intézet számítógépes igei lexikai adatbázisának kibővítését a szintaktikailag releváns szemantikai predikátumosztályok meghatározásával kezdjük. Bemutattunk egy, a predikátumosztályok automatikus kinyerésére irányuló kísérletet, valamint egy másik megközelítést, amely a főnévi csoportok mondatbeli szintaktikai és szemantikai funkciójának vizsgálatából kiindulva próbálja meghatározni a produktívnak tekinthető bővítmények körét. Végül a jelentésegyértelműsítés példája alapján amellett érveltünk, hogy a predikátumok szintaktikai bővíthetőségük szerinti csoportokba sorolása olyan osztályozást eredményez, amely felhasználható e nyelvtechnológiai feladat aktuális problémáinak megoldásához.

Irodalom

Abeillé, Anne – Marie-Hélène Candito 2000. FTAG : A lexicalized tree adjoining grammar for French. In: Anne Abeillé – Owen Rambow (szerk.): Tree Adjoining Grammars. Stanford: CSLI. 305–330.

Aït-Mokhtar, Salah – Jean-Pierre Chanod 1997. Incremental finite-state parsing. In: Proceedings of Applied Natural Language Processing 1997. Washington, DC: Association for Computational Linguistics. 73—79.

Alberti Gábor – Kilián Imre 2010. Vonzatkeretlisták helyett polaritásos hatáslánccsaládok – avagy a ℜeALIS σ függvénye. In: Tanács Attila – Vincze Veronika (szerk.): A VII. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 113–127.

Baker, Collin F. – Charles J. Fillmore – John B. Lowe 1998. The Berkeley FrameNet project. In: Proceedings of the COLING-ACL, Montreal, Canada. Montreal: Association for Computational Linguistics. 86–90.

Baker, Mark 1988. Incorporation: A theory of grammatical function changing. Chicago: University of Chicago Press.

Bloomfield, Leonard 1933. Language. New York: Holt, Rinehart and Winston.

Brent, Michael R. 1991. Automatic acquisition of subcategorization frames from untagged text. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL). Berkeley, CA: Association for Computational Linguistics. 209–214.

Brent, Michael R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. Computational Linguistics 19: 243–262.

Bresnan, Joan W. – Annie Zaenen 1990. Deep unaccusativity in LFG. In: Katarzyna Dziwirek – Patrick Farrell – Errapel Mejias-Bikandi (szerk.): Grammatical relations: A cross-theoretical perspective. Stanford CA: CSLI Publications. 45–57.

Briscoe, Ted – John Carroll 1993. Generalized probabilistic LR parsing of natural language (corpora) with unication-based grammars. Computational Linguistics 19: 25–59.

Briscoe, Ted – John Carroll 1997. Automatic extraction of subcategorization from corpora. In: Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97). Washington, DC: Association for Computational Linguistics. 356–363.

Butt, Miriam 2006. Theories of case. Cambridge: Cambridge University Press.

Chomsky, Noam 1981. Lectures on government and binding. Dordrecht: Foris.

Copestake, Ann 1993. The representation of lexical semantic information. Doctoral dissertation, University of Sussex.

Csendes, Dóra – János Csirik – Tibor Gyimóthy 2004. The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. Lecture Notes in Artificial Intelligence 3206: 41–48.

Eynde, Karel van den – Piet Mertens 2003. La valence: l'approche pronominale et son application au lexique verbal. French Language Studies 13: 63–104.

Fellbaum, Christiane (szerk.) 1998. WordNet: An electronic lexical database. Cambridge MA: MIT Press.

Gábor, Kata – Enikő Héja 2005. A rule-based analysis of complements and adjuncts. In: Radovan Garabík (szerk.): Proceedings of the third international seminar on computer treatment of Slavic and Eastern-European languages, Bratislava. Bratislava: Slovenská Akadémia Vied. 37–50.

Gábor Kata – Héja Enikő 2006. Predikátumok és szabad határozók. In: Kálmán (2006a, 11–28).

Gábor, Kata – Enikő Héja 2007. Clustering Hungarian verbs on the basis of complementation patterns. In: John A. Carrol – Antal van den Bosch – Annie Zaenen (szerk.): Proceedings of the ACL'07 conference, Prague. Prague: Association for Computational Linguistics. 91–96.

Gábor Kata – Héja Enikő – Kuti Judit – Nagy Viktor – Váradi Tamás 2008. A lexikon a nyelvtechnológiában. In: Kiefer Ferenc (szerk.): Strukturális magyar nyelvtan 4. A szótár szerkezete. Budapest: Akadémiai Kiadó. 853–893.

Grimshaw, Jane 1990. Argument structure (Linguistic Inquiry Monograph 18). Cambridge MA: MIT Press.

Gross, Maurice 1975. Méthodes en syntaxe. Paris: Hermann.

Hajič, Jan – Barbora Hladká – Petr Pajas 2001. The Prague Dependency Treebank: Annotation structure and support. In: Steven Bird – Peter Buneman – Mark Liberman (szerk.): Proceeding of the IRCS Workshop on Linguistic Databases. Philadelphia: Univesity of Pennsylvania. 105–114.

Harris, Zellig 1954. Distributional structure. Word 10: 146–162.

Héja, Enikő 2008. The outlines of a hybrid approach to word sense disambiguation. Előadás. Intern's day, Xerox Research Centre Europe, Grenoble, 2008. június 23.

Héja Enikő – Gábor Kata 2008. A vonzatok és szabadhatározók elkülönítéséről. In: Sinkovics Balázs (szerk.): LingDok 7. Nyelvész-doktoranduszok dolgozatai. Szeged: JATEPress. 43–59.

Héja Enikő – Kuti Judit – Sass Bálint 2009. Jelentésegyértelműsítés – egyértelmű jelentésítés? In: Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): A VI. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 348–352.

Jackendoff, Ray 1983. Semantics and cognition. Cambridge MA: MIT Press.

Jackendoff, Ray 1990. Semantic structures. Cambridge MA: MIT Press.

Joanis, Eric – Suzanne Stevenson 2003. A general feature space for automatic verb classification. In: Proceedings of the 10th Conference of the EACL (EACL 2003). Budapest: Association for Computational Linguistics. 163–170.

Kálmán László (szerk.) 2006a. KB 120: A titkos kötet. Nyelvészeti tanulmányok Bánréti Zoltán és Komlósy András tiszteletére. Budapest: MTA Nyelvtudományi Intézet/Tinta Könyvkiadó.

Kálmán László 2006b. Miért nem vonzanak a régensek? In: Kálmán (2006a, 229–246).

Kiefer Ferenc 1992a. Az aspektus és a mondat szerkezete. In: Kiefer (1992b, 797–886).

Kiefer, Ferenc (szerk.) 1992b. Strukturális magyar nyelvtan 1. Mondattan. Budapest: Akadémiai Kiadó.

Kipper, Karin – Dang Hoa Trang – Martha Palmer 2000. Class-based construction of a verb lexicon. In: AAAI-2000 Seventeenth National Conference on Artificial Intelligence, Austin TX, July 30 – August 3, 2000. Austin, TX: AAAI. 691–696.

Koenig, Jean-Pierre – Anthony Davis 2003. Semantically transparent linking in HPSG. In: Stefan Mueller ed (szerk.): Proceedings of the HPSG03 Conference. Stanford: CSLI Publications. 222–235.

Koenig, Jean-Pierre – Anthony Davis 2006. The KEY to lexical semantic representations. Journal of Linguistics 42: 71–108.

Komlósy András 1992. Régensek és vonzatok. In: Kiefer (1992b, 299–527).

Korhonen, Anna – Ted Briscoe 2004. Extended lexical-semantic classification of English verbs. In: Dan Moldovan – Roxana Girju (szerk.): HLT-NAACL 2004 : Workshop on Computational Lexical Semantics, Association for Computational Linguistics, Boston, MA, May 2–7, 2004. Boston, MS: Association for Computational Linguistics. 38–45.

Korhonen, Anna – Yuval Krymolowski – Zvika Marx 2003. Clustering polysemic subcategorization frame distributions semantically. In: Erhard W. Hinrichs – Dan Roth (szerk.): Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan. Sapporo: Association for Computational Linguistics. 64–71.

Kuti, Judit – Károly Varasdi – Ágnes Gyarmati – Péter Vajda 2007. Hungarian WordNet and representation of verbal event structure. Acta Cybernetica 18: 315–328.

Lapata, Mirella 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternation. In: Robert Dale – Kenneth Ward Church (szerk.): Proceedings of the 37th Meeting of the Association for Computational Linguistics, College Park, MD. College Park, MD: Association for Computational Linguistics. 266–274.

Levin, Beth 1993. English verb classes and alternations. Chicago: University of Chicago Press.

Levin, Beth – Malka Rappaport Hovav 2005. Argument realization. Cambridge: Cambridge University Press.

Lightfoot, David W. 1979. Principles of diachronic syntax. Cambridge: Cambridge University Press.

Lopatková, Markéta 2003. Valency in the Prague Dependency Treebank: Building the valency lexicon. The Prague Bulletin of Mathematical Linguistics 79–80: 37–59.

Manning, Christopher D. – Hinrich Schütze 1999. Foundations of statistical natural language processing. Cambridge MA: MIT Press.

McCarthy, Diana 2001. Lexical acquisition at the syntax–semantics interface: Diathesis alternations. Doctoral dissertation, University of Sussex.

Merlo, Paola – Suzanne Stevenson 2001. Automatic verb classification based on statistical distributions of argument structure. Computational Linguistics 27: 373–408.

Miller, George A. 1995. WordNet: A lexical database for English. Communications of the ACM 11: 39–41.

Pinker, Steven 1989. Learnibility and cognition: The acquisition of argument structure. Cambridge MA: MIT Press.

Pustejovsky, James 1995. The generative lexicon. Cambridge MA: MIT Press.

Radford, Andrew 1988. Transformational grammar. A first course. Cambridge: Cambridge University Press.

Rákosi, György 2006. Dative experiencer predicates in Hungarian. Doctoral dissertation, Utrecht. Megjelent: LOT Dissertations, 146. kötet.

Reinhardt, Tanya 2002. The theta system: An overview. Theoretical Linguistics 28: 229.290.

Resnik, Philip 1993. Selection and information: A class-based approach to lexical relationships. Doctoral dissertation, University of Pennsylvania.

Riloff, Ellen – Mark Schmelzenbach 1998. An empirical approach to conceptual case frame acquisition. In: Eugene Charniak (szerk.): Proceedings of the Sixth Workshop on Very Large Corpora (WVLC-98). Montreal: Association for Computational Linguistics. 49–56.

Sass, Bálint 2007. First attempt to automatically generate Hungarian semantic verb classes. In: Proceedings of the 4th Corpus Linguistics Conference, Birmingham.

Schabes, Yves – Richard C. Waters 1993. Lexicalized context-free grammars. In: Lenhart K. Schubert (szerk.): 31st Annual Meeting of the Association for Computational Linguistics (ACL'93). Columbus OH: Association for Computational Linguistics. 121–129.

Schulte im Walde, Sabine 2000. Clustering verbs semantically according to their alternation behaviour. In: COLING. Saarbrücken: Morgan Kaufman. 747–753.

Schulte im Walde, Sabine – Chris Brew 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In: Pierre Isabelle (szerk.): Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA: Association for Computational Linguistics. 223–230.

Vendler, Zeno 1957. Verbs and times. Philosophical Review 66: 143–160.

Véronis, Jean 2003. Sense tagging: Does it make sense? In: Andrew Wilson – Paul Rayson – Tony McEnery (szerk.): Corpus linguistics by the Lune: A Festschrift for Geoffrey Leech. Frankfurt: Peter Lang. 273–290.

The lexical representation of verbs and language technology

Abstract: The present paper investigates the principles and criteria that have to be met while constructing a computational lexicon of verbs. We claim that two major conditions, coherence and explicitness need to be satisfied. The coherence of the lexicon should be ensured by the application of distributional complement tests, while explicitness entails an exact and unambiguous representation of the verbal entries. Since at this time there are no wide-coverage distributional complement tests for Hungarian, we propose a different approach which consists in reversing the order of the description: starting from the enumeration of productive structures, we will consider as complements those constituents for which no productive rules can be found. Our objective is to give a more precise description of productivity. The investigation helps us to make generalizations over the verbal complementation frames, thus increasing the coherence and explicitness of the available verbal database.

Keywords: verbal lexicon, lexical representation, complement test, compositionality, verb classes

A Budapesti Szociolingvisztikai Interjú lexikai és szintaktikai jellemzői*

Váradi Tamás – Oravecz Csaba – Peredy Márta

Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Budapest
varadi@nytud.mta.hu; oravecz@nytud.mta.hu; mperedy@nytud.mta.hu

A dolgozat célja a Budapesti Szociolingvisztikai Interjú (BUSZI) társalgási moduljainak lexikai és szintaktikai elemzése nyelvtechnológiai módszerekkel és ennek segítségével a szóbeli és írásbeli nyelvhasználat közötti különbségek kvantitatív jellemzése. Az elemzés a számítógépes elemzéssel annotált szövegeket elsősorban statisztikai eljárásokkal vizsgálja. A BUSZI társalgási nyelvhasználatát a Magyar Nemzeti Szövegtárból vett minta segítségével az írott nyelvhasználat jellemzőivel veti össze. Ahol erre mód nyílik, a BUSZI által vizsgált társadalmi csoportok közötti lexikális és mondatszerkesztésbeli nyelvhasználati különbségeket is vizsgálja a tanulmány. A bemutatott vizsgálatok nem kimerítőek, inkább kutatási irányokat, lehetőségeket mutatnak be, melyek magyar nyelvre még nagyrészt feltáratlanok, mindazonáltal az írásbeli nyelvhasználat lexikai gazdagsága, a közölt információ tömörítésére való törekvése a szóbeli változattal szemben egyértelműen, kvantitatív mérésekkel kimutatható.

Kulcsszavak: szóbeli és írásbeli nyelvhasználat, beszélt nyelv, korpuszösszehasonlítás, korpuszhomogenitás, jellemzőszó-vizsgálat, mondatszerkesztés, szintaktikai elemzés

1. Bevezetés

Jelen tanulmányban a Budapesti Szociolingvisztikai Interjúnak (Kontra 1990) a BUSZI2 (Váradi 2003) néven ismert részével foglalkozunk, mely öt foglalkozás szerinti társadalmi csoport nyelvhasználatát vizsgálja a szociolingvisztikai interjú Labov (1984) által kidolgozott módszerével. Ennek fontos eleme az irányított társalgás, melynek során a gondosan kiképzett terepmunkások kötelező, illetve tetszőlegesen választott témákat beszéltek meg az adatközlőkkel. A magnóra felvett anyag lejegyzése alapjában véve a helyesírási szabályokat követte, de célul tűzte ki a BUSZI vizsgálati kérdéseit tartalmazó szociolingvisztikai változók, illetve a beszéd prozódiai és paralingvisztikai kísérőjelenségeinek a gondos megörökítését is. Az eredetileg házi norma szerint kidolgozott annotáció az anyag tartalmi felülvizsgálata után XML szabványos alakra lett átalakítva.

* Jelen dolgozat a Váradi et al. (2010) tanulmány – annak szövegét gyakorlatilag teljes mértékben felhasználó, de részben átdolgozott és számos kérdésben kibővített – változata.

A szóbeli és írásbeli nyelvhasználat összehasonlítása fontos módszertani kérdéseket vet fel. Nehezen megvalósítható ugyanis az az egyébként ideálisnak tekinthető helyzet, hogy ugyanazon beszélő írott és beszélt nyelvhasználatát is vizsgálhassuk. Erre a BUSZI2 esetében már csak azért sincs mód, mert az interjúk nem tartalmaztak írásbeli komponenst, másrészt pedig az 1987-ben felvett interjúk alanyai ma már nagy számban elérhetetlenek. Ennél fogva az egyedi adatközlők szintjéről a nyelvhasználati változatok szintjére kell helyeznünk az összevetés alapját. Szerencsére a Magyar Nemzeti Szövegtár (MNSz) (Váradi 2002) összeállításában az írott nyelvhasználat a kizárólagos, ezért könnyen előállítható egy olyan válogatás, amely méretben referenciaként tud szolgálni. Abban az általános (a szóbeli és írásbeli nyelvhasználat különbözőségét általában jellemezni próbáló) perspektívában, amelyből a statisztikai alapú vizsgálatokat végeztük, ez a megközelítés az adott körülmények között elfogadhatónak tűnik.

A tanulmány célja az, hogy bemutassa, hogyan alkalmazható a nyelvtechnológia a különböző nyelvhasználati rétegek szociolingvisztikai vizsgálataiban. Nem egyetlen központi hipotézis eldöntése a szándék, hanem néhány kiszemelt kérdés vizsgálatához mutatja be a nyelvtechnológia hozzájárulását. A dolgozat két fő részre tagolódik: A 2. részben a lexikai vizsgálatok eredményeit mutatjuk be. A szokásos gyakorisági listák mellett kísérletet teszünk a szövegváltozat egyedi jellemzőit tükröző lexikai mintázatok feltárására, valamint azok korszerű módszerrel történő megjelenítésére is. A 3. rész olyan szintaktikai elemzéseket tartalmaz, melyekhez az adatbázis reguláris lekérdező nyelvén definiált lokális grammatikákat használtunk fel. A szófajok és a felszíni szerkezeti minták statisztikai síkon megragadható jellegzetességeit az írott nyelvhasználattal való összevetésben, illetve a BUSZI2 adatközlőcsoportjainak egymás közötti összehasonlításával mutatjuk be. Rövid összefoglalás zárja a dolgozatot a 4. részben.

2. Lexikai vizsgálatok

A lexikai vizsgálatok a szövegek szókincsére irányulnak. Elsősorban sűrűségi *(lexical density)* illetve gazdagsági *(lexical diversity/richness)* elemzéseket végeznek. Az előbbin általában a tartalmas szavak teljes szövegre vetített arányát (Ure 1971; Halliday 1985), utóbbin a szöveg szavainak változatosságát, különbözőségét kifejező valamilyen mértéket értenek. Mindkét esetben, de különösen a szókincsgazdagságot kifejező mutatók tekintetében évtizedek óta tartó kutatás irányul a lehető legelfogulatlanabb – és járulékos paraméterekre (pl. szöveghossz) legkevésbé érzékeny – mérték kidolgozására. A klasszikus mérték a sztenderd **típus/**

token arány,[1] illetve ennek különféle, a hosszfüggést kiküszöbölni próbáló változatai (Johnson 1944; Richards–Malvern 1997; McKee et al. 2000; Malvern et al. 2004), melyeket azonban számos kritika ért, éppen az említett szöveghossztól való függés miatt, ez ugyanis minden próbálkozás ellenére kimutatható maradt (Tweedie–Baayen 1998; McCarthy–Jarvis 2007; 2010).

Az aktuális mérték hosszérzékenysége akkor kritikus, amikor a rendelkezésre álló adatok nem normalizálhatók/sztenderdizálhatók megfelelően, ezért kulcskérdés a lehető legelfogulatlanabb mutató kidolgozása. Esetünkben ez a kényszer nem áll fenn, jelentős mennyiségű adat áll rendelkezésre, így a gyakorlatilag évtizedek óta vizsgált problémák itt nem jelentkeznek. A kvóták egymás közötti esetleges összehasonlításánál viszont már tekintettel kell lenni erre is. Meg kell jegyeznünk, hogy az itt bemutatott vizsgálatok csupán illusztratívak, és kutatási irányokat, lehetőségeket mutatnak be, melyek magyar nyelvre még nagyrészt feltáratlanok, ezért elsődleges céljuk bevezetőként, útmutatóként szolgálni további olyan részletes kvantitatív elemzésekhez, melyekhez számítógépes nyelvészeti eszközöket hatékonyan lehet felhasználni. Alapvető célunk a nyelvhasználati formák, illetve egyes beszélőcsoportok nyelvhasználatának lexikai gazdagságát és változatosságát vizsgálni, és kvantitatív mérőszámokkal igazolni azt a kiinduló feltételezést, hogy egyészt az írásbeli nyelvhasználat lexikai gazdagsága, változatossága nagyobb a szóbeli változaténál, másrészt hasonló különbségek az egyes társadalmi csoportok szóbeli változatai között is mérhetőek.

2.1. Szókincsgazdagsági vizsgálatok

Szókincsgazdagsági elemzésekben a típus/token arány mellett további mutatók is vizsgálhatók (pl. hapax[2] gyakoriság, dislegomenon[3] gyakoriság), melyek több alkalmazásban is gyakran használatosak, például szerző-, illetve műfajazonosításban (Stamatatos et al. 2000), nyelvelsajátítási, nyelvfejlődési, szókincsfejlődési, szociolingvisztikai vizsgálatokban (MacWhinney 2000; Johansson 2008), de megbízhatóságuk éppen az egyszerűségük miatt alacsony. Az egyes szövegtípusok szókincsére vonatkozó néhány szembetűnő különbség azért kiolvasható belőlük. A mérőszámok közül néhány nagyon egyszerű statisztikát foglal össze az 1. táblázat. A kvóták kódjai az alábbi adatközlő csoportokra vonatkoznak:

[1] Az egyes szótípusok halmazának, illetve összes előforduló példányaiknak a hányadosa.

[2] Egyszer előforduló elem.

[3] Kétszer előforduló elem.

KV1: tanárok; KV2: egyetemi hallgatók; KV3: bolti eladók; KV4: gyári munkások; KV5: szakmunkástanulók.

Jellemző	Korpusz						
	MNSz	BUSZI	KV1	KV2	KV3	KV4	KV5
1. szóalak	224128	173331	36846	29278	40994	37116	29097
2. szótípus	52876	26449	8971	6776	8639	8601	6560
3. típ./token	0,2360	0,1526	0,2435	0,2314	0,2107	0,2317	0,2255
4. normált	25000 szóalak						
5. szótípus	10140		6704	6048	5866	6283	5935
6. típ./token	0,4056		0,2682	0,2419	0,2346	0,2513	0,2374
7. főnév	6813		4109	3535	3299	3808	3070
8. ige	3904		4231	3869	4544	4362	4396
9. fn/ige	1,7451		0,9712	0,9137	0,7260	0,8729	0,6983
10. hapax	4402		2416	2082	1912	2189	2021
11. disleg.	1014		564	538	522	577	548

1. táblázat. Szóstatisztikai adatok a különböző szövegeken

A 3. sor magasabb típus/token aránya abszolút mértékben gazdagabb szókincset tükröz (többféle szó fordul elő adott nagyságú szövegben), viszont természetesen a korpusz növelésével a típusok száma nem nő arányosan, ezért a normált korpuszméretből (4. sor, 25 ezer szó) számított érték (6. sor) mutatja pontosan az írott és beszélt változat közötti eltérést ezen mutató tekintetében. Jól látható, hogy az MNSz szövegeit szignifikánsan magasabb érték jellemzi. Szembetűnő az eltérés a főnév/ige-használatban is, itt a szóbeli nyelvhasználat szövegeire mutatható ki egyértelműen az igék használatának magasabb aránya a főnevekéhez képest. Az egyszer, illetve kétszer használatos szótövek (10., 11. sor) előfordulási gyakoriságának különbsége is egyértelműen jelzi a írott változat nagyobb lexikális gazdagságát. Összefoglalásképpen megállapíthatjuk, hogy a írásbeli nyelvhasználat lexikai gazdagsága már egyszerű mérőszámokkal is világosan kimutatható, ugyanakkor a szóbeli változat dinamizmusát, kötetlenségét jellemzi a gyakori igehasználat.

Fontos megjegyezni, hogy a kvóták között is jelentkezik ugyan különbség a mérőszámokban, megbízható eredményekhez azonban részletesebb vizsgálatokra, illetve esetenként nagyobb mennyiségű szövegre lenne szükség, amely jelenleg még nem áll rendelkezésre.

2.2. Jellemzőszó-vizsgálatok eredményei

Számos lehetséges módszer közül (l. pl. Kilgarriff 1996; 2001) az alábbiakban egy olyan eljárás eredményeit mutatjuk be, amely az egyes korpuszok szógyakorisági profiljainak összehasonlításával határozza meg az adott szövegre jellemző lexikai elemeket. Ebben az összehasonlításban azok a nyelvi elemek szerepelnek a rangsor elején, amelyek vagy az egyik, vagy a másik korpuszban jellegzetesek, vagyis az eljárás egy közös listát generál, melyet utána kvalitatív vizsgálatnak lehet alávetni.

A vizsgálatban először a két korpusz nyers gyakorisági listáit állítjuk elő, majd minden listában szereplő szóra log-likelihood statisztikát számolunk (Rayson–Garside 2000). A teszt két modell (hipotézis) adatokhoz való illeszkedését hasonlítja össze, esetünkben azt, hogy a kérdéses szóalakok hasonló relatív gyakorisággal fordulnak elő a korpuszokban, illetve jelentősen eltérnek ebben a tekintetben. A magasabb számított érték jellemzi azokat a szavakat, melyeknek a relatív gyakorisága eltér a korpuszokban. Ha ezen eredmények szerint rendezzük újra a gyakorisági listát, a lista elején megkapjuk az egyik vagy másik korpuszra jellemző szavak halmazát.

MNSz-re jellemző szavak				BUSZI-ra jellemző szavak			
1962	a	C1: 21240	C2: 9687	6143	hát	C1: 107	C2: 4232
689	amely	C1: 696	C2: 13	3341	igen	C1: 118	C2: 2512
505	magyar	C1: 996	C2: 148	3273	én	C1: 307	C2: 2991
414	minden	C1: 402	C2: 5	2688	nem	C1: 2873	C2: 6672
342	kormány	C1: 337	C2: 5	2277	van	C1: 2274	C2: 5438
238	évfolyam	C1: 254	C2: 7	1435	szóval	C1: 21	C2: 973

2. táblázat. MNSz (C1) és teljes BUSZI (C2)

A 2. és 3. táblázatokban szereplő nem kimerítő, csupán illusztratív példákat tartalmazó listákban az első oszlop a számított súlyérték, a második a szó(tő), harmadik az egyik (C1), illetve másik (C2) korpuszbeli gyakorisági érték.

Az eljárás eredményei elnagyolva, de nagyon szemléletesen ábrázolhatók „szófelhők" formájában, melyeket az 1., 2., 3. és 4. ábra illusztrál. Az egyes lexikai elemek az alkalmazott statisztika szerint számított súllyal arányosan kiemelve jelennek meg az ábrákban, így rögtön szembetűnnek azok a nyelvi elemek, amelyek az adott nyelvhasználatra nagyon jellemzők, és jelentős túlsúllyal jelennek meg az összehasonlítás alapját képező másik szövegtípushoz viszonyítva.

KV1-re jellemző szavak				KV5-re jellemző szavak			
88	gyerek	C1: 108	C2: 12	326	hát	C1: 376	C2: 1044
53	ugye	C1: 59	C2: 5	91	akko	C1: 0,5	C2: 70
47	gimnázium	C1: 40	C2: 1	84	meg	C1: 146	C2: 347
42	tanít	C1: 50	C2: 5	68	szóval	C1: 46	C2: 162
42	tanár	C1: 36	C2: 1	61	például	C1: 16	C2: 94

3. táblázat. KV1 (C1) és KV5 (C2)

1. ábra. MNSz vs. teljes BUSZI

2. ábra. Teljes BUSZI vs. MNsz

3. ábra. KV1 vs. KV5

4. ábra. KV5 vs. KV1

2.3. Korpuszok homogenitásának és hasonlóságának vizsgálata

Korpuszok hasonlóságának vizsgálata, illetve mérése leginkább nyelvtechnológiai alkalmazások kontextusában merül fel és arra vonatkozóan adhat hasznos információt, hogy meghatározott típusú szövegekre fejlesztett alkalmazások milyen ráfordítással alakíthatók át újabb szövegek kezelésére. Ha a szövegek hasonlóak, a ráfordítás feltehetően kisebb. Általánosságban ennél többet nem is nagyon

lehet állítani, egyrészt a lehetséges mértékek is nagyon sokfélék, másrészt megbízhatóságuk is jelentősen függ az adott szövegektől, illetve alkalmazásoktól.

A hasonlóság kérdése szervesen összefügg egy adott korpusz homogenitásának kérdésével, hiszen nyilván nem egyértelmű, mit fejezne ki egy ilyen mérték abban az esetben, ha egy homogén korpuszt hasonlítunk össze egy általános, nagy változatosságot mutató heterogén korpusszal (a kérdés részletes tárgyalását l. Kilgarriff 2001). Ideális esetben ugyanazt a mértéket lehet használni korpuszon belüli és korpuszok közötti „távolságok" mérésére, melyek így közvetlenül összehasonlíthatók. A jelen vizsgálatokban alkalmazott perplexitásmérték használható ilyen módon, de természetesen – mint minden más egydimenziós mérték – csak durva közelítésként értelmezhető egy alapvetően többdimenziós összehasonlítási feladatban.

A perplexitás annak mérésére szolgáló érték, hogy mennyire pontosan tudunk modellezni egy ismeretlen valószínűségi eloszlást a rendelkezésre álló tanító adatok alapján. Minél magasabb értéket kapunk a perplexitásra, annál bizonytalanabbak vagyunk az ismeretlen eloszlás tekintetében, annál kevésbé tudjuk megjósolni az ismeretlen eloszlás által generált adatokat. Szövegek vizsgálatára az alábbi módon tudjuk felhasználni: a szöveg egy meghatározott részét (pl. 90%) használjuk fel tanító adatként, melyből például egy trigram (szóhármas) alapú statisztikai nyelvi modellt építünk (vagyis szóhármasok előfordulási gyakorisága alapján számoljuk ki a mondatok valószínűségét).[4] Ezek után lemérjük, hogy az elkészült modell alapján milyen pontosan tudjuk megjósolni a tesztadatként szereplő szövegrész (a maradék 10%) mondatait. Minél nagyobb perplexitásértéket kapunk, annál távolabb áll a tanító és tesztszöveg egymástól, minél kisebbet, annál homogénebb és annál jobban hasonlít a két szövegrész egymásra. Informálisan, a perplexitásra kapott számérték annak a szóhalmaznak a nagyságát határozza meg, amelyből (trigram nyelvmodell esetén) a megelőző két szó ismeretében a következő szót választhatjuk. Minél kisebb ez a halmaz, modellünk annál megszorítottabb (Jelinek et al. 1977). Esetünkben az egyes szövegtípusok önmagukban tekintett és egymáshoz képest mért változatosságáról kaphatunk információt ilyen jellegű vizsgálattal.

Az itt végzett vizsgálatok sztenderd tízszeres keresztvalidációval[5] készültek, a CMU-Cambridge Statistical Language Modeling Toolkit (Clarkson – Rosenfeld

[4] Természetesen – különösen szabad szórendű nyelvek esetén – ezeknek a modelleknek súlyos korlátaik vannak, de az itt használt egyszerű összehasonlító vizsgálatokra alkalmasnak tekinthetők.

[5] A 10 részre osztott szövegből 10 mérés során mindig más 9 rész tanító, és 1 rész tesztadatot veszünk, a végeredményt pedig a 10 mérés átlagaként képezzük.

1997) segítségével, a morfológiai variabilitásból eredő eleve magas értékeket kiküszöbölendő a szokásos gyakorlatnak megfelelően szótövesített szövegekkel. A kapott eredmények a 4. táblázatban láthatók. Az egyes sorokban szereplő szövegekből készült a trigram nyelvmodell, az oszlopok jelzik a tesztadatot. Abban a cellában, ahol mindkét, a sorban és oszlopban szereplő szöveg azonos, ott az adott korpusz homogenitására vonatkozó érték szerepel, a további cellákban pedig a különböző korpuszok hasonlóságát jellemző érték jelenik meg. Mivel a vizsgálat illusztratív, nem törekszik kimerítő jellemzésre, inkább a szembetűnő jellegzetességekhez kíván kvantitatív mérőszámot rendelni, ezért nem minden cellában szerepel (az egyébként minden esetben számítható) mutató. Néhány összehasonlítás a szövegek jellegéből következően nem hordoz lényeges információt, így azokat eleve nem érdemes elvégezni. Mivel az itt szereplő MNSz-minta jól láthatóan igen heterogén, nagy variabilitású szövegeket tartalmaz, a BUSZI-szövegekkel való összehasonlítás nem eredményezne újabb információt azon túl, hogy az írott szöveg a beszélthez képest sokkal változatosabb, ez pedig a homogenitási adatokból is egyértelműen látszik már. A BUSZI-szövegek vizsgálatában pedig informatívabb az egyes kvóták anyagát egymással összehasonlítani, mint a teljes BUSZI anyagot a kvóták anyagával; ez utóbbi esetben sem kapunk az előbbi vizsgálathoz képest új információt.

Tanító korpusz	Tesztkorpusz						
	MNSz	BUSZI	KV1	KV2	KV3	KV4	KV5
1. MNSz	733,618	—	—	—	—	—	—
2. BUSZI	—	121,52	—	—	—	—	—
3. KV1	—	—	123,835	123,462	113,633	118,273	107,597
4. KV2	—	—	122,666	115,782	—	—	—
5. KV3	—	—	124,402	—	101,542	—	—
6. KV4	—	—	130,106	—	—	108,828	—
7. KV5	—	—	127,512	117,237	110,695	116,796	89,401

4. táblázat. Perplexitásértékek a különböző szövegeken

Az egyes korpuszrészek, kvóták homogenitására vonatkozó értékből kiolvasható, hogy az adott kvótához tartozó beszélőknek mennyire változatos a nyelvhasználata. A kvóták egymással történő összehasonlításából kapott értékek arra adnak választ, hogy a kvóták szövegei mennyire állnak közel egymáshoz, illetve az egyik szöveg milyen mértékig „foglalja magában" a másikat. A KV1 és KV2 korpusz például ebben az összehasonlításban viszonylag távol esik egymástól, míg ha a KV5 korpuszhoz hasonlítjuk például a KV1 korpuszt, akkor jelentős távolságot

kapunk, fordított irányban pedig alacsonyat, vagyis a KV1 korpuszból épített modell „magában foglalja" a KV5 korpuszt is. Ezeknek az eredményeknek elsősorban az egyes stílusrétegek nyelvi adatainak specifikus számítógépes feldolgozása során van jelentőségük, mivel – kissé elnagyoltan fogalmazva – az adott nyelvtechnológiai alkalmazásban[6] használt nyelvi modellek a fenti információk birtokában a bemenő nyelvi adatokhoz pontosabban illeszthetők, illetve finomhangolhatók.

A 2. részben elvégzett lexikai vizsgálatok alapján egyszerű modellekkel és mérésekkel is jól kimutatható az írásbeli nyelvhasználat gazdagsága és változatossága a szóbelihez viszonyítva, illetve ilyen különbségek a vizsgált társadalmi csoportok szóbeli változatai között is megmutatkoznak. A jellemzőszó-vizsgálatok igazolják, hogy a tartalmas lexikai elemek, a főnévi csoportokhoz kapcsolódó névelők, választékos kötőszavak túlsúlya az írásbeli nyelvhasználatban egyértelmű, míg a szóbeli változat gazdag a töltelék- és egyéb kötőszavak használatában. Ez a sajátosság az egyes kvóták között is megjelenik, a képzettebb társadalmi csoportok nyelvhasználatát a tartalmas elemek jellemzik az alacsonyabb képzettségűekkel szemben.

3. Szintaktikai elemzések

A szóbeli és írásbeli nyelvhasználat sajátosságaival, az egyes nyelvi szinteken jellemző különbségekkel számos kutatás foglalkozott (l. pl. Lanstyák 2009 és az általa közölt részletes szakirodalmi listát), de a mondattani jelenségek széles körének jelentős méretű korpuszon alapuló, kvantitatív módon történő vizsgálata nemigen volt jellemző. Az MNSz és a BUSZI összehasonlításakor azt a szakirodalomban általánosan elfogadott nézetet kívánjuk ilyen vizsgálatokkal alátámasztani, amely szerint az írott nyelv sokkal inkább igyekszik tömöríteni az információt, mint a beszélt nyelv. A jelen pont kiinduló hipotézisét tehát ez az állítás képezi.

A mondatok szintjén ez azt jelenti, hogy az egyes mondatok jóval hosszabbak az írott korpuszban (3.1.), mert egyrészt több tagmondatból állnak (3.1.1.), másrészt az egyes tagmondatok is több bővítményt tartalmaznak, mint a beszélt nyelvben. A főnévi csoportok szintjén azt figyelhetjük meg, hogy alárendelt tagmondatok és még inkább önálló mondatok helyett gazdag jelzős szerkezeteket használ az írott nyelv (3.3.). A tömörítés és a gazdaságosság jele az is, hogy a redundáns, azaz elhagyható elemeket sokkal ritkábban használja az írott, mint a beszélt nyelv. Ezt bizonyítja a kétféle birtokos szerkezet összehasonlítása (3.3.4.),

[6] Például beszédfelismerés, statisztikus gépi fordítás, szófaji egyértelműsítés.

valamint a topikismétlő névmások használata (3.4.). A beszélt és az írott nyelvhasználat közötti különbség a névmások (l. 3.5., 3.6.) és a névelők (l. 3.3.2.) előfordulási arányaiban is megnyilvánul.

3.1. Mondathossz

A szintaktikai vizsgálatok alapegysége a mondat, így minden szintaktikai elemzés a mondathatárok megállapításával kell, hogy kezdődjön. Az írott nyelvi korpuszban ez nem jelent problémát, a beszélt nyelvi korpuszt tanulmányozva azonban talán éppen ez a korpusz elemzésének legbizonytalanabb pontja. A beszélők ugyanis (szemben az írott szövegek létrehozóival) nem jelzik egyértelműen, hogy hol van szerintük a mondataik vége. A BUSZI-korpusz tagolásánál a szöveget annotáló személyek anyanyelvi intuíciójuk, illetve a Németh T. (1991) által megfogalmazott elvek alapján állapították meg a mondathatárokat.

A két korpusz közti első szembetűnő különbséget az 5. táblázat mutatja. Az írott nyelvi anyag átlagos mondathossza (17,1 szó) kétszerese a BUSZI-adatközlők élőbeszédbeli mondataiénak (8,5 szó). A BUSZI-terepmunkások megszólalásainak célja elsősorban az adatközlők beszédének terelgetése volt, így nem meglepő, hogy az ő megszólalásaik még rövidebb mondatokra tagolódnak. (A teljes BUSZI-beli átlagos mondathossz 6,5 szó.)

	BUSZI		MNSz
	terepmunkások (tm)	adatközlők (ak)	
átlagos mondathossz	4,6	8,5	17,1

5. táblázat. Átlagos mondathossz

3.1.1. Ragozott igealakok – tagmondatok

A mondatszerkezet szempontjából a legfontosabb eltérés a ragozott igealakok számában figyelhető meg. A BUSZI-ban másfélszer annyi ragozott ige van (≈15%), mint az MNSz-ben (≈10%) (l. alább a 6. táblázatot). Ez az adat utal arra a később alaposan vizsgált tényre, hogy az írott nyelv több információt sűrít a főnévi csoportokba jelzős szerkezetek segítségével, míg a beszélt nyelv több alárendelt mondatot, és így több ragozott igét használ. Figyelembe véve, hogy tagmondatonként egy véges alakú igével számolhatunk,[7] megállapítható a tagmondatok

[7] Az egyes szám harmadik személyű, jelen idejű, kijelentő módú főnévi és melléknévi állítmányt tartalmazó tagmondatok ily módon kimaradnak a számolásból.

átlagos hossza. A BUSZI-ban 6,7, az MNSz-ben 10 szó adódik. Ezeket az értékeket összevetve a feljebb említett átlagos mondathosszal (BUSZI: 6,5; MNSz: 17,1) azt kapjuk, hogy a BUSZI mondatai jellemzően egy tagmondatból állnak, hiszen az átlagos mondat- és tagmondathossz gyakorlatilag azonos, míg az MNSz mondatai 1,7-szer hosszabbak, mint a tagmondatai, tehát a tipikus mondat két tagmondatból áll.

3.1.2. A bővítmények száma

Az NP-k számát a tagmondatok számához (azaz a ragozott igékhez) viszonyítva azt látjuk, hogy míg a BUSZI-ban kettőnél kevesebb NP jut egy tagmondatra, addig az MNSz-ben 3,5, vagyis az írott nyelv mondatai több bővítményt tartalmaznak (l. alább: 3.3.1. pont és 7. táblázat).

3.2. Szófajstatisztika

Szófaj	BUSZI				MNSz	
	tm		ak			
	%	∑	%	∑	%	∑
főnevek	12,5	11904	14,0	24345	29,0	87479
névmások	13,7	13013	12,9	22388	5,5	16667
számnevek	2,3	2200	3,7	6435	3,6	10861
egy-ek	1,0	917	1,4	2457	0,6	1930
névelők	6,4	6094	7,0	12058	12,3	37135
melléknevek	5,6	5315	5,6	9649	10,7	32149
határozószavak	20,5	19533	20,0	34722	7,6	22879
finit igék	15,2	14477	15,3	26570	9,9	29943
mn-i igenevek	0,5	512	0,5	946	3,3	9840
fn-i igenevek	1,7	1616	1,7	3010	1,0	3096
hat-i igenevek	0,2	146	0,2	331	0,3	896
kötőszavak	10,9	10393	12,2	21086	7,5	22651
névutók	0,7	634	0,9	1567	1,6	4778
indulatszavak	1,5	1461	0,4	644	0,0	116
egyéb	7,3	6919	4,1	7120	6,9	20881

6. táblázat. Szófajok

Már a legdurvább statisztikai elemzés, a különböző szófajú szavak számának összevetése is sokat elárul a beszélt nyelvi és az írott nyelvi korpusz mondat-

szerkezeti különbségeiről. A 6. táblázat a különböző szófajú szavak megoszlását mutatja a két korpuszban. Láthatjuk, hogy a legtöbb esetben az adatközlők és a terepmunkások szófajarányai közel azonosak még a diskurzusban betöltött eltérő szerepeik ellenére is, míg az írott nyelvi szófajmegoszlás ettől jelentősen eltér. Megjegyezzük, hogy az alább közölt statisztikai eltérések, ha külön nem jelezzük, 5%-os szignifikanciaszint mellett mindig szignifikánsak.

3.3. A főnévi csoport

Az alábbiakban a főnévi csoportok szerkezetével foglalkozunk részletesebben, ugyanis a közölni kívánt tartalom átadásának két véglete közül az egyik az, amikor a közölni kívánt információ számos tagmondatra tagolódik, míg a másik véglet a tömörített szöveg, amelyben az információ minél nagyobb részét egy mondatba kívánja foglalni a beszélő (vagy a szöveg írója), és ezért a tartalom jelentős része a mondaton belüli főnévi csoportokban jelzős szerkezetekbe sűrítve jelenik meg (vö. az (1a)-beli példában szereplő négy tagmondatot az (1b)-beli példa két tagmondatával).

(1) a. Hát, egy bizonyos összegért havonta, amit juttatnak, három évig ott dolgozok és akkor az az összeg az enyém lesz persze fizetés mellett ez egy ilyen külön társadalmi ösztöndíj.

b. Hát, egy külön társadalmi ösztöndíjnak számító bizonyos összegű havi juttatásért három évig ott dolgozok és akkor az az összeg az enyém lesz persze fizetés mellett.

A beszélt és az írott nyelvi korpusz főnévi csoportjainak összehasonlításakor fő hipotézisünk tehát az, hogy az írott nyelvben sokkal inkább megfigyelhető az információ főnévi csoportokba tömörítése, mint a beszélt nyelvben.

3.3.1. A főnévi csoport feje

A főnévi csoport feje főnév vagy névmás lehet és megfordítva: minden főnévre, illetve névmásra épül egy teljes főnévi csoport. A 7. táblázatban a főnévi csoportok számát a főnevek plusz névmások számával azonosítottuk, ami annyiban pontatlan, hogy a jelzőkkel bővített főnévi csoportból olykor el van hagyva a főnévi fej, illetve a mutató névmás nem mindig alkot önálló főnévi csoportot (pl. *ezt a kutyát*). Ezekről az esetekről alább még lesz szó. A főnévi csoportok jellemzően a mondat ragozott igéjének bővítményeiként jelennek meg a mondatban, de melléknévi csoportok (pl.: *büszke* ***a fiára***), más főnévi csoportok (pl.: *a fiú* ***a távcsővel***, ***a fiúnak*** *a távcsöve*) és igenevek (pl.: ***a kertben*** *játszó gyerek,* ***uszodában*** *úszni*) bővítményei is lehetnek.

Összességében több főnévi csoport van az MNSz-ben, mint a BUSZI-ban. A főnevek és névmások összesített aránya a teljes szószámhoz képest rendre 35%, illetve 26%. Ez az adat máris mutatja, hogy az írott nyelvi korpuszban nagyobb szerepe van a főnévi csoportoknak, mint a beszélt nyelvben, összhangban azzal a 3.1.1. pontban említett adattal, hogy a ragozott igék relatív száma viszont a beszélt nyelvben magasabb.

Fontos további jellemzője a beszélt nyelvi korpusznak, hogy a főnévi csoportok között sokkal nagyobb arányban vannak a névmások, mint az írott nyelvben. Míg az írott nyelvben a félreértés elkerülése végett érdemes egy teljes leírással egyértelműsíteni, hogy mire utalunk, addig a beszélt nyelv sokkal inkább támaszkodhat az egyértelműsítés nem nyelvi eszközeire is (pl. mutatás), és esetleges félreértés esetén lehetőség lenne visszakérdezni, így a figyelem középpontjában álló *(salient)* individuumokra elegendő csupán névmással utalni. A főnévi illetve névmási fejek aránya a BUSZI-ban közelítőleg 50–50%, míg az MNSz-ben 84–16% a főnevek javára.

	BUSZI		MNSz
A főnévi csoport feje	tm	ak	
főnevek aránya (%)	47,8	52,1	84,0
névmások aránya (%)	52,2	47,9	16,0
A főnévi csoportok száma			
a szószámhoz képest (%)	26,2	26,6	34,5
a finit igék számához képest (%)	1,7	1,8	3,5

7. táblázat. Főnévi csoportok

Az adatokhoz három pontosító megjegyzést kell fűznünk. Egyrészt meg kell jegyeznünk, hogy a jelzővel bővített NP főnévi feje olykor elmaradhat (pl. *a sárga tulipánból* helyett *a sárgából*), a nem alanyesetű melléknevek csak ilyen esetekben jelennek meg, ezért az esetragos melléknevek és főnevek számának összevetéséből látható, hogy milyen gyakran maradhat el a főnévi fej a főnévi csoportokból. A BUSZI-ban ez az arány 7,2%-nak adódik a terepmunkások és 6,4%-nak az adatközlők esetében, míg az MNSz-ben csupán 3,5%. Az ellipszisek valódi száma azonban ennél alacsonyabb, ugyanis bizonyos főnévként és melléknévként is érthető szavak melléknévként vannak megjelölve a korpuszban, és ezért például az *a törpéket* főnévi csoportban a *törpe* esetragos melléknévként számolódik. Az ebből fakadó hiba vélhetőleg egyformán érinti a BUSZI- és az MNSz-korpuszt, így ha a kapott értékek nem pontosak is, arányuk jól mutatja, hogy az MNSz NP-i

teljesebbek: nemcsak hogy ritkábban fejezhetők ki névmással, de a főnévi fej is kevésbé hagyható el belőlük.

Másrészt, mint említettük, bár az NP-k leggyakrabban a mondat ragozott igéjének bővítményei, ez nem feltétlenül van mindig így, ezért ezek az esetek torzítják az egy ragozott igére eső NP-k számára kapott értéket. Harmadrészt a mutató névmások *(ez, az)* összes előfordulásainak a BUSZI-ban mintegy 20%-a, az MNSz-ben 27%-a nem önálló NP-ként, hanem egy határozott főnévi csoporttal együtt fordul elő, ezeket tehát le kell vonnunk az önálló főnévi csoportként elszámolt névmások közül. Ez a kis korrekció azonban a névmások és főnevek arányára kapott értékeket lényegében nem módosítja.

3.3.2. Névelők

A főnevek számához viszonyítva több a névelő a BUSZI-ban (60%), mint az MNSz-ben (45%). Ennek a különbségnek két alapvető oka lehet. Egyrészt az MNSz-ben több a csupasz NP (pl. *gyermeket nevel*), másrészt az MNSz-ben komplexebb főnévi csoportok fordulnak elő, amelyekben egy névelőre több főnév is jut (pl. *a gyermeküket egyedül nevelő szülők*). Továbbá okozhat némi eltérést az is, hogy a beszélt nyelvben gyakori a személynevek névelős használata (pl. *a Tamás*).[8] A határozatlan névelők aránya az összes névelőhöz képest magasabb a BUSZI-ban, az adatközlők körében a legnagyobb, 17%, a terepmunkások esetében valamivel alacsonyabb, 13%, míg az MNSz-ben csupán 5%.

	BUSZI		MNSz
Névelők	tm	ak	
aránya a főnevekhez képest(%)	58,9	59,6	44,7
határozatlan névelők aránya az összes névelőhöz képest (%)	13,1	16,9	4,9

8. táblázat. Névelők

Míg a beszélt nyelvi korpuszban a határozatlan névelő részaránya nem mutat jelentős kvótánkénti eltérést (a terepmunkásoknál 11,1% és 16,4% között ingadozik, az adatközlőknél 14,6–20,5%), addig az MNSz-ben nagy különbségeket látunk a szövegek műfaja szerint. A hivatalos nyelvben tizedannyi határozatlan névelő for-

[8] A 3.3.4. pontban láthatjuk, hogy a datívuszos birtokos szerkezet, amely eggyel több névelőt tartalmaz, mint jelöletlen birtokost tartalmazó párja (pl. *Tamásnak a kocsija* és *Tamás kocsija*), gyakoribb a BUSZI-ban, ám ezen szerkezetek száma olyan alacsony, hogy nem tehető felelőssé az itt tárgyalt különbségért.

Határozatlan névelők aránya (%)	
szépirodalom	8,7
hivatalos stílus	0,6
sajtó	5,1
tudományos stílus	6,0
összes	4,94

9. táblázat. Határozatlan névelők az MNSz-ben

dul elő, mint a többi írott szövegben (l. 9. táblázat). (Fontos megjegyezni, hogy az *egy* szó számnévként is előfordulhat, ám a statisztikai vizsgálat nem teszi lehetővé e két használat szétválasztását.[9])

3.3.3. Jelzős szerkezetek

Feltevésünk szerint az írott nyelvi korpuszban több és egyben összetettebb jelzős szerkezetet találunk, mint a beszélt nyelvben. Ezt vizsgáljuk alább a névelőt is tartalmazó NP-ken a melléknévi, majd a melléknévi igeneves jelzők esetén.

Halmozott melléknévi jelzők

A BUSZI-ban a névelős főnévi csoportoknak kb. 58%-a bővítetlen, az MNSz-ben hasonló, de ennél valamivel alacsonyabb, 54% az arány. Az egy melléknévi jelzőt tartalmazók közel kétszer annyian vannak az MNSz-ben, mint a BUSZI-ban, a két melléknévvel bővítettek már 2,5-szer, a hárommal bővítettek négyszer annyian. Négy melléknévi jelzőt tartalmazó NP a BUSZI-ban már nem található (10. táblázat).

Melléknévi igenevek

A melléknévi igenevek használata sokkal gyakoribb az MNSz-ben, mint a BUSZI-ban, az adatokkal azonban óvatosnak kell lennünk, mert a melléknévi igenevek közül sok valójában már melléknévként lexikalizálódott (pl.: *elvált*), elkülönítésükre azonban az annotáció nem ad lehetőséget (11. táblázat).

A jelző + főnév szerkezetek között a melléknévi igenévi jelző a BUSZI-ban kb. 11%-ban, míg az MNSz-ben kétszer olyan gyakran, 22%-ban fordul elő (12. táblázat). Egy jellemző példát mutat a (2)-es szerkezet.

(2) az 1930-tól a németországi avantgárd művésziskolában pallérozódott, majd baloldali nézetei miatt a zwickenaui internálótáborba csukott és onnan Olaszországba szökött Tóth figurája

[9] Bizonyos szintaktikai és szemantikai elméletek nem is választják el ezt a két használatot (Szabolcsi 1994).

Halmozott mn.-i jelzők	BUSZI tm	BUSZI ak	MNSz
névelő(ne)+főnév(fn)	60,0	57,3	54,2
ne+mn+fn	8,60	8,77	17,07
ne+2mn+fn	0,90	0,86	2,36
ne+3mn+fn	0,06	0,06	0,23
ne+4mn+fn	0,00	0,00	0,01

10. táblázat. A bővítetlen és a melléknevekkel bővített névelős főnévi kifejezések százalékos aránya a névelők összes számához képest

Mn.-i igenevek	BUSZI tm		BUSZI ak		MNSz	
	%	∑	%	∑	%	∑
folyamatos	0,3	293	0,3	600	1,9	5806
befejezett	0,2	214	0,2	340	1,3	3922
beálló	0	5	0	6	0	112

11. táblázat. A melléknévi igenevek százalékos aránya a szavak számához viszonyítva

Igenévi/melléknévi jelzők	BUSZI tm	BUSZI ak	MNSz
melléknévi igenév+főnév	11,2	10,1	21,8
melléknév+főnév	88,8	89,9	78,2

12. táblázat. A melléknévi és melléknévi igenévi jelzők aránya

A melléknévi igenevek használata kézenfekvő módja az információ NP-n belüli tömörítésének, mivel az ige nemcsak magában, hanem bővítményeivel együtt is megjelenhet így jelzőként (l. (2)). A 13. táblázat adatai alátámasztják azt a feltételezést, hogy az írott nyelvi szöveg sokkal inkább él ezzel a lehetőséggel, ugyanis mintegy négyszer olyan gyakran van bővítménye az igenévi jelzőnek az MNSz-ben, mint a BUSZI-ban.

3.3.4. Birtokos szerkezet

A kétféle birtokos szerkezet – az alanyesetű és a *-nAk* ragos birtokost tartalmazó – megoszlása eltér a két korpuszban. A birtokot közvetlenül megelőző, nem

Bővített mn.-i ign. jelzők	BUSZI				MNSz	
	tm		ak			
	%	∑	%	∑	%	∑
bővítmény+mn.-i igenév+fn	11,5	25	10,5	43	41,3	2271

13. táblázat. A bővítményes melléknévi igenévi jelzők százalékos aránya a melléknévi igenévi jelzők között

Birtokos szerkezetek	BUSZI				MNSz	
	tm		ak			
	%	∑	%	∑	%	∑
alanyesetű birtokos	89,6	146	81,5	211	99,1	5101
részesesetű birtokos	10,4	17	18,5	48	0,9	44

14. táblázat. A birtokot közvetlenül megelőző birtokosok között a datívusz aránya

névmási birtokost tartalmazó szerkezeteket (pl. *a kutyának a szőre* vs. *a kutya szőre*) vizsgálva kitűnik, hogy a *-nAk* ragos (datívuszos) birtokos aránya a beszélt nyelvben lényegesen nagyobb. A BUSZI-adatközlők között több mint hússzor gyakoribb, mint az MNSz-ben (14. táblázat). Az írott nyelv tehát takarékosabb, az esetek 99%-ában elhagyja a datívuszi ragot és a birtok előtti határozott névelőt.[10] A kötetlen beszéd kevésbé takarékoskodik, lásd (3). Továbbá egy közel kétszeres különbség a terepmunkások és az adatközlők adatai között is mutatkozik.

(3) szoval mi az igazi párttitkárnak a szava

Többszörös birtokos szerkezetek a BUSZI-ban nem fordulnak elő, míg az MNSz-ben igen, a birtokos szerkezetek 6%-ában. Például:

(4) az energiahordozók világpiaci árának növekedéséből

3.4. Topikismétlő névmás

Már említettük, hogy a beszélt nyelv kevésbé takarékos a nyelvi elemekkel. Ezt igazolja a topikismétlő (rezumptív) névmások használata is.

[10] Vannak olyan szerkezetek, amikor a jelöletlen birtokos nem használható (pl. **ki kalapja* és **az kalapja* helyett csak a *kinek a kalapja* és *annak a kalapja* datívuszos alakok használhatók), ilyenkor természetesen az írott nyelv is csak a hosszabb formát használja.

(5) a. Hhh a *fiam azzz* vegyészmérnök
b. nekünk nyevészetből az tanítják, hogy [köhhint] *a nyelv az* nem romlik
c. szoval [P] *az egyetem az* valamikor márciusban indult volna be

A terepmunkások beszédében ötször gyakrabban követte az NP-ket topikismétlő névmás, mint az MNSz-ben, az adatközlőkre jellemző érték pedig még a terepmunkásokénak is majd háromszorosa (l. 15. táblázat).

Topikismétlő *az*	BUSZI				MNSz	
	tm		ak			
	%	∑	%	∑	%	∑
névelő+főnév+*az*	0,5	20	1,4	113	0,1	13

15. táblázat. A névelős főnevet követő topikismétlő névmás aránya

3.5. A *hogy*-os tagmondat névmási feje

A BUSZI-ban a *hogy* kötőszó az összes szó 3%-át teszi ki, míg az MNSz-ben csupán 1%-ot. Ez nagymértékben alátámasztja azt a feltevésünket, hogy a BUSZI lényegesen több alárendelt mondatot használ.

Ha egy ige valamelyik vonzata egy teljes propozíció, akkor ezt gyakran *hogy*-os alárendelt mondat fejezi ki. A *hogy*-os tagmondat sokszor névmás bővítményeként jelenik meg (azaz az ige vonzatát megtestesítő NP feje névmás, melynek a *hogy*-os tagmondat a bővítménye). A névmás lehet mutató névmás *(az*, pl.: *tudok arról, hogy …)*, vagy személyes névmás *(ő*, pl.: *tudok róla, hogy…)*. Az ige előtt (így pl. a fókuszált pozícióban is) csak az *arról* forma fordulhat elő (pl. *arról tudok, hogy …*, de **róla tudok, hogy …*), ige utáni helyzetben viszont mindkét névmás használható: *tudok arról, hogy …* és *tudok róla, hogy…* (Kenesei 1992), ezért a mutató névmást hangsúlyosnak érezzük. Sok beszélőnek az a benyomása, hogy nyelvhelyességi kérdésről van szó, és gondozott beszédben mindenképpen az *az* alakjai használandók (vö. Nádasdy 2002). Ennek az is oka lehet, hogy a személyes névmást alany- és tárgyesetben egyáltalán nem tesszük ki: *jó az, hogy jöttél*, *jó, hogy jöttél*, de **jó ő, hogy jöttél* illetve *tudja azt, hogy beteg*, *tudja, hogy beteg*, de **tudja őt, hogy beteg*, így a mutató névmás előfordulásai komoly túlsúlyban vannak.

hogy-os tagmondat névmási feje	BUSZI				MNSz	
	tm		ak			
	%	∑	%	∑	%	∑
mutató névmás+*hogy*	92,1	117	73,3	132	87,6	226
személyes névmás+*hogy*	7,9	10	26,7	48	12,4	32

16. táblázat. A mutató névmás és a személyes névmás aránya a névmás+*hogy*-os szerkezetekben

A 16. táblázat[11] mutatja, hogy a személyes névmás használata az adatközlőknél kétszer olyan magas, mint az írott korpuszban. A (6)-os két jellemző példát mutat a BUSZI-ból. A terepmunkások még az MNSz-beli értéknél is ritkábban használták ezt az alakot (bár az MNSz és a terepmunkások közti eltérésre $p = 0,05$, tehát 5%-os hibahatár mellett az eltérés nem szignifikáns).

(6) a. Kezünk-lábunk meg van kötve, a szülő ragaszkodik *hozzá, hogy* az ő gyereke érettségizzen

b. tisztába volt *vele, hogy* nálunk a legfontosabb tantárgy a matematika

Érdekes továbbá az, hogy a BUSZI-ban sokkal gyakoribb volt az alanyesetű *az, hogy* szerkezet használata, mint az MNSz-ben: az összes *az, hogy*-os szerkezet 42%-a, szemben a 18%-kal. Ennek vélhetőleg az az oka, hogy az alanyesetű propozicionális vonzat általában melléknévi állítmányok mellett jelenik meg (pl. *érdekes az, hogy*), és az írott nyelv ezekben az esetekben inkább igésített szerkezeteket használ (pl. *érdekesnek tartom azt, hogy*).

3.6. Vonatkozó mellékmondatok

A BUSZI-korpusz azt bizonyítja, hogy az *amely* és a *mely* vonatkozó névmás a beszélt nyelvből mára szinte teljesen eltűnt. Az adatközlők közül a *mely*-t senki, az *amely*-t csak a tanárok és az egyetemisták használták, így az adatközlők által használt összes vonatkozó névmásnak csak 0,7%-a volt *amely*, míg az MNSz referenciakorpuszban a *mely* és az *amely* együttesen 41%-ot tesz ki. Az *amelyik* viszont

[11] Az értékek az alany-, tárgy- és részesesetű alakok kivételével értendők. Az alany- és tárgyesetű alakoktól azért kell eltekinteni, mert mint említettük, a személyes névmást ekkor elhagyjuk, tehát ezek az alakok a korpuszban nem kereshetők. A részesesetű alakokat pedig azért hagytuk ki, mert sokszor a részesesetű vonzat után esik az a *hogy*-os tagmondat, ami valójában az alany vagy a tárgy (pl. *mondtam neki, hogy* . . .), tehát hamis találatokat kapunk.

az írott nyelvből hiányzik (0,4%), míg a BUSZI-ban több, mint 3%-ot képvisel. A beszélt nyelvben ugyanis az *amelyik* nem csak kiválasztó értelemben szerepel, hanem az *ami* és az *amely* helyett is, lásd (7). A vonatkozó névmások BUSZI-beli használatáról részletesen ír Szeredi (2008).

(7) s az Árpád Gimnáziumnak akkor még volt egy öööö nagyon jól működő cserkészcsapata, amelyik különböző rendezvényeket öööö gyártott, rendezett

Vonatkozó névmások	BUSZI				MNSZ	
	tm		ak			
	%	∑	%	∑	%	∑
aki/ami	92,4	970	96,1	1978	58,8	1799
amely	4,2	44	0,6	13	33,1	1011
mely	0,3	3	0,0	0	7,7	236
amelyik	3,1	33	3,3	67	0,4	12

17. táblázat. A vonatkozó névmások előfordulásai

A vonatkozó névmások összes száma megadja a vonatkozó mellékmondatok számát.[12] A vonatkozó mellékmondatok összes NP-hez számított aránya az MNSz-ben alacsonyabb, 2,9%, míg a BUSZI-ban a terepmunkások esetében 4,1%, az adatközlőknél 4,3%, vagyis a beszélt nyelv, feltevésünknek megfelelően, valóban gyakrabban fogalmazza a mondanivalót külön tagmondatba.[13]

A vonatkozó mellékmondatok közül csak a mondatkezdő pozícióban állókat vizsgálva szintén érdekes különbségek adódtak a két korpusz között. A vonatkozó mellékmondatok topikalizációval kerülnek a mondat élére. Ha a vonatkozó mellékmondatnak névmási feje van a mondatban, akkor ez a névmás mindig a mutató névmás (*az*). Az MNSz-ben azonban a vonatkozó mellékmondat névmási feje az esetek felében el van hagyva (pl. (8a)), míg a BUSZI-ban szinte mindig megjelenik (pl. (8b)).

(8) a. *Aki* erre jár és körül akar nézni, ~~*azt*~~ szívesen fogadjuk. (MNSz)

b. *Aki* hisz Istenbe, *az* hisz pap nékül is. (BUSZI)

[12] Elvileg nem kizárt, hogy több vonatkozó névmás álljon egy tagmondatban (pl. *Ki mint veti* …), de ilyen esetek ritkán fordulnak elő.

[13] Kis számban előfordulhatnak melléknévi, illetve határozói frázist bővítő vonatkozó mellékmondatok is, ezek ebben a statisztikában nem szerepelnek.

A 18. táblázat első két sora mutatja ezt az eredményt. A terepmunkások szövegében összesen 8 olyan mondat fordult elő, amelybe a mondat eleji vonatkozó mellékmondat után beilleszthető a mellékmondat *az* névmási feje, és ebből csupán egyszer maradt el az *az*, míg az adatközlők esetében 19 esetből egyszer sem. Ezzel szemben az MNSz-ben 43 esetből 21-ben el volt hagyva az *az*. A névmás hiányát tekinthetnénk az elhagyott hangsúlytalan személyes névmási fej eseténekk, ám ekkor az itteni eredmények ellentmondanának a *hogy*-os tagmondatok fejével kapcsolatban tapasztalt tendenciának, amely szerint az írott nyelv sokkal inkább a hangsúlyos *az* névmást használja, míg a beszélt nyelvben gyakrabban előfordul fejként a hangsúlytalan személyes névmás is, és ez utóbbi lenne az, ami (alany- és tárgyesetben) elhagyható. Az adatok helyes értelmezése az, hogy a vonatkozó mellékmondatot már önmagában, a névmási fej nélkül is referáló bővítményként tudjuk értelmezni, és ekkor a névmási fejre nincs szükség. A mondatkezdő vonatkozó mellékmondat után mégis gyakran és legfőképpen a beszélt nyelvben megjelenő *az* inkább topikismétlő névmásnak tekinthető, összhangban a topikismétlő névmásról elmondottakkal.[14]

Vmm kezdetű mondat	BUSZI				MNSz	
	tm		ak			
	%	∑	%	∑	%	∑
vmm + *az*	21,2	7	38,0	19	31,9	22
vmm + elhagyott *az*	3,0	1	0,0	0	30,4	21
visszautaló vmm	27,3	9	42,0	21	8,7	6
kettőspontos értelmezés	3	1	0	0	13	9
egyéb	45,5	15	20	10	15,9	11

18. táblázat. A mondatkezdő vonatkozó mellékmondatok (vmm) típusai

A mondatkezdő vonatkozó mellékmondatoknak a következő csoportja a visszautaló típus, lásd a 18. táblázat harmadik sorát. A mondatkezdő vonatkozó névmás ezekben az esetekben nem egy, a mondatban később következő főnévi csoportra vonatkozik, hanem az előző mondat valamely szereplőjére utal vissza, például (9). Ezek a vonatkozó mellékmondatok sokszor magukban állnak. Ezekben az

[14] Megjegyzendő, hogy ezzel a magyarázattal összecseng az a mai beszélt nyelvben igencsak elterjedt, de a BUSZI anyagában még nem adatolható szerkezet, melyben a főnév előtti vonatkozó mellékmondat tulajdonképpen a névelőt és a jelzőt egyszerre helyettesíti, de az esetek túlnyomó többségében még önállóan nem áll meg, csak topikismétlő névmással kiegészítve. Pl.: *Amit felolvastál vers, az ebben a kötetben is benne van.* Erről a szerkezetről ír Nádasdy (2006).

esetekben a vonatkozó névmás szintén referenciális kifejezésként viselkedik, tulajdonképpen személyes névmási funkcióban jelenik meg. A terepmunkások beszédében a mondatkezdő vonatkozó mellékmondatok 27%-a, az adatközlők beszédében ezek 42%-a utalt előző mondatbeli szereplőre, míg az MNSz-ben csupán 9%.

(9) Előtte Zuglóban laktunk a nagymamáméknál. *Aki* most ott lakik szintén egyedül.

Végül az MNSz-ben több példát is találunk (13%) a mondatkezdő vonatkozó mellékmondat kettőspontos értelmezésére (pl. (10)), míg a BUSZI-ban összesen egy ilyen mondat szerepelt. A kettőspontos értelmezés az *Ami …, az az, hogy …* mondat rövidített változatának tekinthető, ezt a tömörítést jellemzően az írott nyelv alkalmazza.

(10) a. *Ami* tényként leszögezhető: 1612. november 12-én Dersffy Orsolya — meg vele a Mágóchy-vagyon — feleségül ment az akkor harmincesztendős Esterházy Miklóshoz.

b. *Ami* még ennél is fontosabb: a televíziók nem a háború valóságos emberi vonatkozásaira voltak kíváncsiak.

4. Összefoglalás és további feladatok

A tanulmányban az írott és beszélt nyelvhasználat néhány jellemző különbségét illusztráltuk nyelvtechnológiai módszerekkel segített lexikai és szintaktikai elemzés alapján. Viszonylag egyszerű eszközökkel kaptunk nem triviális eredményeket, melyek alapul szolgálhatnak további nyelvi elemzéseknek. Megmutattuk, hogy az írásbeli nyelvhasználat lexikai gazdagsága, a közölt információ tömörítésére való törekvése a szóbeli változattal szemben egyértelműen, kvantitatív mérésekkel igazolható.

A beszélt nyelvi korpusz méretének és a mondatelemzés mélységének a növelésével részletesebb vizsgálatok is elvégezhetők, ilyenek a BUSZI kvóták közötti különbségek nyelvstatisztikai elemzése, vagy a szórendre, összetettebb szintaktikai szerkezetekre vonatkozó elemzések. Meg kell jegyeznünk, hogy a számítógéppel magasabb nyelvi szinteken végzett elemzések (például szintaktikai elemzés) elkerülhetetlen hibái még inkább jelentkeznek, így lényegesen befolyásolhatják a vizsgálatok eredményeit, ezért vagy kézzel ellenőrzött adatokra, vagy jelentősen nagyobb mennyiségű adatra lenne szükség, de ez a beszélt nyelvváltozat esetében jelenleg nem áll rendelkezésre. Mindazonáltal már az itt használt mutatók is felhasználhatók további részletes vizsgálatokhoz, például az egyes szövegfajták típustelítettségének meghatározásához.

Irodalom

Clarkson, Philip R. – Roni Rosenfeld 1997. Statistical language modeling using the CMU-Cambridge toolkit. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH-97), volume 1, 2707–2710.

Halliday, Michael Alexander Kirkwood 1985. An introduction to functional grammar. London: Edward Arnold.

Jelinek, Frederick – Robert L. Mercer – Lalit R. Bahl – James K. Baker 1977. Perplexity – A measure of the difficulty of speech recognition tasks. Journal of the Acoustical Society of America 62: S63.

Johansson, Victoria 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. Lund University Working Papers 53: 61–79.

Johnson, Wendell 1944. Studies in language behavior. I. A program of research. Psychological Monographs 56: 1–15.

Kenesei István 1992. Az alárendelt mondatok szerkezete. In: Kiefer Ferenc (szerk.): Strukturális magyar nyelvtan 1. Mondattan. Budapest: Akadémiai Kiadó. 79–176.

Kilgarriff, Adam 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In: Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition. Sussex, UK.

Kilgarriff, Adam 2001. Comparing corpora. International Journal of Corpus Linguistics 6: 1–37.

Kontra Miklós 1990. Budapesti élőnyelvi kutatások. Magyar Tudomány 5: 512–520.

Labov, William 1984. Field methods of the project in linguistic change and variation. In: John Baugh – Joel Sherzer (szerk.): Language in use. Englewood Cliffs: Prentice-Hall. 28–53.

Lanstyák István 2009. A magyar beszélt nyelv sajátosságai. Pozsony/Bratislava: Stimul.

MacWhinney, Brian 2000. The CHILDES Project: Tools for analyzing talk (Third edition). Mahwah, NJ: Lawrence Erlbaum.

Malvern, David D. – Brian J. Richards – Ngoni Chipere – Pilar Durn 2004. Lexical diversity and language development: Quantification and assessment. Basingstoke: Palgrave Macmillan.

McCarthy, Philip M. – Scott Jarvis 2007. vocd: A theoretical and empirical evaluation. Language Testing 24: 459–488.

McCarthy, Philip M. – Scott Jarvis 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods 42: 381–392.

McKee, Gerard – David Malvern – Brian Richards 2000. Measuring vocabulary diversity using dedicated software. Literary and Linguistic Computing 15: 323–337.

Nádasdy Ádám 2002. A rágondolás tárgya. Magyar Narancs 04/18: 40.

Nádasdy Ádám 2006. Előre, mellékmondat! Magyar Narancs 07/13: 40–41.

Németh T. Enikő 1991. A megnyilatkozás-típus elméleti kérdései és a szóbeli diskurzusok megnyilatkozás-példányokra tagolása. Kandidátusi értekezés. Szeged.

Rayson, Paul – Roger Garside 2000. Comparing corpora using frequency profiling. In: Proceedings of the workshop on comparing corpora (ACL 2000), 1–6. Association for Computational Linguistics.

Richards, Brian J. – David Malvern 1997. Quantifying lexical diversity in the study of language development. University of Reading, Faculty of Education and Community Studies.

Stamatatos, Efstathios – Nikos Fakotakis – George Kokkinakis 2000. Automatic text categorization in terms of genre and author. Computational Linguistics 26: 471–495.

Szabolcsi, Anna 1994. The noun phrase. In: Katalin É. Kiss – Ferenc Kiefer (szerk.): The syntactic structure of Hungarian (Syntax and semantics 27). New York: Academic Press. 179–274.

Szeredi Dániel 2008. Vonatkozó névmások használata beszélt nyelvi korpusz alapján. Szakdolgozat. ELTE, Budapest.

Tweedie, Fiona J. – R. Harald Baayen 1998. How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities 32: 323–352.

Ure, Jean N. 1971. Lexical density and register differentiation. In: George. E. Perren – John L. M. Trim (szerk.): Applications of linguistics: Selected papers of the 2nd International Congress of Linguistics, Cambridge 1969. Cambridge: Cambridge University Press. 443–452.

Váradi, Tamás 2002. The Hungarian National Corpus. In: Mark T. Maybury (szerk.): Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas: European Language Resource Association. 385–389.

Váradi Tamás 2003. A Budapesti Szociolingvisztikai Interjú. In: Kiefer Ferenc – Siptár Péter (szerk.): A magyar nyelv kézikönyve. Budapest: Akadémiai Kiadó. 339–359.

Váradi Tamás – Peredy Márta – Oravecz Csaba 2010. Nyelvtechnológiai módszerek a Budapesti Szociolingvisztikai Interjú lexikai és szintaktikai vizsgálatában. In: Tanács Attila – Vincze Veronika (szerk.): A VII. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 300–313.

Lexical and syntactic properties of the Budapest Sociolinguistic Interview

Abstract: The paper investigates the lexical and syntactic properties of the conversational modules of the Budapest Sociolinguistic Interview using basic language technology methods and tries to capture the differences between spoken and written language use with quantitative measures. The analysis compares the spoken language corpus with automatically annotated samples from the Hungarian National Corpus of written text and, whenever possible, also examines the lexical and structural differences in the language use of the various socio-economic groups in the interview. This study is not comprehensive but the lexical richness of written texts and the effort to convey the message in a more compact form as compared to the spoken language can be clearly shown with quantitative measures.

Keywords: spoken and written language use, corpus comparision, lexical richness, frequency profiling, syntactic properties of spoken langauge

A fogalmi metaforák és a szövegstatisztika szerepe a metaforák felismerésében*

Babarczy Anna[1] – Simon Eszter[2]

[1]*BME, Kognitív Tudományi Tanszék, Budapest*
[2]*Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Budapest*
babarczy@cogsci.bme.hu; simon.eszter@nytud.mta.hu

A tanulmány a metaforikus kifejezések automatikus gépi felismerését vizsgálja. Az emberi metaforaértelmezés két elméleti modelljét, a fogalmimetafora-elméletet és a statisztikai megközelítést vetjük össze. A két elmélet alapján pszicholingvisztikai és korpusznyelvészeti módszerek felhasználásával potenciálisan metaforikusságra utaló nyelvi jelek listáit állítottuk össze, majd a listák valós metaforajósló erejét számítógépes modellel teszteltük. Az eredmények szerint a statisztikai módszer bizonyult a legsikeresebbnek, bár ennek a teljesítménye is elmarad a várakozásoktól. A gyenge teljesítmény okait a metafora jelenségének megfoghatatlanságában, a fogalom meghatározásának elméleti pontatlanságában keressük.

Kulcsszavak: metafora, korpuszelemzés, gépi nyelvfeldolgozás, kognitív nyelvészet, szövegstatisztika

1. A metafora elméleti megközelítései

1.1. Bevezetés

A nem szó szerinti jelentések felismerése és értelmezése mind a pszicholingvisztika, mind pedig a számítógépes nyelvészet egyik megoldatlan problémája. A kérdés arra vonatkozik, hogy milyen ismeretek segítségével tudjuk megállapítani egy-egy kifejezésről, hogy az adott beszédhelyzetben, illetve szövegkörnyezetben nem a szó szerinti jelentés a releváns értelmezés, hanem annak valamiféle metaforikus kiterjesztése. Metaforikus kiterjesztésen itt olyan, többnyire elvont jelentéseket értünk, amelyek egy-egy szóhoz vagy jól körülhatárolható kifejezéshez kötődnek, de nem felelnek meg a kifejezés konkrét szótári alapjelentése(i)nek. Tehát nem többértelmű kifejezések egyértelműsítéséről van szó, hanem olyan jelentések felismeréséről, amelyeket az értelmezés rugalmassága miatt nem lehet szótárszerűen felsorolni.

* A dolgozatot az EU FP6 keretprogram (028714 sz. ösztöndíj) és az MTA Bolyai ösztöndíja támogatta.

Ezen a kérdéskörön belül a tanulmány célja annak vizsgálata, hogy hogyan határozhatjuk meg nyelvi jelek azon listáit, amelyek bármely természetes nyelvi szövegben metaforikus jelentésekre utalnak, és a listák felhasználásával milyen kereső algoritmusok azonosítják legsikeresebben a metaforákat. A kutatás eredményei két irányban is értelmezhetők: egyrészt a metaforák emberi feldolgozásának elméleteire vetnek fényt, másrészt pedig a számítógépes nyelvfeldolgozás módszereit gazdagítják.

A jelen tanulmány a VII. Magyar Számítógépes Nyelvészeti Konferencia kiadványában megjelent értekezés bővített változata (l. Babarczy et al. 2010). Elsőként az absztrakt fogalmak és azon belül a metaforikus kifejezések értelmezésének elméleti modelljeit tárgyaljuk. Ezt követően bemutatjuk a metaforák korpusznyelvészeti kutatásának eddigi eredményeit és megoldatlan kérdéseit. A harmadik részben részletezzük a jelen kutatás módszereit és eredményeit, és megvitatjuk a hiányosságok feltételezett magyarázatait. Végül összegezzük a tanulmány tanulságait.

1.2. A metafora meghatározása és szerepe

A metaforamegértés kérdése két elméleti kérdéskör metszetében helyezkedik el. Az egyik a fent említett, nem szó szerinti jelentések megértésének problémája. A másik arra vonatkozik, hogy hogyan tud az ember olyan tudásra szert tenni, amelyről nem lehet közvetlen, perceptuális tapasztalata. Az ilyen tudást nevezhetjük absztrakt tudásnak, vagy pontosabban elvont fogalmak megismerésének. A metafora a két kérdéskör egyik találkozási pontja, mivel (a) a metaforikus nyelvhasználat definíció szerint nem szó szerinti nyelvhasználat, (b) a metaforikus kifejezések gyakran a fenti értelemben elvont fogalmakat kódolnak, és (c) a fogalmimetafora-elmélet testesültségi hipotézise szerint a metaforák adják a magyarázatot a nem szó szerinti jelentésben megnyilvánuló elvont fogalmak megértésére. Az elvont fogalmak megismerésére létezik azonban egy másik elméleti magyarázat, a statisztikai tanulás hipotézise, amely szerint a nyelv statisztikai tulajdonságai segítségével sajátítjuk el elvont fogalmainkat. A következőkben ezt a két megközelítést tekintjük át.

A testesültségi hipotézis a kognitív nyelvészet egyik alaptétele (Gibbs 2006; Kövecses 2002; 2005; Lakoff–Johnson 1980; 1999; Gentner et al. 2001; Murphy 1996). A hipotézis szerint absztrakt fogalmainkat a konkrét fogalmakról szerzett perceptuális vagy testi tapasztalatok révén sajátítjuk el, illetve tudjuk értelmezni. Ez lenne a magyarázat arra az empirikus megfigyelésre, hogy az emberi nyelvek konkrét jelentésű szavakat vagy kifejezéseket gyakran metaforikus, elvont érte-

lemben használnak. A fogalmimetafora-elmélet szerint egy-egy ilyen kifejezés konkrét szó szerinti értelme a metafora forrástartománya, míg az átvitt, gyakran elvont metaforikus jelentés a céltartomány. Például az egyik ilyen cél- és forrástartomány pár az AZ IDŐ TÉR metafora: az idő absztrakt fogalmáról konkrétabb téri fogalmak segítségével beszélünk (pl. *előtt, után, alatt)*, amelyeket tapasztalati úton, a környezettel való interakció során sajátítunk el. Éppen emiatt az emberi nyelvek nagyrészt közös morfématárral kódolják a téri és az idői elrendeződést (pl. *ház**ban** – június**ban**, ház**ig** – június**ig***). A kognitív nyelvészet számos hasonló fogalmi metaforát azonosított nyelvi elemzés útján.

Az absztrakt tudás mibenlétének másik magyarázata, a statisztikai tanulás elmélete szerint a nyelv statisztikai tulajdonságai, azaz a nyelvtani mintázatok és a lexikális együttelőfordulások gyakoriságai segítségével sajátítjuk el és strukturáljuk absztrakt fogalmainkat (Burgess–Lund 1997; Landauer–Dumais 1997). Ebben az elméleti keretben a metafora nem az elvont tudás strukturálásának szükségszerű eszköze, hanem csupán a nem szó szerinti vagy esetleg homályos, alulspecifikált jelentések egyik válfaja.

A statisztikai megközelítés (Goatly 2002; Kintsch 2000) a metaforák forrás- és céltartománya közötti szemantikai távolság alapján ragadja meg a metafora jelenségét. A szemantikai távolság mérésére használatos a **látens szemantikai analízis** módszere (LSA, l. Landauer–Dumais 1997), ahol egy-egy szó szemantikai terét a környezetében előforduló szavak definiálják. Nagyon leegyszerűsítve, a szemantikai távolságot a szemantikai terek átfedésének mértékével mérhetjük. A modell szerint erősen metaforikusnak tekinthetők azok a kifejezések, melyekben a forrás- és céltartomány közti szemantikai távolság relatíve nagy: pl. *Az ügyvédem egy cápa* mondat esetén az alany *(ügyvéd)* és a predikátum *(cápa)* közti szemantikai távolság jelentősen nagyobb, mint egy szó szerinti mondat esetén, mint pl. *Az alma gyümölcs*. Szembetűnő azonban, hogy a modell nem húz éles választóvonalat a szó szerinti és a metaforikus értelmezés között, és a metaforát a lexikális többértelműség speciális típusának tekinti.

A statisztikai és a fogalmi metaforákra épülő megközelítés elméletileg megfér egymás mellett, hiszen elképzelhető, hogy absztrakt tudásunk mindkét forrást felhasználva alakul ki. Ezt az integratív álláspontot képviselik Andrews és munkatársai (2005; 2007), akik probabilisztikus generatív modelleket használnak a szemantikai reprezentációk kiépülésének szimulálására: egy szó attribúciós (a szóval asszociált nem-nyelvi fizikai tulajdonságok) és disztribúciós (a szó más nyelvi elemekkel való együttes előfordulásai) tulajdonságai egyaránt szerepet játszanak a jelentés lehorgonyzásában. A konkrét fogalmak esetében az attribúciós tulajdonságok bizonyulnak jobb támpontoknak, az absztrakt fogalmak esetében pedig a disztribúciósak.

A testesültségi és a statisztikai elmélet közti különbséget tulajdonképpen arra a kérdésre vezethetjük vissza, hogy lehetséges-e szimbólumlehorgonyzás (jelentés kialakulása) **kizárólag** nyelvi szimbólumokra épülve. A metaforamegértésre nézve pedig az a lényegi kérdés, hogy függetlenek-e az absztrakt fogalmak a konkrét fogalmaktól a nyelvhasználat során. A két szemlélet a különféle kognitív rendszerek és modalitások szerepét és súlyát vitatja az absztrakt tudás reprezentációjának kérdése kapcsán.

2. A metaforák feldolgozása

2.1. A kísérleti eredmények ellentmondásai

A metaforaelméletek tehát számos olyan kérdést vetnek fel, amelyekre a főként intuíción alapuló elméleti modellek már nem szolgálhatnak kielégítő válaszokkal. Ezek közül talán a három legfontosabbat a következőképpen fogalmazhatjuk meg:

- Hogyan dolgozzuk fel a metaforikus kifejezéseket?
- Valóban konceptuális leképezések segítenek hozzá a metaforák megértéséhez?
- A természetes nyelv szövegeinek vizsgálata alapján mit mondhatunk el a fogalmimetafora-elméletről?

Míg az első két problémakört pszicholingvisztikai kísérletek segítségével vizsgálhatjuk, a harmadik kérdés megválaszolásához korpusznyelvészeti módszerek alkalmazására van szükség. Mint látni fogjuk, a metaforikus nyelvhasználat korpuszalapú vizsgálata nemcsak a kognitív elmélet hiányosságaira mutat rá, hanem arra is rávilágít, hogy sokszor a kísérleti megközelítések sem problémamentesek.

A metaforikus kifejezések feldolgozását tekintve két fő elmélettel találkozhatunk. A **standard pragmatikai nézet** *(standard pragmatic view)* szerint – amely alátámasztani látszik a grice-i és neo-grice-i elméleteket (Levinson 2000) – a figuratív nyelv feldolgozása során az átvitt (ez esetben metaforikus) jelentéshez való hozzáférés előtt mindig a szó szerinti jelentés aktiválódik először. Ezzel szemben a **közvetlen hozzáférés elmélete** *(direct access view)* azt feltételezi, hogy a metaforikus kifejezéseket – szó szerinti megfelelőikhez hasonlóan – azonnal elérjük, mindenféle feldolgozásbeli költség nélkül, ami főleg a kontextuális hatásoknak tudható be (pl. Gibbs 1994). Ezeket az eredményeket és elméleti érveket

felhasználva a relevanciaelmélet a metaforákat a homályos kifejezések kategóriájába sorolja, amelyeknek a jelentését az adott beszédhelyzetben a relevancia elveinek megfelelően hívjuk elő (Sperber–Wilson 1986/1995; Wilson–Sperber 2004). Más kutatók ugyanakkor úgy találták, hogy még a viszonylag erős kontextusok sem gátolják meg az irreleváns jelentések aktiválódását (Peleg et al. 2001), ami azt jelenti, hogy a kontextuson kívül egyéb tényezők is közrejátszhatnak a metaforák feldolgozásában.

A **fokozatos kiugróság modellje** *(graded salience model)* (Giora 1997; 2008) éppen ezt állítja: e szerint a nézet szerint a figuratív nyelv megértésekor két, egymással párhuzamosan futó mechanizmussal kell számolnunk, amelyek közül az egyik alulról felfelé ható, ingervezérelt feldolgozás, amely csak a nyelvi tényezőkre érzékeny, a másik pedig fentről lefelé haladó folyamat, amely a nyelvi kontextust és a nyelven kívüli tudást hívja segítségül. Az alulról felfelé ható mechanizmus kiugróságérzékeny: a mentális lexikonban a konvencionalitás, gyakoriság, ismerősség, prototipikusság szerint kódolt nyelvi egységek közül a kiugróbbak gyorsabban hozzáférhetők lesznek a feldolgozás számára. Ez természetesen nem jelenti azt, hogy a kevésbé kiugró jelentéseket soha nem érjük el hamarabb, hiszen egy erősen prediktív kontextus felgyorsíthatja annyira a hozzáférést, hogy ez aktiválódjék először. Ez azonban nem blokkolja a kiugró választ, csupán visszatartja.

Gibbs és Matlock (2008) ugyancsak a szó szerinti jelentés blokkolásának szükségessége ellen érvel, viszont Giora elméletével ellentétben nem nyelvi tényezőkkel, hanem a testesültség hipotézisével, pontosabban a metaforikus szimuláció *(metaphorical simulation)* segítségével magyarázza a jelenséget. Eszerint amikor metaforikus kifejezéseket értelmezünk, mindig egyfajta fizikai mozgásszimulálást végzünk, vagyis elképzeljük a metaforikusan használt szó által leírt cselekvést vagy eseményt. Kísérleti bizonyítékaink vannak arra, hogy például a szenzomotoros tapasztalat nagyban befolyásolja az időről szóló metaforikus kifejezések értelmezését (Boroditsky–Ramscar 2002), akárcsak a párkapcsolatokat utazásként leíró szövegek megértését (Gibbs 2007, 171–173). Ebben az értelemben tehát a szó szerinti jelentés részvétele a feldolgozásban nem olyan nyelvi tényezőkre vezethető vissza, mint a gyakoriság, konvencionalitás vagy ismerősség, hanem inkább arról van szó, hogy konkrét testi szimulációkat alkalmazunk a figuratív nyelv feldolgozásakor.

Ezek az eredmények azt a feltételezést támasztják alá, hogy a metaforák megértésekor valóban az absztrakt céltartománynak a konkrét forrástartományra való leképezése történik meg. Ugyancsak ezt a feltételezést erősítik meg egyes lexikális döntési feladatok, amelyek során azt figyelték meg, hogy az A DÜH EGY TARTÁLYBAN LÉVŐ FELHEVÍTETT FOLYADÉK (ANGER IS A HEATED FLUID IN A CON-

TAINER) vagy AZ OPTIMIZMUS FÉNY (OPTIMISM IS LIGHT) fogalmi metaforákat tartalmazó kifejezések után a kísérleti alanyok gyorsabban döntöttek arról, hogy a *hőség*, illetve a *fény* lexémák valódi szavak-e, mint az ilyen metaforákat nem tartalmazó kifejezések után (Gibbs 2007).

Akadnak azonban olyan vizsgálatok is, amelyek az eddig elmondottakkal ellentétes következtetésre jutottak. Keysar és munkatársai (Keysar et al. 2000) olyan szövegek megértésének feldolgozási idejét mérték, amelyek az A SZERELEM EGY BETEG PÁCIENS (LOVE IS A PATIENT), A VITA UTAZÁS (ARGUMENT IS JOURNEY) és A GONDOLATOK EMBEREK (IDEAS ARE PEOPLE) fogalmi metaforákat tartalmazták. A kísérlet olyan újszerű metaforikus kifejezések megértését tesztelte, amelyeket vagy a már említett metaforatípus konvencionális példái előztek meg, vagy pedig nem metaforikus mondatok. Az eredmények szerint az első esetben a megértés nem volt gyorsabb, mint a másodikban, ami arra utal, hogy a konvencionális kifejezések feldolgozása során nem aktiválódtak a megfelelő konceptuális leképezések. AZ AZ IDŐ TÉR (TIME IS SPACE) fogalmi metafora vizsgálatát célzó kísérletek eredményei szintén nem adnak egyértelmű választ arra a kérdésre, hogy a téri sémák feltétlenül szükségesek-e az időről való gondolkodáshoz (Szamarasz 2006).

2.2. Korpuszelemzések eredményei

A korpusznyelvészeti módszereket segítségül hívó kutatók az elméleti megközelítések sokféleségének és a pszicholingvisztikai kísérletek ellentmondásos eredményeinek problémáját általában abban látják, hogy ezek egyrészt túlságosan a metaforák fogalmi természetével vannak elfoglalva, és így figyelmen kívül hagyják a nyelvi tényezőket, másrészt nem korpuszadatokat használnak a kísérletek lebonyolításához, hanem nyelvi intuíción alapuló kitalált példákat, amelyek gyakran félrevezetőek lehetnek. A kognitív metaforaelmélet szerint például a dühről azért beszélünk a „hőség" vagy a „nyomás" forrástartományok segítségével, mert ebben az állapotban testhőmérsékletünk megemelkedik, arcunkba vér szökik, és így ez a valós fiziológiai folyamat befolyásolja a „düh" absztrakt céltartományról való beszédünket. A Deignan (2008) tanulmányában összefoglalt korpusz- és szövegnyelvészeti vizsgálatok azonban azt mutatják, hogy nem annyira az egyének érzéseit fejezzük ki a „hőség" és „nyomás" fogalmakkal, hanem inkább egyfajta kollektív dühöt. Erre az utal, hogy az adatok alapján a céltartomány a legtöbb esetben egy tömeg vagy embercsoport, és a *heves*, *tüzes* kifejezések inkább a viták, megbeszélések metaforikus jellemzői, semmint az emberek érzelmeinek kifejezői. Ilyen értelemben az A DÜH EGY TARTÁLYBAN LÉVŐ FELHEVÍTETT FO-

LYADÉK NYOMÁSA (ANGER IS THE PRESSURE OF HEATED FLUID IN A CONTAINER) konceptuális leképezés helyett sok esetben inkább az A DÜHÖS TÖMEG FUTÓTŰZ (AN ANGRY GROUP OF PEOPLE IS A WILDFIRE) metaforával van dolgunk, amelyben a „futótűz" azért szerepelhet forrástartományként, mert a tűzről tudjuk, hogy irányíthatatlanná válhat, pusztító hatású, egy kisebb mennyiség is kiválthatja az elterjedését stb., és ezek mind jellemzőek lehetnek egy lázadó tömegre.

Stefanowitsch (2006) az érzelmekkel kapcsolatos metaforák korpuszalapú elemzése során azt találta, hogy az ún. „metaforikus sablon" módszer *(metaphorical pattern method)*, amely a metaforák céltartományára jellemző szavak korpuszokban való vizsgálatát jelenti, jóval hasznosabb lehet az elméleti kutatók által használt introspekciónál – két okból is: az egyik, hogy ezzel a módszerrel olyan metaforatípusokat is fel lehet lelni, amelyekről eddig nem esett szó a szakirodalomban, a másik pedig, hogy a gyakorisági mutatókat figyelembe véve meg lehet határozni, hogy az egyes céltartományokat mely leképezések jellemzik leginkább. A fogalmimetafora-elmélet szerint például a „boldogság" céltartományt a következő forrástartományok strukturálják: „fent", „fény", „melegség", „természeti erő" stb. Stefanowitsch az általa használt módszerrel további forrástartománytípusokat határozott meg, amelyek szintén a „boldogság" absztrakt kategóriát írják le: „folyadék", „összetörhető tárgy", „betegség", „agresszív állati viselkedés", „organizmus" stb.

Deignan (2005; 2008) főként a metaforikus kifejezésekben szereplő szavak grammatikai és kollokációs természetét vizsgálva arra mutatott rá, hogy a pszicholingvisztikai kísérletekben használt példák problémákhoz vezethetnek. Keysar et al. (2000) kísérleteinek azt a példáját elemezve, amelyben a *latest child* 'legutóbbi gyerek' angol kifejezés metaforikusan a történet szereplőjének legutóbbi szellemi munkájára referált, a szerző arra az eredményre jutott, hogy az általa vizsgált korpuszban a *latest* + *child* kollokáció szinte soha nem fordul elő, ugyanis ilyen esetben a *child*nak a *youngest* 'legkisebb' szokott lenni a megfelelő jelzője, nem pedig a *latest*. A szerző szerint ez megértésbeli nehézséget okozhatott a kísérleti alanyok számára. A kollokációk márpedig fontosak, hiszen informálhatják a nyelvhasználót arról, hogy metaforikus kifejezéssel van-e éppen dolguk, vagy szó szerinti jelentéssel. Vagyis nemcsak az önálló szavak jelentéséről van tudásunk, hanem két vagy több, egymás környezetében gyakran előforduló szót is sokszor egyetlen egységként kezelünk, és úgy is dolgozunk fel.

A nyelvi metaforák grammatikai viselkedésének vizsgálata is olyan fontos részletekre világít rá, amelyeket a fogalmimetafora-elméletben figyelmen kívül hagynak. Ugyancsak Deignan (2005) elemzéseiből derül ki, hogy a különböző szavak, kifejezések többnyire más-más grammatikai jellemzőkkel, illetve logikai relációkkal rendelkeznek a szó szerinti és a metaforikus használatban. AZ AZ

EMBERI VISELKEDÉS ÁLLATI VISELKEDÉS fogalmi metafora esetén például azok a szavak, amelyek a forrástartományban szerepelnek és entitásokat jelölnek, metaforikus használatukban többnyire igeként vagy melléknévként fordulnak elő. A szerző egyéb metaforatípusok vizsgálata alapján számos példával mutatja meg, hogy metaforikus használatban a szavak jóval kevesebb grammatikai szabadsággal rendelkeznek, mint amikor szó szerinti jelentésükben jelennek meg. Ez azt jelenti, hogy a forrástartományban lévő entitások közti logikai reláció nem egyszerűen megismétlődik a céltartományban, ahogyan a kognitív metaforaelmélet jósolná, hanem át is alakul: a szavak metaforikus jelentésükben önálló életet kezdenek élni. Ez pedig a konceptuális leképezés elmélete helyett inkább az ún. elegyítéselméletet *(blending theory),* azaz a fogalmi integráció elméletét támasztja alá, amely azt mondja ki, hogy a metaforikus nyelvhasználat során a forrás- és céltartományok felhasználásával egy harmadik, kevert tartományt hozunk létre, amely saját struktúrával és relációkkal rendelkezik, s ezeket sajátos nyelvi jegyekkel fejezzük ki (Fauconnier–Turner 2002; Kövecses 2005).

Természetesen olyan elemzések is léteznek, amelyek alátámasztják a kísérletek során kapott eredményeket. Martin (2006) a metaforákat megelőző kontextusokat vizsgálva úgy találta, hogy azok a kontextusok jósolják meg leginkább a célmetaforát, amelyek ugyanolyan típusú metaforikus kifejezéseket tartalmaznak, a legkevésbé pedig azok, amelyekben a forrástartomány szavai szó szerinti jelentésükben fordulnak elő. A szerző szerint ez az eredmény azt a korábbi kísérletet erősíti meg, amelyben úgy találták, hogy a metaforikus kontextus felgyorsítja, a forrástartomány szavainak szó szerinti jelentésben való használata pedig gátolja a célmetafora megértését.

A fentebb bemutatott korpusznyelvészeti elemzések alapján a fogalmimetafora-elméletet és a pszicholingvisztikai kísérleteket érő egyik legfontosabb bírálat abban áll, hogy nem fektetnek elég hangsúlyt a metaforikus nyelvhasználat nyelvi jellemzőire. Ezek az adatok azonban – mint kiderült – igen fontosak, hiszen rámutatnak, hogy olyan egyéb tényezők is szerepet játszanak a figuratív nyelvhasználatban, mint a gyakoriság, a kollokáció, a nyelvi sablonok, a grammatikai formák, továbbá a nyelvi és szövegtípusbeli változatosságok. A mentális leképezés elmélete tehát önmagában nem tudja megmagyarázni a nyelvben fellelt mintákat.

3. A korpuszépítés

3.1. A korpuszelemzés módszertani kérdései

A metaforák korpuszalapú vizsgálata ugyanakkor korántsem olyan egyszerű, mint első látásra tűnik. Egyrészt a korpusz kiválasztása önmagában is meghatározó jelentőségű lehet, másrészt pedig a metaforikus kifejezések szövegekben való azonosítása sem problémamentes. Ez utóbbi azért okoz nehézséget, mert a kognitív szakirodalomban tárgyalt konceptuális leképezések általában nincsenek sajátos nyelvi formákhoz kötve, s így nem könnyű meghatározni azokat a nyelvi jegyeket, amelyek leginkább jellemezhetik az egyes tartományokat.

Az egyik lehetséges módszer így a kézi keresés, amelynek során a kutatók saját nyelvi intuíciójukra támaszkodva próbálják összegyűjteni egy adott korpuszból a szerintük metaforikusnak ítélt kifejezéseket. Mivel ez az eljárás meglehetősen idő- és munkaigényes, legalább részben automatizált módszerekkel is érdemes próbálkozni. Ilyen módszer a forrástartomány szókincsére való rákeresés (pl. Deignan elemzései). Ebben az esetben összegyűjtjük az adott metaforatípus forrástartományára potenciálisan jellemző szavakat, majd megnézzük, hogy milyen arányban fordulnak elő ezek metaforikus értelemben. Egy harmadik módszer a céltartomány szókincsére való rákeresés (pl. Stefanowitsch elemzései), amely talán azért lehet sikeresebb, mint az előző kettő, mert azokban a metaforikus mondatokban, amelyek tartalmaznak egy céltartományi kifejezést, általában egy forrástartományi kifejezés is megjelenik, s így nagyobb az esély az ún. metaforikus sablonok fellelésére. Végül negyedik módszerként olyan mondatokra is rákereshetünk, amelyek egy adott metaforának mind a forrás-, mind pedig a céltartományára jellemző szavakat is tartalmazzák (pl. Martin módszere). Ennek az eljárásnak az a hátránya, hogy így csak előre meghatározott metaforikus leképezéseket tudunk tesztelni, és a Stefanowitsch-féle módszerrel szemben az új metaforatípusok fellelése eleve kizárt. Ezzel szemben nagy előnye, hogy gyorsabban megy az annotálás, így nagyobb szövegeken is alkalmazható.

Természetesen mindegyik esetben szükség van egyrészt megfelelő szólisták összeállítására, másrészt pedig annak explicit meghatározására, hogy mi számít metaforikus kifejezésnek, és mi nem.

Az eddigi korpusznyelvészeti kutatások nagyrészt a metaforikus kifejezések nyelvi jellemzőire voltak kíváncsiak, ezért általában az első három elemzési módszer valamelyikét alkalmazták. Ezzel szemben a jelen tanulmány elsősorban arra keresi a választ, hogy a metaforák szövegekben való automatikus megtalálása mennyire sikeres a fogalmi metaforák hipotézisét, illetve a statisztikai tanulás

elméletét alapul véve. A megfelelő korpusz és elemzési módszer kiválasztására nézve ez a következőket jelentette:

– többféle szövegtípusból álló korpusz vizsgálata;

– a fogalmimetafora-elméletben elfogadott metaforatípusok tesztelése;

– olyan mondatok keresése, amelyeken mind a forrás-, mind pedig a céltartományhoz tartozó szó előfordul;

– a lehetséges forrás- és céltartománybeli kifejezéspárok listájának összeállítása többféle módszerrel.

Ennek megfelelően Lakoff–Johnson (1980), valamint Kövecses (2002) metaforaindexéből 12 széles körben ismert fogalmi metaforát választottunk ki, melyek közül az egyiknek mindkét irányú megvalósulását külön vizsgáltuk (A TÖBB FENT VAN/A KEVESEBB LENT VAN), így tulajdonképpen 13-féle annotáció lehetséges (a példák az általunk annotált szövegekből származnak):

(1) A VÁLTOZÁS MOZGÁS (CHANGE IS MOTION): **jön** a **hideg**; rohamléptekkel **közeledik** a **szünidő**; mélységes **szomorúság járta át** a lelkem

(2) AZ IRÁNYÍTÁS FENT VAN (CONTROL IS UP): **magas rangú** katonatiszt; az amerikai parti **őrség** és a haditengerészet járőrei **felügyelik** a houstoni csatornát

(3) A TÖBB FENT VAN (MORE IS UP): **magasabbra** kúszik az **átlaghőmérséklet**; **mértéken felül** bosszantott az ismeretlen idétlen tréfája

(4) A KEVESEBB LENT VAN (LESS IS DOWN): mély **hangját lehalkítva** folytatta; **leszállították igényüket** kétszáz rúpiáról egy fémpumpára; **lelohad** a **szerelem**

(5) A HALADÁS ELŐRE MOZGÁS (PROGRESS IS MOTION FORWARD): a műszaki **haladás** […] **előrevihet** bennünket ezen az úton; **rendbe jönnek** a dolgok

(6) AZ ERŐFORRÁSOK ÉTELEK (RESOURCES ARE FOOD): rengeteg **áramot fogyaszt**; finom artériahálózat **táplálja vérrel** a sebészek beható vizsgálata alatt álló régiót

(7) AZ ELME GÉPEZET (THE MIND IS A MACHINE): fokozni akarják **szellemi kapacitásukat**; **kattant** valami Mihail Alekszandrovics **fejében**

(8) AZ IDŐ PÉNZ (TIME IS MONEY): nem **pazarolom** az **időmet**; mennyi **időbe kerül** a kivitelezés

(9) A DÜH HŐSÉG (ANGER IS HEAT): a **vita hevében** elfelejtettem bemutatkozni; a **lobbanékony** helytartó milyen különös formában **torolja meg**

(10) A KONFLIKTUS TŰZ (CONFLICT IS FIRE): le akartad **rombolni** a templomot, s erre **tüzelted** a népet; **kitört** a **háború**

(11) AZ ELMÉLETEK ÉPÜLETEK (THEORIES ARE BUILDINGS): a **genetika alapjai**; az öreg előbb **megdöntötte** mind az öt **bizonyítékot**, és aztán […] ő maga **felállított** egy hatodikat

(12) AZ ALKOTÁS ÉPÍTÉS (CREATION IS BUILDING): alapjaiban kell **átalakítanunk** az **életünket**; így **formálhattak jogot** a meghódított területekre

(13) A POLITIKA HÁBORÚ (POLITICS IS WAR): a velejéig korrupt **kormány** és rendőrség **kirabolja** a népet; csakis az **ellenséges propaganda** állíthatja

Mivel széles körben használt fogalmi metaforákat választottunk a testesültség hipotézisének korpuszalapú vizsgálatára, minél reprezentatívabb korpuszt kellett építenünk, amelyben mindegyik választott metaforatípus megfelelő gyakorisággal fordul elő. A projekt eredeti célkitűzései között szerepelt, hogy a metaforákat többnyelvű párhuzamos korpuszon vizsgáljuk, ezért olyan szövegeket kellett szereznünk, melyek mind a négy előirányzott nyelven (magyar, angol, spanyol, olasz) elérhetők és szabadon felhasználhatók kutatási célokra. Jogtiszta szövegeket gyűjteni mind a négy nyelven meglehetősen nehéz, sokszor kivitelezhetetlen feladatnak bizonyult, így végül magyar nyelvű korpuszunk csak három szövegtípust tartalmaz: regények, *National Geographic*-cikkek és filmfeliratok az alábbi arányban:

Szövegtípus	**Szövegszavak száma**
National Geographic-cikkek	68 997
Filmfeliratok	32 148
Regények	208 384
Összesen	309 529

1. táblázat. A korpusz összetétele

Mivel a szövegek különböző formátumokban kerültek a birtokunkba, először is egységesíteni kellett őket: minden dokumentumot UTF-8 karakterkódolású sima szöveggé alakítottunk. A korpuszt szövegszavakra és mondatokra bontottuk, majd minden szövegszóhoz egyértelmű morfológiai elemzést rendeltünk a HunPos morfológiai egyértelműsítő (Halácsy et al. 2007) segítségével.

3.2. A gold standard korpusz

A cél tehát olyan elemző rendszer megépítése volt, amely a korpuszban azonosítja a metaforikus mondatokat, és hozzájuk rendel egy-egy címkét, amely azt jelzi, hogy a tizenkét fogalmi metafora közül melyikhez tartozik az adott mondat. Feltételezésünk szerint az a mondat metaforikus, amelyben egyaránt megtalálható valamely fogalmi metafora forrás- és céltartományához tartozó szó is. Az automatikus elemzés eredményeinek teszteléséhez szükségünk volt egy kézi erővel elemzett **gold standard** korpuszra, mellyel összevetettük az automatikus címkézés kimenetét.

Erre a célra építettünk egy minikorpuszt a teljes korpusz 10 százalékából. A minikorpusz kb. 30.000 szövegszóból áll, és a teljes korpuszt arányosan reprezentálja. A minikorpuszt három részre osztottuk; mindegyik részben két annotátor kézzel bejelölte az általa metaforikusnak ítélt mondatokat. A metaforikusság meghatározásához a Pragglejaz (2007) csoport kritériumait vettük alapul, néhány ezek közül: az állandósult kifejezéseket, „halott metaforákat" vagy azokat, amelyek csak etimológiai szempontból számítanak metaforáknak, nem vettük figyelembe (pl. a *depresszió* nem számít metaforikusnak); az igekötők számítanak (a *le* vagy *fel* mint „fent", illetve „lent" forrástartományok); az allegóriák nem; ha egy metaforának az ellentettjét találtuk, akkor azt nem vettük bele az adott metaforatípusba. Ezenkívül mind a tizenkét vizsgált típusnál külön-külön röviden össze is foglaltuk a fontosabb útmutatásokat. Például az A TÖBB FENT VAN konceptuális metaforánál a következő útmutatót használtuk: „Minden olyan mennyiséget jelentő kifejezést annotálunk, amelyet vertikális skálán képzelünk el, pl. *ár, bér, hőmérséklet*. Minden olyan kifejezést annotálunk, amelyben szerepel a *csúcs* szó: *csúcstermelés, csúcstechnológia* stb. Az olyan kifejezések, amelyek arról szólnak, hogy valamiből sok van, és nagy kupacot alkot – pl. *hegyekben áll, tornyosul* –, nem metaforák, ezeket nem annotáljuk."

Az annotátorok közötti egyetértést a számítógépes nyelvészetben általánosan használt módszerrel, a kappa-mértékkel számítottuk ki: $K = (Pr(a) - Pr(e))/(1 - Pr(e))$, ahol $Pr(a)$ = az A és B annotátorok által egyaránt metaforikusnak ítélt mondatok száma + mindkét annotátor által nem-metaforikusnak ítélt mondatok száma/50, és $Pr(e)$ = a két annotátor metaforikussági ítélete valószínűségének szorzata. Mind az annotátorok közötti egyetértés kiszámításánál, mind pedig az automatikus annotálás eredményének értékelésénél a metaforikus/nem-metaforikus kétértékű döntést vettük csak figyelembe. Az annotálás első fordulójában az annotátorok közötti egyetértés értéke átlagosan 17,25 lett, ezért az annotálási útmutató finomítása után végrehajtottunk egy második fordulót, melynél az egyetértés 47,75 lett, vagyis még mindig igen alacsony. Ezek az eredmények

azt mutatják, hogy a metaforikusság definíciója eleve kérdéses, nehezen meghatározható. Ezért úgy döntöttünk, hogy az annotátorok által metaforikusnak ítélt mondatok unióját vesszük – így 155 metaforikusnak annotált mondat szerepel a **gold standard** minikorpuszunkban.

3.3. Az automatikus azonosításhoz használt szólisták összeállítása

A metaforák automatikus azonosításához a 3.1-ben említett módszerek közül a negyediket alkalmaztuk, vagyis olyan mondatokat kerestünk, amelyekben mindkét tartomány kifejezései szerepeltek egyazon mondaton belül. A hipotézis alapján azt feltételeztük, hogy ha egy mondat tartalmaz egy forrás- és egy céltartományi kifejezést is, akkor jó eséllyel metaforikus lesz. Ehhez szükségünk volt forrás- és céltartományi szavakat tartalmazó szólistákra. Illusztrációként álljon itt egy minta az AZ ELME GÉPEZET fogalmi metaforához tartozó listákból:

Forrástartomány	**Céltartomány**
erő	képzelet
kapacitás	szellemi
élesít	memória
működik	agykéreg
feldolgoz	információ
aktiválódik	agyterület
végrehajt	homloklebeny
létrehoz	fej

2. táblázat. Az AZ ELME GÉPEZET fogalmi metaforához tartozó forrás- és céltartományi szólisták részlete

A forrás- és a céltartomány szólistáinak összeállítását három különböző módszerrel végeztük: (a) asszociációs kísérlet alapján, (b) szinonimaszótár alapján és (c) referenciakorpusz alapján.

Az első módszer esetében a pszicholingvisztikai vizsgálatok körében bevett asszociációs kísérletet választottuk. 138 egyetemi hallgató végezte el a kísérletet, ami a következőképpen zajlott: a kiválasztott fogalmi metaforák hívószavai megjelentek a képernyőn, majd a kísérleti személynek egy perc állt a rendelkezésére, hogy olyan szavakat írjon, amelyek a tesztszóról eszébe jutnak. Például az A VÁLTOZÁS MOZGÁS metafora esetében a *változás* szó jelent meg a képernyőn mint forrástartományi, és a *mozgás* szó mint céltartományi tesztszó. Az így kapott listákat normalizáltuk: kiszűrtük a többszavas kifejezéseket, a tulajdonneveket és

az ellentéteket, feloldottuk a rövidítéseket, majd töveltük a szavakat a Hunmorph morfológiai elemző (Trón et al. 2005) segítségével.

A második módszer során az asszociációs kísérletből nyert szólistákat kibővítettük a szavak szinonimáival a *Magyar szókincstár* alapján (Kiss 2007). Ennek hatására a listák mérete a sokszorosára nőtt, annak ellenére, hogy a szinonimák közül a népnyelvi, szleng és ritkán használt szavakat kihagytuk.

A harmadik módszer keretében Martin (2006) alapján tudatosan válogattunk össze szavakat mindegyik forrás- és céltartományhoz az előzőleg kézzel annotált **gold standard** minikorpuszból. Ebből következőleg ezt a módszert a későbbiekben a korpusznak egy másik 10 százalékán teszteltük.

Mindhárom szólista esetében a következő lépésben az eredeti és a morfológiailag egyértelműsített szövegekből, valamint a szólistákból XML-fájlokat állítottunk elő, amelyekben az eredeti szövegek az egyes szólistáknak megfelően annotációkkal vannak ellátva, azaz a szólisták alapján feltételezett fogalmi metaforák jelölve vannak. Az XML-fájlok korpusznyelvészeti feldolgozásának minden további lépését a GATE-alkalmazás, egy könnyen kezelhető grafikus felülettel ellátott szövegfeldolgozó szoftvercsomag segítségével végeztük (Cunningham et al. 2002). Az automatikusan annotált szöveget kézzel ellenőriztük, és korrigáltuk a szavak többértelműségéből adódó hibákat, azaz kitöröltük azokat a címkéket, amelyek a szólistán szereplő szóval megegyező alakú, de más jelentésű és/vagy szófajú szót jelöltek, pl. az A DÜH HŐSÉG fogalmi metafora esetében az *ég* és *nap* szavaknak az 'égbolt' és '24 óra' jelentésű előfordulásait.

3.4. Eredmények

A három módszer eredményeit a számítógépes nyelvészetben általánosan alkalmazott pontosság *(precision)* és fedés *(recall)* alapján értékeltük. A pontosság ebben az esetben azt mutatja, hogy az automatikus felismerő rendszer által metaforikusnak ítélt mondatoknak mekkora hányada ténylegesen metaforikus. A fedés értékéből azt tudhatjuk meg, hogy az emberi elemzők által metaforikusnak ítélt mondatok közül hányat talált meg a rendszer. Az F-mérték *(F-measure)* pedig ezek súlyozott harmonikus közepe, vagyis a hatékonyság végső mérőszáma.

Az eredményekből látható, hogy az asszociációs módszerrel lényegesen kevesebb olyan mondatot találtunk, amely forrás- és céltartományi kifejezést is tartalmaz, mint a másik két módszerrel. Pontosság tekintetében az asszociációs kísérlet valamivel jobb eredményre vezetett, mint a szótáralapú módszer, de mindkettő messze elmarad a korpuszalapú annotáció eredményétől. Ez utóbbi módszer bizonyult a legeredményesebbnek a fedés tekintetében is, vagyis a metafori-

Módszer	Fedés		Pontosság		F-mérték
Asszociáció	6/155	(3,8%)	6/80	(7,5%)	5,65%
Szótár	28/155	(18,06%)	28/617	(4,5%)	11,28%
Korpusz	41/131	(31,29%)	41/74	(55,4%)	43,34%

3. táblázat. A három módszer eredményei

kusság gépi azonosításában – ellentétben a célzott kézi válogatással – az asszociációs módszeren alapuló pszicholingvisztikai megközelítés nem célravezető.

3.5. Problémás esetek

Az eddigiekből is tisztán látszik, hogy nem könnyű feladat egy mondatról eldönteni, hogy metaforikus-e, vagy sem. Általános tapasztalat, hogy ha emberi erővel nehéz megtalálni egy szövegben bizonyos elemeket, akkor azok automatikus azonosítása sem fog jó eredményt hozni. Összességében ki kell mondanunk, hogy a testesültség elméletén alapuló feltételezésünk, hogy egy metaforikus mondatban meg kell jelennie mindkét tartományi kifejezésnek, nem helytálló. Ezt erősítik a korpusz kézi annotálása során gyűjtött példák is, melyek metaforikusak ugyan, de nem szerepel bennük mindkét tartományhoz tartozó kifejezés, vagyis a fedés szempontjából problémásak.

Bizonyos mondatokban csak forrástartományi szót találunk: *Aztán egy nap **lelépett*** (A VÁLTOZÁS MOZGÁS). Itt csak egy mozgást kifejező szó van a mondatban, míg a változásra explicit módon nem utal semmi, mégis pontosan tudjuk, hogy nem a járdáról való lelépésről van például szó, hanem az esemény szereplőinek életében bekövetkezett változásról. Más esetekben a céltartományi kifejezés szerepel ugyan a szövegben, de nem a célmondatban, hanem az azt megelőző szövegkörnyezetben: *Van Toch kapitány majd megfulladt a felháborodástól* […] *arca bíborszínt öltött* […] *feje **elkékült*** (A DÜH HŐSÉG).

Olyan esetek is léteznek, amikor legkevésbé sem a mondaton múlik annak metaforikussága, hanem csupán egy szón, amely magába foglalja a forrás- és céltartományi jelentést is: ***előléptetés*** (A HALADÁS ELŐRE MOZGÁS).

Továbbá olyan mondatokból is sok van, amelyekben szerepel ugyan mindkét tartományi kifejezés, mégsem metaforikusak: *Mérnökök és **vezetők** tanakodnak kisebb csoportokban a 23 emelet **magas** fúrótorony tövében* (AZ IRÁNYÍTÁS FENT VAN). Ez utóbbi mondatok felelősek az alacsony pontossági értékekért.

4. Tanulságok, összegzés

Mivel kutatásunk célja elsősorban a fogalmi metaforák szövegekben való automatikus azonosítására irányult, nem tértünk ki a talált példák grammatikai elemzésére és a szövegek típusainak a különböző metaforákkal való összefüggéseire sem. Első ránézésre azonban úgy tűnik, vizsgálatunk eredményei megerősítik az előzőekben bemutatott korpusznyelvészeti elemzések eredményeit, főként a kollokációkat és a metaforikus kifejezések nyelvi formáját illetően. Erre utal az is, hogy míg az asszociációs kísérlet segítségével összeállított listák jósló ereje nagyon gyenge volt, addig a tartományokra jellemző szavak szövegből való célzott összeválogatása hozta a legjobb eredményt. Ez azt jelenti, hogy nem bármilyen asszociáció vezet metaforához, hanem csak bizonyos szavak, kifejezések együttes előfordulása. Például a *pazarol* és *idő* vagy a *gerjeszt* és *harag* szavak egy mondaton belül szinte mindig metaforát eredményeznek. Ezenkívül a grammatikai forma fontosságát érintő adatokat is említhetünk: AZ ERŐFORRÁSOK ÉTELEK konceptuális metaforánál a referenciakorpusz alapján a forrástartományt leginkább igék jellemzik (pl. *fogyaszt*, *felfal*, *táplál)*; ezzel szemben az asszociációs módszerrel összegyűjtött szavak többsége főnév (pl. *edény, fagylalt, reggeli)*. Ez megint csak a Deignan (2005) által kapott eredményt támasztja alá: a metaforikus kifejezések jelentős hányadában a forrástartományt képviselő szavak többnyire igeként vagy melléknévként jelennek meg. Ennek az lehet a magyarázata, hogy a metaforikus beszédben általában az absztrakt entitásokat próbáljuk leírni, s így a konkrét forrástartományból leginkább viselkedést, tulajdonságot, cselekvést leíró szavakat veszünk át. Az asszociációs kísérlet egyik gyenge pontja tehát az lehetett, hogy bármilyen szót figyelembe vettünk, függetlenül annak szófajától.

Természetesen ezeknek a feltevéseknek az alátámasztásához a talált metaforák teljesebb elemzésére van szükség. Ugyanakkor az eredeti célkitűzést követve ugyanezeknek a szövegeknek az angol, spanyol és olasz nyelvű változatát is érdemes lesz a jövőben megvizsgálni, és az eredményeket összevetni a magyar adatokkal, hiszen ebből újabb következtetéseket vonhatunk le a nyelvi tényezőkre és a fogalmi metaforák természetére vonatkozóan.

Irodalom

Andrews, Mark – Gabriella Vigliocco – David Vinson 2005. The role of attributional and distributional information in semantic representation. In: Bruno Bara – Lawrence Barsalou – Monica Bucciarelli (szerk.): Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society. Hillsdale NJ: Lawrence Erlbaum. 127–132.

Andrews, Mark – David Vinson – Gabriella Vigliocco 2007. Evaluating the contribution of intra-linguistic and extra-linguistic data to the structure of human semantic representations. In: Danielle S. McNamara – J. Greg Trafton (szerk.): Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society. Hillsdale NJ: Lawrence Erlbaum. 767–772.

Babarczy Anna – Bencze Ildikó – Fekete István – Simon Eszter 2010. A metaforikus nyelvhasználat egy korpuszalapú elemzése. In: Tanács Attila – Vincze Veronika (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem. 145–156.

Boroditsky, Lera – Michael Ramscar 2002. The roles of body and mind in abstract thought. Psychological Science 13: 185–188.

Burgess, Curt – Kevin Lund 1997. Representing abstract words and emotional connotation in high-dimensional memory space. In: Michael G. Shafto – Pat Langley (szerk.): Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society. Hillsdale NJ: Lawrence Erlbaum. 61–66.

Cunningham, Hamish – Diana Maynard – Kalina Bontcheva – Valentin Tablan 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In: Pierre Isabelle (szerk.): Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics. 168–175.

Deignan, Alice 2005. Metaphor and corpus linguistics. Amsterdam & Philadelphia: John Benjamins.

Deignan, Alice 2008. Corpus linguistics and metaphor. In: Gibbs (2008, 280–294).

Fauconnier, Gilles – Mark Turner 2002. The way we think: Conceptual blending and the mind's hidden complexities. New York: Basic Books.

Gentner, Dedre – Brian Bowdle – Phillip Wolff – Consuelo Boronat 2001. Metaphor is like analogy. In: Dedre Gentner – Keith J. Holyoak – Boicho N. Kokinov (szerk.): The analogical mind: Perspectives from cognitive science. Cambridge MA: MIT Press. 199–253.

Gibbs, Raymond W. Jr. 1994. The poetics of mind: Figurative thought, language, and understanding. Cambridge: Cambridge University Press.

Gibbs, Raymond W. Jr. 2006. Embodiment and cognitive science. Cambridge: Cambridge University Press.

Gibbs, Raymond W. Jr. 2007. Experimental tests of figurative meaning construction. In: Günter Radden – Klaus-Michael Köpcke – Thomas Berg – Peter Siemund (szerk.): Aspects of meaning construction. Amsterdam & Philadelphia: John Benjamins. 19–33.

Gibbs, Raymond W. Jr. (szerk.) 2008. The Cambridge handbook of metaphor and thought. Cambridge: Cambridge University Press.

Gibbs, Raymond W. Jr. – Teenie Matlock 2008. Metaphor, imagination and simulation. Psycholinguistic evidence. In: Gibbs (2008, 161–176).

Giora, Rachel 1997. Understanding figurative and literal language: The graded salience hypothesis. Cognitive Linguistics 8: 183–206.

Giora, Rachel 2008. Is metaphor unique? In: Gibbs (2008, 143–160).

Goatly, Andrew 2002. Text-linguistic comments on metaphor identification. Language and Literature 11: 70–74.

Halácsy, Péter – András Kornai – Csaba Oravecz 2007. HunPos – An open source trigram tagger. In: Sophia Ananiadou (szerk.): Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume. Proceedings of the Demo and Poster Sessions. Prague: Association for Computational Linguistics. 209–212.

Keysar, Boaz – Yeshayahu Shen – Sam Glucksberg – William S. Horton 2000. Conventional language: How metaphorical is it? Journal of Memory and Language 43: 576–593.

Kintsch, Walter 2000. Metaphor comprehension: A computational theory. Psychonomic Bulletin & Review 7: 257–266.

Kiss Gábor (szerk.) 2007. Magyar szókincstár. Budapest: Tinta Könyvkiadó.

Kövecses, Zoltán 2002. Metaphor. A practical introduction. Oxford: Oxford University Press.

Kövecses Zoltán 2005. A metafora. Gyakorlati bevezetés a kognitív metaforaelméletbe. Budapest: TypoTEX.

Lakoff, George – Mark Johnson 1980. Metaphors we live by. Chicago: The University of Chicago Press.

Lakoff, George – Mark Johnson 1999. Philosophy in the flesh. The embodied mind and its challenge to western thought. New York: Basic Books.

Landauer, Tom – Suzanne Dumais 1997. A solution to Plato's problem: The Latent Semantic Analysis Theory of the acquisition, induction and representation of knowledge. Psychological Review 104: 211–240.

Levinson, Stephen C. 2000. Presumptive meanings: The theory of generalized conversational implicature. Cambridge MA: MIT Press.

Martin, James H. 2006. A corpus-based analysis of context effects on metaphor comprehension. In: Stefanowitsch – Gries (2006, 214–236).

Murphy, Gregory L. 1996. On metaphoric representation. Cognition 60: 173–204.

Peleg, Orna – Rachel Giora – Ofer Fein 2001. Salience and context effects: Two are better than one. Metaphor and Symbol 16: 173–192.

Pragglejaz Group 2007. MIP: A method for identifying metaphorically used words in discourse. Metaphor and Symbol 22: 1–39.

Sperber, Dan – Deirdre Wilson 1986/1995. Relevance: Communication and cognition. Cambridge MA & Oxford: Blackwell.

Stefanowitsch, Anatol 2006. Words and their metaphors: A corpus-based approach. In: Stefanowitsch – Gries (2006, 63–105).

Stefanowitsch, Anatol – Stefan Th. Gries (szerk.) 2006. Corpus-based approaches to metaphor and metonymy. Berlin & New York: Mouton de Gruyter.

Szamarasz Vera Zoé 2006. Az idő téri metaforái: a metaforák szerepe a feldolgozásban. Világosság 47: 99–109.

Trón, Viktor – László Németh – Péter Halácsy – András Kornai – György Gyepesi – Dániel Varga 2005. Hunmorph: Open source word analysis. In: Martin Jansche (szerk.): Proceedings of the ACL Workshop on Software. Stroudsburg, PA: Associaton for Computational Linguistics. 77–85.

Wilson, Deirdre – Dan Sperber 2004. Relevance theory. In: Laurence R. Horn – Gregory Ward (szerk.): The handbook of pragmatics. Oxford & Malden MA: Blackwell. 607–632.

Conceptual and statistical factors in metaphor identification

Abstract: The paper discusses a model of automatic metaphor identification in Hungarian text corpora. Theories attempting to characterise the nature of abstract knowledge are discussed and two approaches to metaphor acquisition are compared and tested: conceptual metaphor theory and statistical language processing theory. Using the principles of the two approaches, lists of linguistic items potentially signalling metaphoricity were complied. The likelihood of these items correctly predicting metaphoricity was tested in a computer model. The results reveal that the statistical method is more successful but it still shows unexpectedly poor performance. The poor performance of the models may be explained by the fuzzy nature of the concept of metaphoricity and the lack of a precise and formalizable definition of metaphorical meaning.

Keywords: metaphor, corpus-based analysis, natural language processing, cognitive linguistics, text statistics

Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből*

Simon Eszter – Sass Bálint

Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Budapest
simon.eszter@nytud.mta.hu; sass.balint@nytud.mta.hu

A nyelvi kulturális örökség elérhetővé tételében kulcsfontosságú szerep jut a nyelvtechnológiának, melynek módszereivel a kutatók egységes, következetes, nyelvi információval ellátott adatbázisokhoz juthatnak. A nyelvtörténészek és nyelvtechnológusok egyik legfontosabb együttműködési terepe a történeti korpuszok építése, melyek kiváló alapanyagot szolgáltatnak az elméleti kutatásoknak. Cikkünkben egy ómagyar nyelvtörténeti adatbázis létrehozásáról számolunk be, bemutatjuk a teljes korpuszépítési munkafolyamatot a szkenneléstől a korpuszlekérdező eszközig.

Kulcsszavak: kulturális örökség, nyelvtechnológia, történeti korpusz, szövegnormalizálás, korpuszépítés

1. Bevezetés: nyelvtechnológia és kulturális örökség

A társadalom- és bölcsészettudományok területén tevékenykedő kutatók korábban elsősorban papíralapú forrásokból: kéziratokból, könyvekből dolgoztak. Az elmúlt évtizedek során azonban az információhoz való hozzáférés módja a számítógépek és az internet használatának elterjedésével merőben megváltozott. Ma már a könyvtárban sem kell katalóguscédulákat átbogarászni, ha meg akarjuk tudni egy könyv elérhetőségét, hanem viszonylag könnyen és egyszerűen tudunk az interneten keresztül keresni a könyvtári adatbázisokban a könyvekhez tartozó metaadatok (szerző, kiadó, kiadás ideje és helye stb.) alapján. A humán tudományok és az információs technológiák találkozásával egyre több adat válik digitálisan is elérhetővé, akár a nagy mértékű digitalizációs törekvéseknek köszönhetően, akár amiatt, hogy az adat eleve digitális formában jön létre.

A nyelvi kulturális örökség elérhetővé és feldolgozhatóvá tételében kulcsfontosságú szerep jut a nyelvtechnológiának. Az egyszerű digitalizálás, ami ál-

* Az ómagyar korpusz építése a Magyar Generatív Történeti Szintaxis projekt keretében valósul meg. A projektet az OTKA NK 78074. számú pályázata támogatja. Köszönettel tartozunk azoknak a nyelvtörténészeknek és kiadóknak, akik rendelkezésünkre bocsátották az általuk előkészített szöveges kódexátiratokat; továbbá mindazoknak, akik a manuális és/vagy automatikus szövegfeldolgozásban részt vettek. Külön köszönet Novák Attilának, aki a morfológiai elemzést és egyértelműsítést, valamint a Jakab-féle táblázatok átalakítását végzi.

talában kimerül a primér adat képként való beszkennelésében, nem nyújt elég széleskörű és szofisztikált keresési lehetőséget. Az olyan szöveges adatbázisok, melyekben az elemek különféle nyelvészeti (és/vagy történeti, paleográfiai stb.) információval vannak ellátva, sokkal kifinomultabb kutatási alternatívákat kínálnak.

A humán tudományok és a nyelvtechnológia ötvözése mindkét tudományterületnek nagy hasznot hozhat. A kutatók az egyik oldalon időt nyernek a hatékonyabb adateléréssel. A számítógépes feldolgozás támogatja a következetességet, az egységességet és a metaadatok könnyebb kezelését. A digitalizált adat nem helyhez kötött, vagyis a kutatók bárhonnan hozzáférhetnek – akár egy időben, párhuzamosan is.

Ami a dolog nyelvtechnológiai oldalát illeti: a nyelvtechnológusok az elmúlt évtizedekben jellemzően relatíve kicsi, szűk tartományra specializált és szűrt adathalmazokkal dolgoztak. A nyelvi kulturális örökség területén viszont elsősorban a sztenderdtől eltérő, illetve archaikus nyelvváltozatokkal találkozunk, amelyek számos kihívást állítanak a nyelvtechnológusok elé. A korpuszépítési munkálatok során elsősorban már digitalizált szövegekből indulnak ki – de nem ez a helyzet a történeti dokumentumokkal. Az elektronikus formátumok (sőt az elektromosság) előtti korból származó szövegekből való korpuszépítés sokkal idő- és munkaigényesebb folyamat, és bizonyos esetekben más módszereket is igényel, mint a mai szövegek esetében. Már az alapszintű szövegfeldolgozó lépések (szavakra és mondatokra bontás, morfológiai elemzés és egyértelműsítés) során az eddigieknél robusztusabb vagy teljesen új módszerekre van szükség. Az ezen a területen kifejlesztett eszközök valószínűleg a nyelvtechnológia más területein is sikerrel alkalmazhatóak. Vagyis a kulturális örökség digitalizálása során nemcsak a már bevált módszerek új területeken való alkalmazása történik, hanem az új módszerek új kutatási kérdéseket is felvetnek. Ezek megoldásához a különböző tudományterületek képviselői közötti szoros együttműködésre van szükség.

A nyelvtörténészek és nyelvtechnológusok egyik legfontosabb együttműködési terepe a történeti korpuszok építése. A kilencvenes, de legfőképp a kétezres években sorra indultak olyan projektek, melyek egy adott nyelv valamely régebbi változatának digitalizálását és feldolgozását célozzák (Kroch–Taylor 2000; de Sousa–Trippel 2006; Kunstmann–Stein 2007; Thomas et al. 2007). Ezek a korpuszok természetesen sok paraméterükben különböznek: teljes szövegeket vagy csak részleteket tartalmaznak; egy korszak teljes lefedésére törekszenek, vagy egy nagyobb kor szövegeiből kívánnak reprezentatív válogatást adni; morfológiai és szintaktikai annotációt is tartalmaznak, vagy a puszta szöveget adják szövegegységekre tagolva stb. Annyiban azonban megegyeznek, hogy valamilyen szintű

nyelvi információt mindenképpen tartalmaznak, és szofisztikált kereséseket tesznek lehetővé, hogy minél inkább megkönnyítsék a nyelvészeti, irodalmi vagy történelmi célú kutatásokat.

Cikkünkben egy, a fenti trendbe illeszkedő projektet mutatunk be, melynek célja, hogy diakrón szintaktikai vizsgálatokat végezzen magyar nyelvű szövegeken, amihez elsődleges fontosságú egy elektronikus nyelvtörténeti adatbázis létrehozása. A *Magyar Generatív Történeti Szintaxis* című projekt keretein belül felépítünk egy olyan korpuszt, amely tartalmazza az összes fennmaradt ómagyar kori (896–1526) szövegemléket, és amely nyelvészeti információkat tartalmaz elektronikusan előhívható és interpretálható módon.

A cikkben a teljes korpuszépítési munkafolyamatot bemutatjuk. A 2. pontban a korpusz anyagának összegyűjtését írjuk le, majd a 3. pontban a feldolgozási lépéseket a szkenneléstől a betűhű szöveg előállításáig. A 4. és az 5. pont a kézi és a gépi normalizálást mutatja be. A 6. pont a morfológiai elemzés és egyértelműsítés feladatkörét tárgyalja. A 7. pontban azt vizsgáljuk, hogy hol kaphatnak helyet az automatikus, félautomatikus és manuális nyelvfeldolgozó eljárások a korpuszépítési munkálatokban. A 8. pont a korpusz felépítését, a 9. pont pedig a hozzá készült lekérdező eszközt mutatja be. Ugyanitt néhány példán keresztül azt illusztráljuk, hogy a korpusz segítségével milyen típusú nyelvészeti kérdéseket tudunk megválaszolni. Végül az összegzés előtt a korpuszépítéssel kapcsolatos további feladatainkról esik szó.

2. A korpusz anyagának összegyűjtése

A reprezentativitás, de legalábbis a kiegyensúlyozott szövegválogatás a korpuszépítés fontos elve. Ez azonban háttérbe szorul, ha eleve korlátozott az elérhető nyelvi anyag mennyisége (például ha egy holt nyelv vagy egy nagyon speciális nyelvi réteg adja a korpusz anyagát). Ez a helyzet az ómagyar korpusz esetében is, amely – célkitűzésének megfelelően – az összes ómagyar korból fennmaradt szövegemléket tartalmazza. Szövegemléken az összefüggő mondatokat tartalmazó nyelvemlékeket értjük; az ún. szórványemlékekkel, amelyekben csak sporadikusan fordulnak elő magyar szavak vagy nevek, ebben a projektben nem foglalkozunk. Nem szerepelnek továbbá a korpuszban azok a szövegek sem, amelyeket még soha nem adtak ki nyomtatásban, vagyis a nyelvtörténeti átírási munkát nekünk kellene elvégezni.

A fenti megszorításokat figyelembe véve a feldolgozandó ómagyar anyag 48 kódexet, 27 rövidebb szövegemléket és 244 misszilist (elküldött levelet) foglal magában, vagyis mindösszesen körülbelül 2 millió szövegszót.

A korpuszépítés első lépése a valamilyen elektronikus szöveges formátumban már meglévő nyelvtörténeti anyagok összegyűjtése. A különböző forrásokból (kiadóktól, nyelvtörténészektől) származó, változatos fontkészleteket használó dokumentumokat egységes, UTF-8 kódolású, sztenderd Unicode-karaktereket tartalmazó sima szövegfájlokká alakítjuk (l. 3.3. pont).

Másik forrásunk a *Számítógépes Nyelvtörténeti Adattár*, amelyben több ómagyar kódex ábécérendes adattára elérhető (Jakab–Kiss 1994; 1997; 2001; Jakab 2002). A kódexfeldolgozási munkálatok még a hetvenes években kezdődtek a Debreceni Egyetemen Jakab László vezetésével. Az adattárban a kódex címszavai (a szövegszavak tövei mai magyar átírásban) ábécérendbe rendezve szerepelnek. A hozzájuk tartozó betűhű szövegszavakat a lelőhely (lapszám, sorszám) megjelölésével közlik, mellettük számokkal rögzítették az adatra vonatkozó helyesírástörténeti, szótörténeti, hangtani, szófajtani, jelentéstani és alaktani tudnivalókat. A szövegben sokszor előforduló szavakat egy függelékben különítették el, melyeket a lelőhely alapján visszahelyezünk az eredeti kódexbeli helyükre. Az egyes szövegszavak soron belüli sorrendjét nem közlik, ezért a sorbarendezést is elvégezzük. Ezután a többféle fontkészletet alkalmazó táblázatot UTF-8 kódolású sima szöveggé alakítjuk, majd ebből állítjuk vissza a kódexek eredeti betűhű szövegét. Az egyes szövegszavakhoz tartozó morfológiai elemzést az általunk használt morfológiai elemző kimeneti formátumára alakítjuk, továbbá a mai magyar tövek és az elemzés alapján rekonstruáljuk a normalizált szóalakot (l. 4. pont). Ennek a konvertálási munkafolyamatnak a végén megkapjuk az adott kódex szavainak betűhű és normalizált alakját, valamint a hozzájuk tartozó egyértelmű morfológiai elemzést (a feldolgozási szintekről részletesen l. a 8. pontot).

Az ómagyar szövegek nagy részének azonban nincsen elektronikusan elérhető szöveges változata, így ezeket a számítógép által olvasható és feldolgozható formára kell hoznunk. Ez a rövidebb szövegek esetében általában begépeléssel, a hosszabbak esetében szkenneléssel, optikai karakterfelismerő (OCR) program alkalmazásával és kézi ellenőrzéssel történik.

3. A korpusz anyagának feldolgozása

3.1. Szkennelés

Néhány kódex beszkennelt verziója megtalálható a Magyar Elektronikus Könyvtárban, sőt ezek egy része ún. „szendvics" PDF, vagyis a kép mögött megtalálható az OCR-ezett szöveg is. Ennek ellenére ezeket nem tudtuk használni: a mögöttes

szöveg nem esett át kézi ellenőrzésen, vagyis meglehetősen sok benne a hiba, a képek felbontása pedig nem elég jó az OCR-ezéshez.

Így minden kódexet, amelyet nem tudtunk szöveges formában megszerezni, minimum 300 dpi felbontásban beszkenneltünk.

3.2. Optikai karakterfelismerés

Az ómagyar kódexekben található nagyszámú különleges karakter kezelése miatt az OCR programmal szemben alapvető elvárásunk volt a **taníthatóság**. Ez utóbbi azt jelenti, hogy a program nem zárt karakterkészlettel dolgozik, hanem meg lehet neki adni bármilyen új karaktert. A szóba jöhető nyílt forráskódú szoftverek közül a *Tesseract*ot próbáltuk ki, amelynek az a hátránya, hogy az összes felismerendő dokumentum alapján egy egész karakterkészletet (nyelvet) kell megtanítani neki. Ezért végül az *Abbyy FineReader 9.0 Professional edition* mellett döntöttünk. Ez ugyan nem nyílt forráskódú, de karakterről karakterre, interaktív módon tanítható, és elég jó minőségű kimenetet ad.

Az OCR program teljesítményét szópontossággal *(word accuracy, WAcc)* mértük, amely egy dokumentumban a helyesen felismert szavak és az összes szó számának az aránya. Az előzetes elvárásoknak megfelelően az eredmények azt mutatják, hogy a pontosság nagyban függ a kódexekben alkalmazott helyesírástól. Kniezsa (1952) az ómagyar kori kódexek kezeinek helyesírását három nagy típusba sorolja; a kiértékelésnél ezt a kategorizálást követtük. A mellékjel nélküli helyesírás a latinban nem szereplő magyar hangokat több betű kombinációjával írja le, például: *cs* → *ch* ~ *cz* ~ *chy* ~ *chi* ~ *cy*. A mellékjeles helyesírás egy rokon hang betűjének mellékjeles változatával jelöli ezeket, például: *cs* → *č* ~ *ć*. A harmadik típus pedig ezek keveréke, amely egy hang jelölésére karakterkombinációkat és diakritikus jeleket használ (akár egyszerre is), például: *cs* → *ch* ~ *chy* ~ *cyh* ~ *c* ~ *chi* ~ *č* ~ *ch*'. A kiértékeléshez három kódexet választottunk a három különböző típusból, továbbá összehasonlítási alapként egy rövidebb mai magyar szövegen is kiértékeltük a szoftver teljesítményét.

Az 1. táblázatból kiolvasható, hogy legjobban a mellékjel nélküli helyesírással boldogult a program: ez nagyjából megegyezik a mai magyar szövegek felismerésében nyújtott pontossággal. A mellékjeles és keverék helyesírású kódexekben használt speciális karakterek nagy száma a tanítás ellenére is közel 30%-kal rontotta a pontosságot. A mellékjel nélküli kódexek esetében a latin ábécé betűit kell felismerni, ezért itt az OCR program jó teljesítményt nyújt. A bonyolult, akár többszörös, illetve egymáshoz hasonló ékezetek elkülönítése viszont problémát okoz. A jelentős teljesítménycsökkenés hátterében tehát ezeknek a diakriti-

kus jeleknek a nem kielégítő kezelése állhat, ahogy erről például Volk et al. (2010) is beszámol.

Kódex	Helyesírás	Tokenszám	Felismert	WAcc (%)
Kulcsár	mellékjel nélküli	36.321	35.258	97,07
Müncheni	mellékjeles	74.657	50.790	68,03
Czech	keverék	11.478	7.910	68,91
–	mai magyar	5.121	5.068	98,97

1. táblázat. Az OCR szópontossága helyesírási típusok szerint

3.3. A betűhű szöveg előállítása

A betűhű szöveg elkészítésekor nem a kódexek kézzel írott változatát, hanem az általunk használt átirat szerkesztőjének konvencióit követjük, vagyis nem feltétlenül törekszünk tökéletes paleográfiai pontosságra. Például a Jókai-kódex esetében a Jakab-féle adattárból (Jakab 2002) indultunk ki, amely nem jelöli külön a korban gyakran használt, ám a nyelvtörténészek nagy része szerint jelentésmegkülönböztető szereppel nem rendelkező hosszú *s*-t (ſ). Így ebben a kódexben mi sem jelöljük ezt a karaktert, annak ellenére, hogy a kódexek jelentős hányadában jelölve van. Ahol egyedi indokkal mégis eltérünk a szerkesztő közlésétől, azt mindig külön jelezzük.

A szabványosság előnyei miatt a teljes korpuszt UTF-8 kódolású sztenderd **Unicode-karakterekkel** tároljuk és jelenítjük meg. A nemzetközi Unicode szabvány (http://unicode.org) éppen azért jött létre, hogy a világ összes nyelvének összes karakterét egy kódolási rendszerbe foglalja, lehetővé téve minden ma használatos karakter egységes megjelenítését. Mivel minden platformon elérhető, széles körben elterjedt és elfogadott szabvány, érdemes volt az ómagyar karakterek tárolására és reprezentálására is az UTF-8 kódolású Unicode-ot választani. A Unicode nagy előnye, hogy az alapkaraktereket és a diakritikus jeleket külön egységekként (külön kóddal) tárolja, és lehetőséget nyújt ezek szabad összeépítésére. Így nemcsak az *a*-ból és a vesszőből (´) gyárthatunk *á*-t, hanem például az *y*-ból és az umlautból (¨) is előállíthatjuk az ómagyar kódexekben nagyon gyakori *ÿ* karaktert. A hozzáadott ékezetek halmozhatók is, így ezen a módon a kódexek különleges karaktereinek jelentős részét szabványos kódolással tudjuk reprezentálni.

Mindenképpen szükséges egy, az egész korpuszra kiterjedő, szigorúan **egységes** formátum; ez teszi lehetővé, hogy a lekérdezéseket az egész anyagra vonatkoztathassuk. A korpuszok egyik haszna, hogy nem csak példákat szolgáltatnak bizonyos jelenségekre, hanem adott lekérdezésre az **összes** találatot megad-

ják, ezáltal lehetővé teszik a jelenségek statisztikai vizsgálatát is. A korpusz ezen fontos tulajdonságát csak úgy biztosíthatjuk, ha következetesen betartjuk azt az alapelvet, hogy azonos dolgokat mindig ugyanúgy, különbözőeket pedig mindig eltérően jelölünk. Ugyanakkor viszonylag nagy erőfeszítést kíván ennek az egységességnek a megvalósítása, mert előfordulnak olyan régi magyar karakterek is, melyek a sztenderd kódtáblában nincsenek reprezentálva. Ezeket a karaktereket egy kiválasztott Unicode-karakterrel helyettesítjük, mégpedig úgy, hogy az adott helyettesítő karaktert kizárólag az adott hiányzó eredeti karakter helyett használjuk a korpuszban. Jó példa erre az ún. **huszita *cs***, amely megjelenésében leginkább egy kiskapitális L-hez hasonlítható, és amelyet Volf (1874)-et követve rendre *č*-vel helyettesítünk.

Éppen a Unicode-táblában nem szereplő különleges karakterek teszik szükségessé, hogy a háttérben egy másik fajta kódolást is alkalmazzunk. Az ún. **Prószéky-kódban** a különböző diakritikus jelekkel ellátott és speciális történeti karaktereket betűk és számok kombinációjával jelöljük: például az *á*-t a1, az *ö*-t 02, az *ű*-t u3 jelöli. A Magyar Történeti Korpusz számítógépes adatbázisának előállításakor használt kódtáblából (Kiss – Pajzs 2001) indultunk ki, amelyet az ómagyar kori speciális karakterek nagy száma miatt folyamatosan bővítünk. Minden szöveget a Unicode-változat mellett Prószéky-kódokkal is rögzítünk, amivel a Unicode hiányosságai ellenére is rögzíteni tudunk minden információt. A betű–szám kombinációk alkalmazása a szövegbevitel és -javítás során is hasznos, mivel így a begépelők és a nyers OCR-kimenet javítását végzők operációs rendszertől és szövegszerkesztőtől függetlenül, egyszerűen be tudják vinni a speciális karaktereket is.

A betűhű szövegváltozat előállításakor a korabeli írásjeleket, elválasztásokat (illetve azok hiányát), egybe- és különírást, a mondat- és tulajdonnévkezdő kis- és nagybetűket megtartjuk úgy, ahogy a kódexkiadásban szerepelnek. Az eredeti kódexbeli színezéseket, betűvastagításokat és kiemeléseket nem őrizzük meg, és a nyomtatott kiadás során belekerült, sor- és oldaltörést jelölő virgulákat is elhagyjuk.

4. Normalizálás

Az ómagyar kori szövegemlékeket és kódexeket a latin nyelvű és vallásos tárgyú irodalom fordításának igénye hívta életre, de a latin ábécé magyarra alkalmazása számos problémát vetett fel. A legfőbb gond abból fakadt, hogy a magyar hangrendszer több eleme a latinban ismeretlen, így ezek jelölésére új jeleket kellett bevezetni. Az ómagyar kor több mint hat évszázadot fog át, amelynek során nem

volt egységes hangjelölési rendszer, sőt egy kódexet akár több kéz is jegyezhetett, ami további egyenetlenségeket okoz a szövegekben. A különböző helyesírási rendszerekben is ritka az egy hang–egy betű megfelelés (vagyis amikor egy hang jelölésére mindig ugyanaz a betű használatos, és az adott betűnek mindig egy hangértéke van), de egy alakulóban levő helyesírási rendszerben ilyenfajta következetesség még kevésbé van jelen. Sőt inkább az a tipikus, hogy egy emléken belül is ingadozik egy-egy hang jelölésmódja (pl. HB: *kinec* [kinek]), vagy többes hangértéke van egy-egy betűnek (pl. HB: *gimilcíctul* [gyümölcsöktől]). Tovább bonyolítja a helyzetet, hogy néhány betű egyaránt utalhat magánhangzóra és mássalhangzóra is, például az *u,v,w* több évszázadon át jelölhette az *u, ú, ü, ű, v, β* hangok bármelyikét (Korompay 2003).

E probléma megoldása céljából szükség van egy ún. **normalizálási** lépésre, amelynek során az eredeti betűhű szóalakokat mai magyar helyesírású szavakra alakítjuk át. A többféle, különböző nyelvtörténeti szakmai érvekkel alátámasztható lehetséges feldolgozási forgatókönyvek egyik gyakori közös átalakító lépése ez a fajta normalizálás (pl. McEnery–Hardie 2003). A szövegfeldolgozásnak ez a lépése kritikus fontosságú, enélkül ugyanis a (fél)automatikus annotáció hatékonysága a következő lépésekben drámaian visszaesik (Rayson et al. 2007).

A normalizálás során két alapelvet tartunk szem előtt. Első elvünk, hogy az összes ma nem létező szót, toldalékot, morfológiai konstrukciót megtartjuk, vagyis morfémát nem toldunk be, és nem hagyunk el. A 2. táblázat utolsó sora kiváló példa erre a jelenségre: a *-va/-ve* végű határozói igenév személyragozható volt, sőt a teljes paradigmája megvolt ebben a korban (Adamikné Jászó 1992). Ha a normalizálás során ezt az alakot a ma használatos *-va/-ve* végű alakra írnánk át, nyilvánvalóan elvesztenénk a morfológiai információt.

Betűhű	**Normalizált**	**Értelmezés**
villamik	villamik	villámlik/villanik
isa	isa	bizony
iesek	jeszek	jövök
ymaduam	imádvám	imádva E/1.

2. táblázat. A normalizálás első alapelve

A normalizálás második alapelve, hogy elhagyunk minden fonológiai és helyesírási esetlegességet, vagyis egységes, és amennyire lehet, a mainak megfelelő helyesírásra törekszünk. Ez utóbbi azt is jelenti, hogy egy adott szót mindig ugyanúgy írunk le – ez is az egységesség elvének egy megnyilvánulása (vö. 3.3. pont).

A normalizálási lépés során történik meg a szöveg szavakra és mondatokra való bontása is – mindkettő manuális munkával. Az ómagyar szövegekben a sza-

Betűhű	**Normalizált**
mēden	minden
menden	minden
minden	minden
algyu	ágyú
agyu	ágyú
strumlast	ostromlást

3. táblázat. A normalizálás második alapelve

vak egybeírása és elválasztása nem a mai szabályokat követi. Ezért a tokenizálás, vagyis a szöveg szavakra szegmentálása során az ómagyar szövegben a szavakat a mai helyesírásnak megfelelően összevonjuk, illetve szétválasztjuk, természetesen jelölve a változtatásokat.

A ma használatos logikai-grammatikai írásjelezés kibontakozása csak a 17. században kezdődik, vagyis a korabeli központozásra nem támaszkodhatunk a mondatra bontásnál. Ezért a mai értelemben vett automatikus mondatra bontás lehetetlen vállalkozásnak tűnik, így ezt a szövegfeldolgozási lépést is manuálisan végezzük el. Természetesen a kézi mondatra bontás sem mindig egyértelmű – kétséges esetben inkább nem teszünk mondathatárt, vagyis azt az elvet követjük, hogy a mondat legyen inkább hosszabb, mint rövidebb. Alapesetben az alárendelő tagmondatot nem választjuk el a főmondattól, míg a mellérendelő tagmondatot igen. A feladat végrehajtása során a mai központozási alapelvekhez igazodunk.

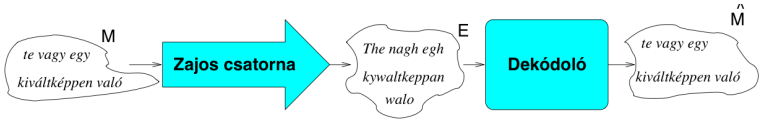
Mivel a korabeli szövegek jó része vallási tárgyú, nagyon sok bibliai nevet találunk bennük. Az egységesség jegyében a különböző bibliafordításokban és bibliai históriákban említett tulajdonneveket is normalizáljuk, vagyis az adott nevek különbözőképpen használt alakjait egységesítjük. Ehhez a Szent István Társulat bibliafordítását használjuk: minden tulajdonnevet abban az alakban normalizálunk, ahogy ebben a kiadásban szerepel. Természetesen ez sem mentes a következetlenségektől: bizonyos neveket ebben a kiadásban sem közölnek mindig egységesen. Ilyen esetekben a kétféle névhasználat közül a gyakoribbat választjuk.

5. Gépi normalizálás

Mivel a normalizálás rendkívül időigényes manuális munka, megpróbáltuk kiváltani automatikus eljárással. A folyamat számítógépes modellezésének célja az volt, hogy választ kapjunk arra a nagyon fontos gyakorlati kérdésre, hogy a szükséges emberi erőforrás alkalmazása leszűkíthető-e a teljes anyagnál nagyságren-

dekkel kisebb méretű kézzel normalizált részkorpusz előállításának feladatára, mely az automatikus módszerhez **tanítókorpuszként** szükséges. Mivel ez a szövegnormalizáló konverzió analóg több klasszikus nyelvfeldolgozási probléma során jelentkező feladattal, így érdemesnek tűnt az azokban sikerrel alkalmazott módszerek adaptálása és eredményességének vizsgálata.

Fő kérdésünk az volt, hogy az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe, és melyek azok a jegyek, amelyeknek a felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását eredményezi. Ennek érdekében szükség volt az adott modellben használt jegyeket tartalmazó, specifikusan annotált tanító szövegekre, melyekből korlátozott mennyiség áll rendelkezésünkre – éppen a normalizálás szakértelmet kívánó, időigényes volta miatt. A fentebb leírt szövegbeli egyenetlenségek miatt nehéz egyértelmű konverziós szabályokat meghatározni, és emiatt kritikus kérdés az is, hogy a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvemlékekre. Mindezek miatt célszerű a problémát valamilyen valószínűségi alapú paradigma keretei között vizsgálni. Az átírás (transzliteráció) nyelvtechnológiai szempontú kutatásának igen gazdag eszköztára van, a különféle módszerek közül mi Shannon zajoscsatorna-modelljét (Shannon 1948) választottuk. (A feladat lehetséges megközelítéseiről bővebben l. Oravecz et al. 2009; 2010.)



1. ábra. Szövegnormalizálás zajoscsatorna-modellben

Az 1. ábrán látható modellben az eredeti szöveget úgy tekintjük, mint a normalizált változat egy zajos kommunikációs csatornán átment „eltorzított" változatát. M jelöli a normalizált szövegváltozat egy részét (a példában egy részmondatot), E pedig ennek betűhű átiratát. A dekódoló feladata annak az M karaktersorozatnak a megtalálása, amelyre a $P(M|E)$ feltételes valószínűség maximális, vagyis a Bayes-tételbe behelyettesítve:

(1) $\hat{M} = \text{argmax}_M P(M|E) = \text{argmax}_M P(E|M)P(M)$

A feladat egyrészt a $P(E|M)$ csatornamodell, másrészt a $P(M)$ forrásmodell meghatározása.

A **csatornamodell** az „eredeti betűhű szöveg → normalizált változat" leképezésekből áll elő. Ehhez szükségünk volt egy tanítókorpuszra, amely két óma-

gyar kori szövegemlék (Müncheni emlék, Szabács viadala) nyelvész szakértők által kézzel normalizált változatából állt elő. A két nyelvemlék tokenszáma (a nem magyar nyelvű részek elhagyásával) összesen 1525. Gépi eszközökkel és kézi ellenőrzéssel karakterszinten párhuzamosítottuk a betűhű és a normalizált szövegváltozatokat, így a tanítókorpusz körülbelül 17.000 megfeleltetést tartalmaz. Ebből már kiszámítható az egyes megfeleltetések valószínűsége.

A **forrásmodell** azt modellezi, hogy a normalizált szövegben milyen valószínűséggel szerepelnek bizonyos karakterszekvenciák. Mivel a normalizált szöveg a mai magyarhoz nagyon hasonló, a forrásmodell előállításához a rendelkezésünkre álló mai magyar szövegeket tartalmazó korpusz megfelelő. Ezért ezt a Magyar Nemzeti Szövegtár (Váradi 2002) egyik alkorpuszából, mintegy 10 millió szóból, 65 millió karakterből állítottuk elő.

Adott E sztring esetén az (1) képlet szerinti $\hat{M}$ értéket kellett kiszámítanunk. Ehhez az eredeti betűhű szöveg minden tokenjéből a csatornamodell megfeleltetései alapján a lehetséges normalizált változatokat legeneráltuk, melyekhez a modell hozzárendelte a valószínűségüket is. Ennek alapján kaptunk egy rangsort a lehetséges változatokra, amelyet aztán a forrásmodell segítségével újrarendeztünk – így alakult ki az eljárás kimenete. (Az eljárás teljes leírásához l. Oravecz et al. 2009; 2010.)

A kimenet minden egyes ómagyar szóhoz a legjobb n normalizált alakot tartalmazó lista. Ennek illusztrációja a 2. ábrán látható. A módszer valós haszna abban mutatkozik meg, hogy a manuális annotáció redukálható a felkínált alakok közötti választásra, ami jelentősen felgyorsítja a normalizálási munkát.

fwl (fül)=>		ygen (igen)=>	
-8,80780895229285	föl	-10,8729908279143	igén
-10,7227286786192	fel	-11,3178857141749	igen
-11,0558158154337	fül	-11,5989613202567	igény
-11,2756412387919	föl	-13,4229320257043	igyen
-12,4574295350367	fol	-14,3578433608162	igin
-12,790296695296	ful	-14,478835649955	igyén
-13,519092302452	fely		
honneg (honnét)=>		sabach (szabács)=>	
-19,1117218113907	honneg	-17,2582527599661	szabács
-19,5230300429664	honnég	-18,1187648297282	sabács
-20,8376176340216	honnét	-18,6771909747334	szabacs
-21,8538140705439	honyneg	-19,1848409742852	sábacs
-22,2098585020436	honynég	-19,5520665992527	szabach
-22,5639991398073	hónneg	-19,9685260661797	szabách

2. ábra. Legjobb n listák különböző bemenetekre

6. Morfológiai elemzés és egyértelműsítés

A normalizálásnak két fő célja van: egyrészt ez teszi lehetővé, hogy a sokféleképpen írt szavak összes előfordulását megtaláljuk, másrészt a normalizált szövegváltozat képezi a morfológiai elemző bemenetét. Mivel a normalizálás során az ómagyar szöveget mai magyarra írjuk át, az ez utóbbira kifejlesztett automatikus morfológiai elemzőt viszonylag könnyen tudjuk alkalmazni a nyelvemlékek feldolgozására. Jelen projektben a **Humor** elemzőt használjuk (Prószéky–Kis 1999). Az egyik normalizálási alapelvünk, hogy minden morfológiai konstrukciót megtartunk, ezért természetesen ki kell bővítenünk a lexikont és a szabályhalmazt bizonyos ma már nem létező, de az ómagyarban még használt nyelvi jelenségek leírásával.

A morfológiai elemző kimenetének egyértelműsítését automatikusan végezzük, utólagos kézi ellenőrzéssel. A 2. pontban ismertetett Jakab-féle táblázatok konvertálásával előállt normalizált és morfológiailag egyértelműsített anyag tanítókorpuszként tud szolgálni egy gépi egyértelműsítő számára. Ennek a kimenetét aztán – a gépi normalizáló kimenetének kezeléséhez hasonlóan – kézzel ellenőrizzük.

Már a normalizálás során felmerül az a probléma, hogy vannak olyan ómagyar szóalakok, amelyeket a szövegkörnyezet alapján sem lehet egyértelműen normalizálni. Például: BécsiK 253. o.: *kic nē hallottac* [kik nem hallottak/hallották]. Mivel ebben a korban a magánhangzó hosszúságát nem jelölték, és a mondat itt véget ér, nem tudjuk, hogy a *hallottac* szóalak határozott vagy határozatlan ragozású. Az ilyen esetekben a normalizálás, valamint a morfológiai elemzés és egyértelműsítés során is megőrizzük a szóalak alulspecifikáltságát.

7. Automatikus vagy manuális?

Amint láttuk, egy jelentős méretű korpusz előállítása során számos nagy munkaigényű feldolgozó lépést kell megvalósítani. Az egyik lehetőség, hogy aprólékos manuális munkával szavanként dolgozzuk fel és ellenőrizzük a korpuszt. Ugyanakkor a nyelvtechnológia célja éppen az, hogy bizonyos feladatokat a számítógép segítségével meggyorsítson, vagy egészében automatikusan megoldjon. A modern nyelvtechnológiai eszközök az alapszintű feldolgozó lépéseket (szavakra és mondatokra bontás, morfológiai elemzés) nagy sebességgel, nagy mennyiségű (akár milliárd szónyi) szöveget feldolgozva jó minőségben oldják meg.

Az automatikus nyelvtechnológiai módszerek két nagy csoportra oszthatók: szabályalapú, valamint statisztikai, gépi tanulási módszerekre. Mindkét eset-

ben valamilyen módon a szabályszerűségeket próbáljuk feltérképezni; a két megközelítés között lényegében az a különbség, hogy az ember vagy a gép alakítja-e ki a szabályrendszert. A gépi tanulási módszerek egy jelentős csoportjában az algoritmusok egy mintahalmaz (tanítókorpusz) alapján fedezik fel az összefüggéseket. Ezek az algoritmusok tehát a megfelelő nyelvi információval felcímkézett korpuszok segítségével taníthatók és tesztelhetők.

Az automatikus módszerek jó teljesítményt nyújtanak, de nem hibátlanok. A teljes hibamentesség nem érhető el, de bizonyos területeken (pl. tulajdonnévfelismerés) 95% fölötti teljesítmény is elérhető. Fontos látni, hogy az automatikus módszerek alkalmazása sok esetben egyáltalán nem jelent kompromisszumot a minőség tekintetében, mivel a manuálisan végzett elemzés, címkézés szintén nem hibamentes. Véletlenül is előfordulhatnak hibák az elemző, annotátor figyelmetlensége miatt, ennél fontosabb azonban, hogy a manuális elemzésnek is van egy minőségi határa. Azokban az esetekben, amikor ugyanazt a szövegrészt több ember párhuzamosan annotálja, egyértelműen megmutatkozik, hogy minél nehezebb(en megfogalmazható) egy annotálási feladat, annál kisebb az egyetértés az annotátorok között. Ilyen feladatok esetén már az emberi munka hibaszintjét közelítő automatikus megoldás is jelentős eredmény.

Abban, hogy egy nyelvfeldolgozási lépés megvalósításakor automatikus vagy manuális megoldáshoz folyamodunk, természetesen számít a feldolgozandó anyag mérete is. Kis méretnél reális alternatíva a manuális munka, illetve az automatikus elemzés manuális ellenőrzése, nagy méretnél azonban kizárólag az automatikus feldolgozásra hagyatkozhatunk. Bizonyos speciális vagy újszerű feladatoknál megbízható automatikus eszközök hiányában nagyobb a manuális munka létjogosultsága.

A továbbiakban a jelen projektben alkalmazott szövegfeldolgozási lépéseket tekintjük át automatizáltságuk szempontjából (vö. 4. táblázat).

Az optikai karakterfelismerés (l. 3.2. pont) feladatára a mai nyelvekre kifejlesztett megbízható automatikus eszközök állnak rendelkezésre. A fő nehézséget az ómagyar anyagban található különleges karakterek: a kombinált diakritikus jelek és a latin ábécén kívüli karakterek kezelése jelenti. Amint ez az 1. táblázatból látható, a tanítható OCR program a latin alapkaraktereket kiválóan felismerte, a mellékjeles karakterek esetén azonban jóval gyengébb teljesítményt mutatott. Az OCR kimenetét hibamentessé kellett tennünk, hogy a további feldolgozó lépések tiszta, zajmentes adatokon dolgozhassanak, ezért a hibákat kézi erővel javítottuk. A fenti két lépés együttese tekinthető **félautomatikus** karakterfelismerésnek, amely (a hosszabb szövegek esetében) a begépelésnél gazdaságosabbnak bizonyult.

A normalizálás átfogó nyelvtörténeti szakértelmet igényel, és rendkívül időigényes, emiatt megkíséreltük a manuális munkát automatikus eszközzel segíteni. A statisztikai algoritmus (l. 5. pont) nehezen boldogul az egységes írásmód hiánya miatt nagyon szabálytalan ómagyar szöveg kezelésével, ezért azt a megoldást választottuk, hogy automatikusan felkínálunk valószínű normalizált alakokat, és az ezek közül való választás már kézzel történik. A normalizálás tehát szintén **félautomatikus**.

A meglévő robusztus mai magyar morfológiai elemzőre támaszkodva a morfológiai elemzés **automatikusan** történhetett. Az elemző adaptálásával megbízható ómagyar elemzőhöz jutottunk. Az adaptálás során egyrészt új tövekkel bővítettük az elemző szótárát, másrészt pedig új alakok kezelésére tettük alkalmassá az ómagyar ragozási paradigmáknak megfelelően.

Az utolsó feldolgozási lépést, az egyértelműsítést – melynek során az egyes szóalakokhoz rendelt több alternatív morfológiai elemzés közül választjuk ki a valóban érvényeset –, az OCR-ezéshez és a normalizáláshoz hasonlóan **félautomatikusan** végezzük.

Összefoglalva elmondhatjuk, hogy ha kellően robusztus eszközök állnak rendelkezésre, akkor előnyösebb a gazdaságos, automatikus megoldás választása. De a különféle automatikus módszerek megfelelő eszközök hiánya esetén is segíthetik a kézi munkát, azaz ilyenkor a félautomatikus megoldást választjuk. A tisztán manuális megoldáshoz akkor folyamodunk, ha különösen fontos a hibamentesség, illetve nincs elegendő/megfelelő tanítóanyag az automatikus módszerek tanításához.

8. A korpusz felépítése

A korpusz felépítése, vagyis az egyes szövegszavakhoz tartozó annotációs szintek párhuzamosan alakulnak a szövegfeldolgozottsági szintekkel, melyeket a 4. táblázatban láthatunk. Ezek alapján hat annotációs szintet és öt feldolgozó lépést különíthetünk el.

Ahhoz, hogy a korpuszban a nyelvi jelenségek kereshetők legyenek, vagyis az adatbázis használható segédeszköze legyen az elméleti nyelvészeti és nyelvtörténeti kutatásoknak, a releváns információkat elektronikusan előhívható és interpretálható módon kell tárolni. A kifinomult, nyelvészetileg releváns lekérdezések sok esetben különféle nyelvi szinteken megjelenő információra hivatkoznak. Hogy ezek mind elérhetőek legyenek, adatbázisunk párhuzamosan tartalmazza a 4. táblázatban látható szövegfeldolgozottsági szinteknek megfelelő nyelvi adatokat. Vagyis minden egyes szövegszóhoz a következő adatok tartoznak:

(1)	kiadott kódex szkennelve → *automatikus* OCR
(2)	nyers OCR-kimenet → *kézi* javítás, kódolás
(3)	betűhű elektronikus forma → *félautomatikus* normalizálás
(4)	normalizált forma → *automatikus* morfológiai elemzés
(5)	szótövesített és morfológiailag elemzett forma → *félautomatikus* egyértelműsítés
(6)	egyértelműsített korpusz

4. táblázat. Szövegfeldolgozottsági szintek

- betűhű forma (3): *adÿad*
- normalizált alak (4): *adjad*
- szótő (6) alapján: *ad*
- morfológiai elemzés (6): *V.Sub.S2.Def*

A korpusz anyaga vertikális fájlok formájában készül el. Ezek `.tsv` formátumú táblázatok, amelyek soronként egy szövegszót tartalmaznak. Az egyes szövegfeldolgozottsági szintekhez tartozó információkat a megfelelő oszlopokban kódoljuk, ahogy az 5. táblázat mutatja (a példa a Bécsi kódexből származik, amelynek a morfológiai elemzése és egyértelműsítése még nem készült el, ezért nem szerepel benne a szótő és a morfológiai információ).

kéz	**könyv**	**oldal**	**fejezet**	**vers**	**betűhű**	**norm**	**ért**	**megj**
1	Rut	4	2	8	Es	és		
1	Rut	4	2	8	monda	mondá		
1	Rut	4	2	8	Booz	Boász		
1	Rut	4	2	8	[Noėminèc]			FAIL
1	Rut	4	2	8	Rutnac	Rutnak:		
1	Rut	4	2	8				
1	Rut	4	2	8	Halgassad	hallgassad,		
1	Rut	4	2	8	leaṅom·	leányom:		

5. táblázat. A vertikális fájlformátum

A korpusz a különböző szinteken feldolgozott szövegen kívül számos metaadatot tartalmaz. Az elsődleges metaadatok az ún. **lókuszjelölők** (l. az 5. táblázat első öt oszlopát), melyek megadják, hogy a dokumentumban hol szerepel az éppen aktuálisan keresett szövegszó. A lókuszjelölők szövegenként változnak, de annyiban

megegyeznek, hogy mindig az eredeti kódex helyeire vonatkoznak, nem a nyomtatott kiadáséira. A például hozott Bécsi kódex esetében rögzítjük a kódexmásoló kezek sorszámát, valamint a bibliai könyv- és versszámozást is, hogy más bibliakiadásokban is visszakereshető legyen az adott szövegrész.

A vertikális fájl tartalmaz egy **értelmezés** mezőt is, amelybe a normalizált alak mai magyarra való „fordítása" kerülhet, például az ómagyar *jonh* szó mai magyar *szív* megfelelője. Az a tény, hogy külön mezőben rögzítjük az értelmezést, természetesen nem jelenti azt, hogy a normalizálás során nem történik értelmezés. Normalizálás és értelmezés szorosan összefüggenek, az utóbbi feltétele az előbbinek. Például az Ómagyar Mária-siralom *buthuruth* szavát csak akkor tudjuk normalizálni, ha rájövünk, hogy ennek a jelentése 'bútőr, a fájdalom tőre' (Korompay 2003).

A **megjegyzés** rovat egyrészt szabad szöveges megjegyzések rögzítésére alkalmas, másrészt ide kerülnek a szöveghez tartozó egyéb metaadatok is különböző kódok formájában. A korpusz az alábbi metaadatokat tartalmazza:

- Ha a cím a szöveg része, akkor szövegként kódoljuk, és a megjegyzés rovatba `TITLE` kód kerül. Ha a cím nem a szöveg része, akkor lókuszjelölőként funkcionál, vagyis külön oszlopot kap.

- A szövegekben előforduló idegen nyelvű szavakat, amelyek a szöveg részét képezik, felvesszük a korpuszba, és a `LANG{nyelv}` címkét adjuk nekik, amellyel egyben azt is jelezzük, hogy ennek a szónak nem lesz morfológiai elemzése. Ha az idegen nyelvű szó magyarul ragozódik, akkor magyar szóként kezeljük, vagyis normalizáljuk, és elemezzük.

- A betűhű szövegváltozat a szkriptor javításait is tartalmazza. Ezeket a következőképpen jelöljük: szkriptor általi utólagos **betoldás** (kód: `ADD`), **szövegrekonstrukció** eredményeként létrejött betoldás (kód: `RECO`), az eredeti szövegben szereplő **áthúzott** szöveg (kód: `STRIKE`), a szkriptor által **elírt**, de nem áthúzott szó (kód: `FAIL`), **töredékes** szó (kód: `FRAG`). Ha csak a szó egy részét érinti a felsorolt jelenségek valamelyike, akkor kerek zárójellel megjelöljük a betűhű mezőben – és lehetőség szerint a normalizált mezőben is – a megfelelő szórészt. Például:

Betűhű	**Normalizált**	**Megjegyzés**
uimagg(om)uc	imádju(n)k	ADD
sumha	soha	
nym	nem	
kyul		FAIL
hyul:	hűl.	
teun	tőn	
l		FRAG

A metaadatokkal ellátott vertikális fájlt XML-lé alakítjuk, így végezzük el a validációs lépéseket, melyek az adatbázis konzisztenciáját ellenőrzik. Egy követke-

ző átalakító lépés során alakul ki az alkalmas bemenet a korpuszkezelő rendszer számára.

9. A korpuszlekérdező eszköz

A korpusszal párhuzamosan készül a hozzá tartozó korpuszlekérdező felület, amelynek segítségével a teljes ómagyar korpuszt kutathatjuk. Ez jelenlegi állapotában az **Emdros** (Petersen 2004) korpuszkezelő rendszerre épül. A korpusztalálatok megjelenítése független a lekérdezéstől, abban az értelemben, hogy igény szerint bármilyen – a lekérdezésben esetleg nem is szereplő – szövegfeldolgozottsági szintet is megjeleníthetünk. Ezenfelül lehetővé tesszük a több szintre való egyidejű hivatkozást akár egy kérdésen belül is. Ha például az a kérdésünk, hogy milyen szavak szerepelnek egy igealak és egy igekötő között, akkor az elemzések szintjén (6) kell megfogalmazni a kérdést. Ha gyakorisági listát készítünk a korpusz egy részéből, akkor ezt megtehetjük például a szótövekből kiindulva, de rá lehet kérdezni közvetlenül az *nç* végű szavakra is, ekkor a (3) szinthez fordulunk (vö. 4. táblázat).

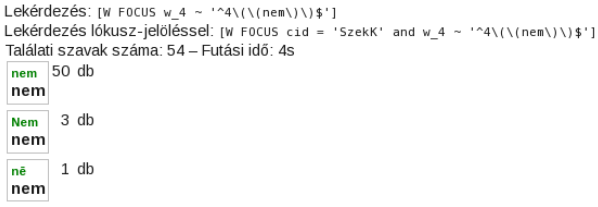
3. ábra. A korpuszlekérdező felület. A példában azokra a szavakra keresünk, melyeknél a normalizált alak kezdete a *jonh* sztring

A lekérdező felület a 3. ábrán látható. A felület középső részén adhatjuk meg a lekérdezést, melyben hivatkozhatunk az egyes szövegfeldolgozottsági szintekre, akár többre is egyszerre. Az itt megadott adatokból az *OK* gomb megnyomásával áll elő maga a lekérdezés a bal oldali szövegmezőben az Emdros lekérdezőnyelvén. Ez utóbbi még utószerkeszthető, és a *Mehet* gombbal futtatható.

A 3. ábrán bemutatott lekérdezés eredménye a 4. ábrán látható. A találatok felett a lókuszjelölő található, mely a kódex azonosítójából, az oldalszámból és az adott szó egyedi azonosítójából áll. Az egyes találatokat táblázatos formában jelenítjük meg: fent a betűhű alakot (a felületen zölddel), alatta a normalizált alakot (feketével), majd az értelmezést (kékkel). A felületen (jobb oldalt) a konkordan-

4. ábra. A 3. ábrán látható lekérdezés eredményének részlete: korpuszpozíciók, ahol a normalizált alak kezdete a *jonh* sztring, konkordancia formájában megjelenítve

cia mellett alternatív megjelenítési formátumként a gyakorisági lista is beállítható. Az 5. ábrán erre látunk egy példát: a Székelyudvarhelyi kódexben kerestünk rá a *nem* normalizált alakra, és az eredményben a betűhű alakot is megjelenítettük. Láthatjuk, hogy ezen a kódexen belül szinte egységes (és a maival egyező) ennek a szónak a helyesírása, de egy esetben azért előfordul a nazalitást makronnal jelölő régies forma (*nē*) is.



5. ábra. Példa a gyakorisági listás megjelenítésre

Végül lássunk három, az ómagyar szintaxisra vonatkozó elméleti nyelvészeti kutatási kérdést, melynek megválaszolásához segítséget nyújthat a korpusz. Mindhárom esetben a (6)-os szintre vonatkozik a lekérdezés, amely a szótövet és a morfológiai elemzést tartalmazza – ennek használatával lehet a nyelvészetileg leginkább releváns kérdéseket feltenni.

A mai magyarban tagadás esetén az igekötő követi az igét *(nem jön be),* az ómagyar viszont az igekötő + tagadószó + ige *(be nem jön)* sorrendet használja legtöbbször (É. Kiss 2010). Ezt a jelenséget mutatja az alábbi példamondat is:

JókK 69. o.: *Ver touaba kÿ nem futott* [Vér továbbá ki nem futott]. A szófajok sorozatára vonatkozó megfelelő lekérdezés a mai magyar szórendre:

```
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
[W FOCUS w_6e ~ 'Vpfx']
```

A lekérdezés az ómagyar szórendre:

```
[W FOCUS w_6e ~ 'Vpfx']
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
```

A `w_6e` jellemzővel a (6) szinten elérhető morfológiai elemzésre kérdezhetünk rá, a tagadószó kódja `Mod`, az ige kódja `V`, az igekötőé pedig `Vpfx`.

Az ómagyarban a mai magyartól eltérő a névelőhasználat: sok helyen nem használnak névelőt, ahol ma igen (Egedi 2010). Hogyan tudunk alátámasztani egy effajta hipotézist korpusz segítségével, azaz hogyan tudunk rákeresni arra, ami nincs ott? A megoldás az lehet, hogy két olyan szó kombinációjára keresünk rá, melyek között mai intuícióval várnánk a névelőt, de az ómagyarban a két szó névelő nélkül közvetlenül követi egymást. Ilyen konkrét helyzet lehet, amikor definit ige után tárgyesetű főnév áll, mint például ebben a mondatban: JókK 140. o.: *Es azert ewkewztewk zent ferencz czudalatost gÿczerÿuala teremtewtt* [És azért őköztük Szent Ferenc csodálatost dicséri vala Teremtőt]. Ilyen esetekre a megfelelő lekérdezés:

```
[W FOCUS w_6e ~ 'V.*Def']
[W FOCUS w_6e ~ 'N.*Acc']
```

A használt morfológiai kódok: ige: `V`; határozottság: `Def`; főnév: `N`; tárgyeset: `Acc`.

A harmadik kutatási kérdés a *se*-névmások tulajdonságairól szól. Míg a mai magyarban a tagadószó hordozza a tagadást, a *se*-névmások pedig csupán a tagadószóval egyeztetett alakok, a korai ómagyar korban a *se*-névmásoknak önmagukban is lehetett tagadó erejük (É. Kiss 2010). Ha a *senki/semmi* után közvetlenül egy tagadószótól különböző szót találunk a korpuszban, akkor jó eséllyel erre a jelenségre találtunk példát. Az alábbi szövegrészlet éppen ilyen: JókK 8. o.: *mendenestewlfoguan maganac semÿtt meg tarttuan* [Mindenestül fogván magának semmit megtartván]. Ebben az esetben a lekérdezés a következőképpen néz ki:

```
[W FOCUS w_6s ~ '^6s\(\(se[nm][km]i\)\)$']
[W FOCUS NOT(w_6e ~ '^6e\(\(Mod\)\)$')]
```

A *Régi Magyar Konkordancia* nevet viselő lekérdezőfelület szabadon elérhető a http://rmk.nytud.hu címen.

10. További feladatok

Elsődleges feladatunk a teljes ómagyar anyag betűhű szöveges formában való előállítása és kereshetővé tétele. A normalizálást, valamint a morfológiai elemzést és egyértelműsítést csak a korpusz egy részén fogjuk végrehajtani.

Az ómagyar szövegek eleve adott heterogenitása mellett további problémákat okoz az is, hogy a különböző korokban kiadott nyomtatott kódexátiratok tipográfiai kényszerűségek miatt azonos karaktereket eltérően jelenítenek meg. Terveink között szerepel ezen esetlegességek kiküszöbölése, vagyis a különbözőképpen jelölt karakterek azonos sztenderd Unicode-karakterrel való lecserélése.

A projekt vállalásai közé tartozik, hogy a korpusz arányos válogatást tartalmazzon a középmagyar kor (1526–1772) szövegeiből is. Ezen anyagok esetében már fontos szerepet játszik a reprezentativitás kérdése, ugyanis ebből a korból lényegesen több nyelvemlékünk származik, vagyis a teljes anyag feldolgozására ebben a projektben nem vállalkozhatunk. A középmagyar szövegemlékek kiválogatásánál két fő szempontot tartunk szem előtt: csak a már szöveges formátumban elérhető dokumentumokkal foglalkozunk, és ezeket Dömötör (2006) műfaji beosztását követve kategorizáljuk úgy, hogy minden regiszter képviselve legyen a korpuszban.

11. Összegzés

A nyelvi kulturális örökség feldolgozhatóvá és elérhetővé tételében kulcsfontosságú szerep jut a nyelvtechnológiának, amely (fél)automatikus módszereivel hozzásegíti a humán tudományok kutatóit olyan adatbázisokhoz, melyekben a nyelvészeti (és/vagy történeti, paleográfiai stb.) információk elektronikusan előhívható és interpretálható módon vannak tárolva. Az ilyen korpuszok sokkal kifinomultabb keresési lehetőségeket kínálnak, mint az egyszerű digitalizálás, amely általában kimerül a primér adat képként való beszkennelésében. A nyelvtechnológiai eszközökkel feldolgozott történeti szövegeknek további előnyei közé tartozik, hogy a kutatók egységes, akár egy egész korra jellemző, átfogó keresési eredményhez jutnak, amellyel elméleti feltevéseik könnyebben igazolhatóvá válnak.

A nyelvi kulturális örökség feldolgozása a nyelvtechnológusok elé számos kihívást állít. Az elektronikus formátumok előtti korból származó szövegek esetében az eddigieknél robusztusabb vagy teljesen új módszerekre van szükség. Vagyis a kulturális örökség digitalizálása során nemcsak a már bevált módszerek új területeken való alkalmazása történik, hanem az új módszerek új kutatási kérdéseket is felvetnek. Ezek megoldásához a különböző tudományterületek kép-

viselői közötti szoros együttműködésre van szükség, amelyből meggyőződésünk szerint hosszú távon minden résztvevő profitálhat.

Irodalom

Adamikné Jászó Anna 1992. Az igenevek. In: Benkő Loránd – E. Abaffy Erzsébet (szerk.): A magyar nyelv történeti nyelvtana. 2/1. kötet: A kései ómagyar kor: morfematika. Budapest: Akadémiai Kiadó. 319–352.

Dömötör Adrienne 2006. Régi magyar nyelvemlékek. Budapest: Akadémiai Kiadó.

É. Kiss Katalin 2010. A tagadás. Előadás a Mondattani jelenségek a Jókai-kódexben műhelykonferencián.

Egedi Barbara 2010. A határozott névelő. Előadás a Mondattani jelenségek a Jókai-kódexben műhelykonferencián.

Jakab László 2002. A Jókai-kódex mint nyelvi emlék szótárszerű feldolgozásban (Számítógépes Nyelvtörténeti Adattár 10). Debrecen: Debreceni Egyetem Magyar Nyelvtudományi Tanszék.

Jakab László – Kiss Antal 1994. A Guary-kódex ábécérendes adattára (Számítógépes Nyelvtörténeti Adattár 6). Debrecen: KLTE Magyar Nyelvtudományi Tanszék.

Jakab László – Kiss Antal 1997. Az Apor-kódex ábécérendes adattára (Számítógépes Nyelvtörténeti Adattár 7). Debrecen: KLTE.

Jakab László – Kiss Antal 2001. A Festetics-kódex ábécérendes adattára (Számítógépes Nyelvtörténeti Adattár 9). Debrecen: Debreceni Egyetem.

Kiss, Gabriella – Júlia Pajzs 2001. An attempt to develop a lemmatiser for the Historical Corpus of Hungarian. In: Proceedings of CL 2001. University of Lancaster. 443–451

Kniezsa István 1952. Helyesírásunk története a könyvnyomtatás koráig. Budapest: Akadémiai Kiadó.

Korompay Klára 2003. Helyesírás-történet [az ómagyar korban]. In: Kiss Jenő – Pusztai Ferenc (szerk.): Magyar nyelvtörténet. Budapest: Osiris Kiadó. 281–300.

Kroch, Anthony – Ann Taylor 2000. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania, second edition. CD-ROM. (http://www.ling.upenn.edu/hist-corpora/)

Kunstmann, Pierre – Achim Stein 2007. Le Nouveau Corpus d'Amsterdam. In: Pierre Kunstmann – Achim Stein (szerk.): Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23–26 février 2006, 9–27. Stuttgart: Steiner.

McEnery, Tony – Andrew Hardie 2003. Lancester Newsbooks Corpus. (http://www.lancs.ac.uk/fass/projects/newsbooks/default.htm)

Oravecz Csaba – Sass Bálint – Simon Eszter 2009. Gépi tanulási módszerek ómagyar kori szövegek normalizálására. In: Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): A VI. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 317–324

Oravecz, Csaba – Bálint Sass – Eszter Simon 2010. Semi-automatic normalization of Old Hungarian codices. In: Proceedings of the ECAI 2010 workshop on language technology for cultural heritage, social sciences, and humanities (LaTeCH 2010). Lisbon, Portugal: Faculty of Science, University of Lisbon.

Petersen, Ulrik 2004. Emdros – a text database engine for analyzed or annotated text. In: Coling 2004, 1190–1193.

Prószéky, Gábor – Balázs Kis 1999. A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. College Park, Maryland. 261–268.

Rayson, Paul – Dawn Archer – Alistair Baron – Jonathan Culpeper – Nicholas Smith 2007. Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In: Proceedings of corpus linguistics. University of Birmingham.

Shannon, Claude Elwood 1948. A mathematical theory of communication. Bell System Technical Journal 27: 379–423.

de Sousa, Maria Clara Paixao – Thorsten Trippel 2006. Building a historical corpus for classical Portuguese: Some technological aspects. In: Proceedings of the Vth International Conference on Language Resources and Evaluation (LREC 2006). Genova: ELRA.

Thomas, Peter Wynn – D. Mark Smith – Diana Luft 2007. Rhyddiaith Gymraeg 1350–1425. (http://www.rhyddiaithganoloesol.caerdydd.ac.uk)

Váradi, Tamás 2002. The Hungarian National Corpus. In: Mark T. Maybury (szerk.): Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas: European Language Resource Association. 385–389.

Volf György 1874. Nyelvemléktár I. Budapest: A Magyar Tudományos Akadémia Könyvkiadó Hivatala.

Volk, Martin – Torsten Marek – Rico Sennrich 2010. Reducing OCR errors by combining two OCR systems. In: ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010). Lisbon, Portugal: Faculty of Science, University of Lisbon.

Human language technology and cultural heritage: corpus building from Old Hungarian codices

Abstract: Human language technology plays a key role in digitisation of our cultural heritage. With the help of robust computational methods researchers can clean, enrich, search and mine digitised data. A pivotal area of the cooperation of historical and computational linguists is corpus building from historical language data, which can serve as a base of theoretical research. This paper presents the whole workflow of corpus building from Old Hungarian codices.

Keywords: digital humanities, human language technology, historical corpus, normalisation, corpus building

Az ember–gép kommunikáció elméleti–technológiai modellje és nyelvtechnológiai vonatkozásai

Hunyadi László – Földesi András – Szekrényes István – Staudt Alexandra – Kiss Hermina – Abuczki Ágnes – Bódog Alexa

Debreceni Egyetem, Általános és Alkalmazott Nyelvészeti Tanszék, Debrecen
hunyadi@ling.arts.unideb.hu; andras.foldesi1@gmail.com; xepenator@gmail.com; crysaetos@gmail.com; kissh3@gmail.com; abuczki.agnes@gmail.com; alexab@unideb.hu

A tanulmány a nyelvészet, a kommunikációkutatás, a pszichológia, az informatika és a kognitív robotika találkozásánál elhelyezhető kutatás első eredményeiről kíván számot adni. A kutatás komplexitását a feladatmeghatározás indokolja: szeretnénk megérteni a dialógusokban megjelenő ember–ember kommunikáció alapszerkezetének azon aspektusait, amelyek relevanciával bírnak az ember–gép interakcióban és amelyekről úgy gondolhatjuk, hogy technikailag megvalósíthatók. Ezen fő szempontok vezérlik egy elméleti–technológiai modell fő vonalainak a meghatározását éppúgy, mint egy multimodális korpusz létrehozását. A tanulmányban először e modellt mutatjuk be, majd a korpusz fő moduljait ismertetjük és elemezzük a korpusz alapján létrehozott HuComTech multimodális adatbázis egyes, az ember–ember kommunikációra vonatkozó adatait. Az eredendően nyelvészek által kezdeményezett kutatás egyben láttatni engedi a nyelvtechnológia multimodális kiterjesztési lehetőségeit is.

Kulcsszavak: ember–gép kommunikáció, multimodalitás, korpuszépítés, annotáció, prozódia, szintaxis, pragmatika

1. Bevezetés

Évtizedekkel ezelőtt, amikor a számítógép először megjelent a környezetünkben, ez a környezet szakemberek (elsősorban matematikusok, egyéb természettudósok, csak később informatikusok) szűk csoportjára szorítkozott, azokéra, akik „értették a számítógép nyelvét". Eme exkluzivitás következtében a „kívülállók" a számítógépet valami megközelíthetetlennek, misztikusnak és olyannak tekintették, ami az ő világuktól merőben különböző, és szinte adottnak tekintették, hogy annak (meg)léte nem befolyásolja saját hétköznapjaikat. Pedig ha jobban belegondolunk, a számítógép is az ember teremtménye, így az embertől nehezen különíthető el, ráadásul olyan feladatokat bízunk rá, amelyek ilyen vagy olyan

mértékben, de mégis az ember szolgálatát jelentik. Egy hosszú evolúció eredményeként ezek a feladatok egyre inkább hétköznapi életünkhöz kezdtek közeledni, azaz egyre határozottabban érzékelhettük, hogy használatával „rólunk van szó" (egy levél megszerkesztésétől adóbevallásunkig vagy egy vonatjegy megvásárlásáig). Mivel most már nem csupán egy szűk kört érintő feladatról van szó (mint amilyen például egy termelési folyamat irányítása vagy az élve születések száma trendjének kiszámítása), mi, a hétköznapi felhasználók valóban úgy kívánunk tekinteni a számítógépre, mint ami emberi létezésünk része. Azaz – ha már bennünket szolgál – legyen a gép „gondolkodásában" is olyan, mint amilyenek mi vagyunk.

Nos, ez az a terület, ahol a technológiai fejlődés messze előtte jár a felhasználó hétköznapi élményeinek. Azt várjuk, hogy a gép szinte előre kiszámítsa vagy kövesse gondolatainkat, szándékainkat, ugyanúgy, ahogy ezt egy humán partnertől is elvárjuk. Természetesen azt is tudjuk, hogy egy humán partnerhez is alkalmazkodni kell, úgy, hogy lépéseinket a másik fél lépései (netalán megfejtett szándékai) is befolyásolják, azaz miközben elvárjuk, hogy a géppel való interakció számunkra könnyebb, emberközelibb legyen, ezen interakciónak mégis szigorú szabályokon kell alapulnia. Ezektől a szabályok azonban elvárhatjuk, hogy nekünk szóljanak, és így legfőképp az emberi kommunikáció szabályain alapuljanak.

De vajon milyenek az emberi kommunikáció szabályai? Vajon ismerjük őket? Úgy gondolhatjuk: persze hogy ismerjük, hiszen nélkülük nem tudnánk embertársainkkal sikeresen kommunikálni. De vajon meg tudjuk-e fogalmazni e szabályokat úgy, hogy a gép is megértse őket és így szinte természetes partnerünk lehessen? A kérdés nem is olyan egyszerű, miközben a megálmodott feladatok ezt kívánnák: azt, hogy például egy robot szobában elrejtett szemeivel-füleivel segítsen egy magányos arra rászorulót, vagy azt, hogy egy repülőtér forgatagában bennünket fáradhatatlanul és mindig ugyanolyan megbízhatósággal egészítsen ki egy keresett személy megtalálásában. Kövesse a hangunkat? A tekintetünket? A mozdulatainkat? Válassza ki az elérhető számtalan jel közül a relevánsakat, értse meg ezek összefüggéseit és ezekre alapozva jusson el egy döntés pillanatáig? Bizony, ezek az igények messze meghaladják azt a feladatot, amit egy ipari gép egyszerű vezérlő gombokkal történő irányítása jelent. Az ember hangsúlyozott részvételét az ember–gép interakcióban azon feltétel teljesülése mellett várhatjuk, ha mi, emberek megismerjük az ember–ember kommunikáció alapvető rendjét és ezeket a technológia számára is fogyasztható formában szabályokba foglaljuk, majd implementáljuk az ember–gép kommunikáció valamely technológiai alkalmazásában.

Az itt következőkben egy olyan projekt eredményeiről és jövőbeni terveiről számolunk be, amelynek a középpontjában az ember–gép kommunikáció technológiájának a továbbfejlesztése áll. E cél érdekében arra törekszünk, hogy jobban megismerjük az ember–ember kommunikáció alapvetőnek tekintett, ugyanakkor technológiai szempontból is releváns tulajdonságait, továbbá, hogy mindezen ismereteket olyan rendszerben reprezentáljuk, ami a technológia számára is elérhető lehet.[1] A projekt megtervezésekor kézenfekvőnek tűnt, hogy munkánk középpontjában a nyelv vizsgálata, ezen belül annak nyelvtechnológiai megközelítése álljon. Így céljaink között szerepelt, hogy a gyakorlatban akár közvetlenül is alkalmazható módon leírjuk és vizsgáljuk a spontán beszéd bizonyos akusztikai fonetikai tulajdonságait (F0, intenzitás, tempó), hogy a prozódia rendszeres leíásával hozzájáruljunk az automatikus beszédfelismerés és -szintézis tökéletesítéséhez. Ez utóbbihoz arra is szükség volt, hogy egy jelentős méretű korpuszon vizsgáljuk és gépi tanítás számára elérhetővé tegyük a spontán beszélt nyelvi szintaxis és a prozódia interfészét. Úgy gondoljuk, és első eredményeink is azt mutatják, hogy ez a spontán beszédre irányuló megközelítés a nyelvelméleten túl a nyelvtechnológia számára is számos új lehetőséget rejt magában. Bár projektünkben nem térünk ki olyan, jól megalapozott nyelvtechnológiai vizsgálatokra, mint az automatikus morfológiai és szintaktikai elemzés, korpuszunk multimodális jellege számos olyan izgalmas kérdés feltevésére ad lehetőséget, amelyek egyrészt kiterjeszthetik a nyelvtechnológia fogalmi és cselekvési körét, másrészt utat nyitnak további interdiszciplináris kutatásokhoz. Így a nem verbális gesztusok annotálása lehetővé teszi, hogy további, komplex adatokat szolgáltassunk a kommunikáció pszichofizológiai és kognitív vizsgálatához, különösen a spontán beszédben oly gyakori megszakadt vagy megszakított szintaktikai szegmensek és a gesztusok együtt járásának feltérképezésén keresztül. Ezt szolgálja a korpusz dialógusainak pragmatikai és funkcionális feldolgozása is, amelynek során a nyelvi, verbális elemeket szoros formális és funkcionális összefüggésben vizsgáljuk a nem verbális elemekkel. Az ilyen vizsgálatokra egy tágabb értelmű nyelvtechnológiának is nagy szüksége van, hiszen például egy avatárt nem elég „beszéltetni", hanem az adott kommunikatív helyzetben harmonizálni kell annak verbális és nem verbális viselkedését egyaránt.

Megközelítésünkben tehát a nyelvtechnológia tulajdonképpeni tárgya, a nyelv nem választható el a beszéd során megjelenő, azzal összefüggő egyéb, nem verbális jelektől, ami által a nyelvtechnológia is kiterjesztett értelmezést kap.

[1] A HuComTech munkacsoport munkája a TÁMOP 4.2.1 2008/9 projekt keretében indult és jelenleg a TÁMOP-4.2.1/B-09/1/KONV-2010-0007 projekt támogatásával zajlik. Egyes eredményei elérhetők a honlapján is: http://hucomtech.unideb.hu/hucomtech.

A jelen dolgozatban a HuComTech-team sokszálú tevékenységei közül a 2. pontban felvázoljuk azt az elméleti–technológiai modellt, amelynek kidolgozása mind eredménye, mind további alapja az ember–ember kommunikáció technológiai célú vizsgálatának. Ezt követően röviden bemutatjuk a létrehozott korpuszt, majd végigvezetjük az olvasót a különböző annotálási szinteken, bemutatva bizonyos kezdeti eredményeinket: a 3. pontban a vizuális szint, a 4. pontban az akusztikai szint (szöveg és prozódia) annotálását, az 5. pontban bemutatjuk egy, a nem szándékolt ismétlésekre irányuló vizsgálat multimodális adatokra támaszkodó eredményeit, a 6. pontban ismertetjük a spontán beszéd szintaxisának annotálását célul kitűző újszerű rendszert, majd a 7. és a 8. pontban, ugyancsak újdonságként, bemutatjuk egy tetszőleges kommunikációs esemény pragmatikai vonatkozásainak annotálását unimodális és multimodális megközelítésben.

2. A kommunikáció egy elméleti–technológiai modellje

A bevezető gondolatok szellemét követve egy technológiailag implementálható modellnek az ember–ember kommunikáció elméletén kell alapulnia. Olyan elméleten, amely kiindulásként figyelembe veszi a technológia, a gépi irányítás alapvető megkötöttségeit (ezáltal redukálva a fenti bevezető alapján szabadon ereszthető kívánságlistánkat) és ilyen megkötöttségek között keresi a humán kommunikáció szabályszerűségeit és írja le annak lehetséges működését.

A kommunikáció alapvető komplexitása megkívánja, hogy egy ilyen modell, miközben kellő általánossággal bír, figyelembe vegye az ember–gép interakció adott szándékolt alkalmazási területét. Egy ilyen kézenfekvő terület a humán–gépi interfészeké. Az ember–gép kommunikáció elméleti kutatásának a jelentősége is elsődlegesen a gépi interfészek kutatásában emelkedik ki (a teljesség igénye nélkül vö.: Wahlster 1991; Dix et al. 2003; Oviatt 2003). Ezek között Wahlster (1991) a kommunikáció komplex, dialógusalapú felhasználói modelljét írja le (XTRA rendszer), melyben – egyebek között – kimutatható a természetes és a gépi deiktikus jelek közötti különbség és igen tanulságos a gépi jelek használhatóságának kísérleti értékelése. Az interfészkutatáson belül külön hangsúlyt kap az interakció multimodalitása (vö. Flanagan 1997; Mariani 1997; Bernsen–Dybkjær 2005; Kuppevelt et al. 2005). A multimodális kommunikáció modelljeinek egy széles körű áttekintését adja Wahlster (2006), tanúsítva a modellek megfeleltetését az alkalmazások speciális céljainak. Az általunk is választott modellhez általános szemléletében Thórisson (2008) modellje áll közel, aki olyan absztrakt modulhierarchiákat tételez fel, amelyek kifejezetten a technológia (ezen belül a robotika) szempontjait veszik figyelembe.

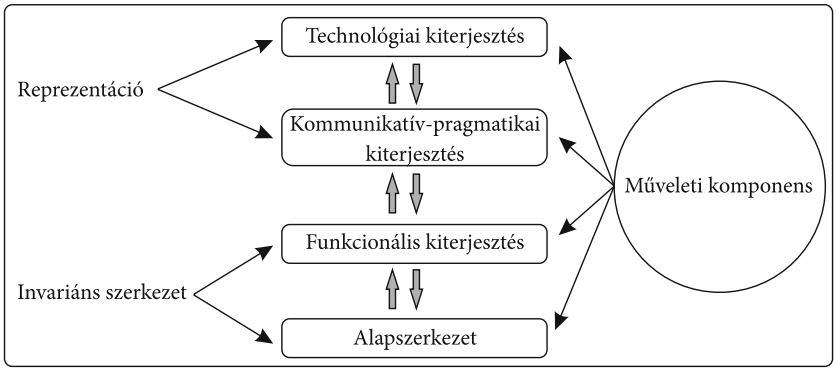
Az itt javasolt modell messzemenően figyelembe kívánja venni azokat a szempontokat (valójában kihívásokat), amelyeket a technológia állít egy, az ember–ember kommunikáció jellegét megközelíteni szándékozó alkalmazás elé. Így kiindulásként lényeges figyelembe vennünk, hogy a technológiai folyamatok lényegüket tekintve szigorúan szekvenciálisak. Ez azt jelenti, hogy egy esemény létrejöttének (a technológiai szabályozás egy bizonyos állapotának) mindig vannak előfeltételei és ezt követő következményei, úgy, hogy az ilyen szabályozás mindig egyirányú. Lényeges továbbá, hogy a szabályozás során mindig valamely egyetlen paraméter diszkrét értékének a beállítására kerül sor. Bár az lehetséges, hogy ennek során egynél több paraméter értékének a beállítása történjék meg egy időben, ez nem változtat azon, hogy a szabályozásnál egymástól jól elkülöníthető paraméterek értékei egyenkénti beállításáról van szó. Ilyen ételemben a szabályozás **moduláris.** Ezzel szemben a humán kommunikáció egyik fő jellegzetessége, hogy az egy időben bejövő jelek sokaságát (például köszönéskor az üdvözlő szavakat kéz- és testmozdulat, tekintet és fejtartás is kíséri, amelyek közül egyesek opcionálisak lehetnek) egyazon időben dolgozzuk fel, így a kommunikációt **holisztikusnak** érzékeljük. Ezért az elméleti–technológiai modellnek egyszerre kell egyrészt szekvenciálisnak és modulárisnak, másrészt holisztikusnak lennie. Ezen látszólagos ellentmondást azzal oldhatjuk fel, ha a modellünk szekvenciálisan moduláris, egyben multimodális lesz.

A humán kommunikáció főbb modelljei természetes módon nem a technológia követelményeire összpontosítanak, így eredményeik csak részben elérhetőek számunkra.[2] A **kódmodellek** (Shannon–Weaver 1949; Jakobson 1969) magára a kódhasználatra korlátozódnak, azaz arra, hogy a két információfeldolgozó eszköz közötti sikeres kommunikáció feltétele az azonos kód ismerete és a reprezentáció azonossága. E megközelítésben a kontextus nem játszik szerepet. A következtetéses modellek (Grice 1957; 1975; Lewis 1969) szerint a kommunikáció akkor sikeres, ha a kommunikációs partner meg tudja fejteni a beszélő által konkrét szituációkban, azaz kontextusokban létrehozott megnyilatkozások jelentését, azonban nem írják le kellően azt a folyamatot, amelynek során a kódhasználat beágyazódik magába a kontextusba. Az osztenzív-következtetéses modellek (Sperber–Wilson 1986/1995) már kezelik mind a kódhasználatot, mind a kódhasználat nélküli osztenziót és következtetést, azonban – hasonlóan az előző modellekhez – nem támaszkodnak a kommunikáció multimodalitására és szekvencialitására.

[2] Az itt említendő főbb modellek részletes vizsgálatára és az itt következő sorokban is felhasznált jellemzésére, továbbá a jelen tanulmányban javasolt elméleti-technológiai modellel való összevetésére l. Németh T. (2011a).

A javasolt modell alapját a generatív grammatika moduláris szemlélete képezi (vö. Chomsky 1965; 1977; 1986). Felteszi, hogy a kommunikáció meghatározó aspektusai megragadhatóak egymásra épülő modulokban zajló folyamatok leírásán keresztül. Ugyancsak felteszi, hogy az egyes modulok alapján egységesen ún. primitívek, tovább már nem bontható elemi részekből állnak és egy általános, a modulok fölött álló műveleti komponens ezen primitívekből összetett, immáron nem primitív szerkezeteket hoz létre. A műveleteket rekurzív módon alkalmazva elvileg végtelen összetettségű kommunikatív szerkezetek állhatnak elő. A legalsó szinten lévő modul, az invariáns Alapszerkezet a kommunikáció „szintaxisát" állítja elő, az erre épülő, ugyancsak invariáns Funkcionális kiterjesztés az Alapszerkezet által lehetségesnek (jól formáltnak, „grammatikusnak") tekintett formális szerkezetekhez ugyancsak lehetséges funkciókat rendel (azaz a „szintaxist" ellátja bizonyos mértékű „szemantikával"), végül a Kommunikatív-pragmatikai kiterjesztés képezi a reprezentáció szintjét, azt a szintet, ahol az immáron formailag és szűken vett értelemben funkcionálisan lehetséges (jól formált) absztrakt kommunikatív szerkezetek felszíni, azaz az adott kontextusban pragmatikailag aktualizált (az adott kommunikatív aktusra érvényesített) reprezentációt kapnak. Ezen, a modell elméleti komponense értelmében felszíni reprezentációnak valóságos, tárgyiasult felszíni reprezentációt a modell technológiai komponense, a Technológiai kiterjesztés ad, ahol a modell elméleti komponensének kimenete technológiai megfogalmazást nyer. Így tehát a Kommunikatív-pragmatikai kiterjesztés mintegy interfész-szerepet tölt be a modell elméleti és technológiai aspektusai között.

A modell sematikus felépítését mutatja az 1. ábra.



1. ábra. Az ember–gép kommunikáció generatív technológiai–elméleti modelljének sémája

A modell modularitásából és szekvencialitásából következik, hogy megvan annak az elvi lehetősége, hogy egyszerre szolgálja a szintézist (egy kommunikatív aktus létrehozását) és az analízist (egy észlelt konkrét aktus – funkcionális és pragmatikai – értelmezését), amit az ábrán a két irányú nyilak is sugallnak. Így a modell egységes elméleti–technológiai keretet biztosít arra, hogy a kommunikációban egy időben jelen lévő két, ellenkező irányú funkciót (a résztvevők főként váltakozó aktív és passzív szerepét – mint pl. a beszélő és a hallgató között) egyszerre ragadja meg és kezelje.

Néhány példa a modell egyes moduljainak alapjául szolgáló primitívekre:

Alapszerkezet: a kommunikáció kezdete, vége, fenntartása, időleges felfüggesztése és újrakezdése

Funkcionális kiterjesztés:

- strukturális funkcionális primitívek: a kezdés módja, a befejezés módja, a felfüggesztés módja, az újrakezdés módja,
- logikai funkcionális primitívek közé tartozik az állítás, tagadás, kérdés, kondicionális, kvantifikáció,
- holisztikus funkcionális primitívek: mellérendeltség, alárendeltség, fölérendeltség, a részvétel létrehozása és fenntartása, beszélőváltás, folyamatosság, érzelmek és szándékok

Kommunikatív-pragmatikai reprezentáció:

- *nem verbális*: vizuális, auditoros, vagy egyéb érzékeléssel, pl. érintéssel vagy szaglással közvetített formák, mint elemi mozdulatok, test- és arckifejezések
- *verbális*: szintaktikai (pl. állítás, kérdés, felkiáltás, óhajtás stb. kifejezése, valamilyen logikai funkció kifejezése, mint feltétel, kvantifikáció, negáció), lexikális (pl. szcenáriótól függő kifejezések), fonetikai–fonológiai (intonáció, beszédtempó, szünet)

Technológiai kiterjesztés:

- *interfész a modell elméleti oldala felé – markerek*: az Alapszerkezet, a Funkcionális kiterjesztés és a Kommunikatív-pragmatikai kiterjesztés primitívjeinek felszínen megjelenő, fizikailag mérhető megvalósulásai
- *interfész a modell technológai oldala felé – paraméterek*: a markerek technológiafüggő megvalósításának diszkrét bemeneti adatokat váró alapeszközei

A modulokra egységesen érvényes műveleti komponens olyan műveleteket tartalmaz, amelyekkel elvileg végtelen összetettségű kommunikatív szerkezetek állhatnak elő. Ilyen műveletek lehetnek: konkatenáció, iteráció, beágyazás, kiágyazás, közbeékelés, megszakítás, kombináció, és mindezek rekurzív alkalmazása.

A fentiekből kitűnik, hogy ahhoz, hogy egy kommunikatív eseményt akár szintetizálhassunk, akár analizálhassunk, minimálisan az szükséges, hogy az

adott esemény felszínen megjelenő összetevő szerkezeti elemeit, a markereket azonosítsuk. Ahhoz, hogy – a kommunikáció multimodalitásából kiindulva – ezen markerek együttállása alapján valamiféle kommunikatív értelmezést, valamint a technológia számára paraméterekben megfogalmazható vezérlést is adjunk, szükséges ezen markerek strukturális viszonyainak a meghatározása mind az eseménytípus, mind az aktuális esemény számára. Így az ember–ember kommunikáció tanulmányozása során az adatainkat egy olyan adatbázisba kell rendeznünk, amely alkalmas mind a kommunikáció általános jellegzetességeinek a meghatározására, mind az egyes konkrét esemény valamely konkrét mozzanatának az értékelésére. A HuComTech korpusz alapján létrejött adatbázis ezt a célt szolgálja. Az alábbiakban az ezen adatbázisból nyert adatok alapján rámutatunk az ember–ember kommunikáció során megfigyelhető bizonyos összefüggésekre. Ezen összefüggések feltárása vezethet el a bemutatott elméleti–technológiai modell technológiai alkalmazásához.

A HuComTech korpuszt egy 112 beszélő (egyetemi hallgatók) részvételével készült, összességében 50 órányi, beszélőnként fél-fél óra hosszú audió- és videófelvétel alkotja, amiből felvételenként kb. 5–5 perc felolvasás, 25–25 perc dialógus. A dialógusok során két személy spontán párbeszédét rögzítettük, egy formális és egy informális társalgási szcenárió keretei között. Az első (formális) dialógus egy szimulált állásinterjú, a második (informális) pedig egy, különböző témákat körbejáró irányított beszélgetés formájában valósult meg. A korpusz számítógépes feldolgozhatóságát a felvételekhez készült annotációk biztosítják, amelyek többféle megközelítésben (vizuális jelek, nyelvi egységek és kommunikációs események megfigyelése) címkézik fel a korpuszban vizsgált jelenségeket. Magában a korpuszban az egyes címkéket időjelekkel *(timestamp)* láttuk el, ami lehetővé teszi a kommunikáció számos jelenségének multimodális vizsgálatát és e jelenségek technológiai célú leírását. Az annotálást főként manuálisan végeztük úgy, hogy minden, előre megfogalmazott irányvonal mentén történt annotálás független ellenőrzésen ment át. Az adatbázis-lekérdezést felhasználtuk magának az annotálásnak az utólagos ellenőrzésére is. Az annotálás vonatkozott mind fizikai jellegű, de egyszerű leírást igénylő (pl. a tekintet iránya vagy kézi gesztusok jellege), mind interpretációt feltételező adatok (pl. érzelmi kifejezések) kinyerésére. Elkezdődött a korpusz automatikus annotálása is olyan területeken, ahol jól definiálható és igen pontos fizikai adatok kinyerésére van szükség (a beszédprozódia területén, elsősorban különböző felismerő algoritmusok számára).

3. A korpusz videoannotálásának elvei és egyes tapasztalatai

A fizikai jellegű, egyszerű leírást igénylő adatok (mint a tekintet iránya vagy kézi gesztusok jellege), valamint az interpretációt feltételező adatok (pl. érzelmi kifejezések) fentebb említett kinyerésére többfajta annotációt alkalmazunk.

A multimodális korpusz annotálása során kiindulásként kézenfekvő volt a videó- és az audiócsatornákat egymástól elválasztva kezelni. Ennek elsődleges oka az, hogy egy technológiai alkalmazásnál a kétféle csatornából érkező jeleket természetük különbözősége folytán különböző eszközökkel (szenzorokkal) tudjuk érzékelni, de az is, hogy egy szintézis során ugyancsak külön-külön kell létrehozni videó- és audiómintázatokat. A videó kézi annotálásánál egyrészt, mint mondtuk, diszkrét, fizikai természetű adatokat figyeltünk meg és annotáltunk, másrészt egyes mintázatokhoz interpretatív jegyeket rendeltünk. Az annotálás eme kettőssége (deskripció és interpretáció) előrevetítette az így kinyert értékek kettős felhasználását: ezen adatok megléte lehetőséget teremt a szintézis (fizikai elemek egyenkénti manipulálásával történő) technológiai megvalósítására és – ugyanezen jegyek együttállásának vizsgálatával – a kommunikatív esemény multimodális értelmezésére. (A videoannotáció ezen utóbbi mozzanata összetettebb szinten visszaköszön a multimodális pragmatikai annotáció esetében; vö. 8. pont.)

A projekt keretében fejlesztett Qannot annotáló program (Pápay et al. 2012) lehetővé teszi, hogy egy videót a lehető legkisebb részekre *(frame)* bontva megfigyeljük az esemény lelfolyásának videómozzanatait és az észlelt jelenségek időtartamát (kezdetét és végét), valamint annak milyenségét megfelelő címkével jelöljük. A különböző szempontoknak megfelelően az annotálás különböző szinteken történt.

A videoannotáció szintjei három csoportba sorolva (az angol nevek értelmezése utánuk zárójelben áll) az alábbiakban láthatók:

1) basic (a videófelvételt technikailag azonosító szintcsoport):

- comevent class (itt jelölendő a teljes kommunikációs eseménysor, az egész interjú kezdete és vége)

2) physical (a videoannotációban fizikai természetű jegyek alapján azonosítható szintcsoport):

- facial expression class (az arckifejezés szintje)
- gaze class (a tekintet iránya)
- eyebrows class (a szemöldök mozgása)

- headshift class (fejmozgás)
- handshape class (a kézfejek formája)
- touchmotion class (adott testrész érintése)
- posture class (testtartás)
- deictic class (utalás valamire kézzel)

3) functional (a videoannotációban funkcionális természetű jegyek alapján azonosítható szintcsoport; ezeken hallható az alany hangja is):

- emotion class (érzelemkifejezés)
- emblem class (ez az ún. emblémaszint az egyetértés és az egyet nem értés szintje)

Anélkül, hogy az annotálás folyamatát részletesebben ismertetnénk (erre l. Földesi 2011, 39), a következőkben néhány kiragadott példával érzékeltetni kívánjuk a videofelvételeken észlelhető modalitáskombinációk sokrétűségét, úgy, hogy párokat vagy hármasokat adunk meg, amelyeknek a tagjait egy arckifejezés-típus és egy tekintetirány, vagy az előbbi kettő és esetleg egy jellegzetes kéztartás képezik. Először az interjúkban elhangzott üzenetek információtartalmához való hozzáállás többsíkú (több modalitásban történő) kifejeződését vizsgáljuk, de tovább is óhajtunk lépni annyiban, hogy rátérünk a még nem kezelt, egy unimodális annotáció keretein belül is kérdéses mozdulatokra vagy mozdulatsorokra, címkével nem jelölhető arckifejezésekre. Tehát olyan, az adott megfigyelt személyre (interjúalanyra) jellemző, az egész formális vagy informális interjú alatt visszatérő mozdulatsorokról, arckifejezésekről stb. is lesz szó, amelyekhez nem rendelhető funkció.

3.1. A szegmentálás és következményei

A videoannotáló program segítségével a korpuszunkban rögzített interjúanyagot a benne fellelt kommunikációs események részmozzanataira daraboljuk fel. A feladat lényege a videofelvételek alatti, annotációs szinteket tartalmazó szalagon (egyenlő egységekre – *frame*-ekre – tagolt elemző felületen), időhatárok kijelölése után, minden határozottan észlelhető történés megcímkézése.

Az annotátorok munkájuk végeztével elérik, hogy ne csak maguk a videófelvételeken látott események és részmozzanataik legyenek láthatóak, illetve láthatóbbak, mint annotáció nélkül, hanem ezek időbeli kiterjedése is, ami a kész annotációkról leolvasható.

Vigyáznunk kell azonban, hogy mekkora szeletekre aprítunk fel egy mozdulatsort annak felcímkézéséhez. Mozdulatsoron az egymást követő azonos, hasonló vagy különböző mozgásokat értjük. Nincs okunk rá, hogy a mozdulatsor egyes részeit feltétlenül egymáshoz tartozónak vegyük.

Fontos megkülönböztetni egyszeri és periodikus mozgást is. Ugyanis előfordul, hogy a megfigyelt személy nem határozott, teljes mozdulatot tesz, hanem csak megkezd egy mozdulatot, azaz mozgást végez, de mozdulatot nem hajt végre.

Egyszeri (határozott) mozdulat a meghatározott számú *frame*-en keresztül (hosszabb időn át) végzett, az adott testrészt a nyugvópontjára visszatérítő mozgás, pl.: amikor a személy a címzettre (kérdezőre) mutat, vagy osztenzív mutatás történik, mert a beszélő a felemelt kezét visszaviszi a törzse mellé vagy ismét a combján nyugtatja, esetleg ujjait megint összekulcsolja.

Egyszeri mozgás a pillanatnyi időre megemelt kéz vagy láb, a testtagok rándulásai: ezeket ugyan annotálhatjuk, de felesleges funkciót tulajdonítani nekik.

Periodikus mozgás például a lábrázás vagy -lóbálás, ahol a mozgás, amely ismétlődik, sokszori és oly rövid, hogy megcímkézése legfeljebb felületesre sikerülhet, mivel nem feleltethető meg neki az „egy *frame* jobbra, egy *frame* balra" – különben rendkívül körülményes – címkézési módszere.

Abban az esetben, ha az illető csak az ujjával malmozik, ez a módszer még célravezető lehet, és a mozdulatsornak szó szerint teljes (címkékkel való) lefedettséget biztosít. A valódi, azaz gyors periodikus mozgás ellenben saját címkét venne igénybe, mivel egy frame-en belül a mozgás többszöri megismétlése is lehetséges (egy frame a Qannotban kb. 247 millisecundum).

A túl aprólékos vagy éppen felszínes szegmentálás másik következménye az, hogy igen rövid, egy vagy két egységnyi időszakasz alatti történések nem ugyanazokat a címkéket kapnák, mint amelyeket egy fokozott sebességű annotáció alatt. Megfordítva: amennyiben az annotátor mindig hosszú, több perces szakaszokat határol el, a kommunikációs események és az őket alkotó mozdulatok, mozdulatsorok vagy a változó arcjáték részei nem különülnek el megfelelő élességgel. Feltétlenül igaz viszont az, hogy nem szabad (a fentiek ismeretében nem is lehet) részeire tagolni egy periodikus mozgást, mint amilyen a fej ingatása (címkéje: „sideways"), a bólogatás („nod") vagy a fej rázása („shake"). Nyitva hagyható az a kérdés, hogy a fej csóválása (a rázásnál jóval lassabb jobbról balra és balról jobbra fordítása) periodikus mozgásnak számítson-e. Ha igen, ugyanúgy „shake" címkével kell ellátni; ha nem, jobbra fordításnak („turn right") és balra fordításnak („turn left") kell annotálni.

A periodikus mozgások gyakorta a pótcselekvéseknek feleltethetők meg, bár a kéz tördelése („broke") kivétel.

3.2. Az arcon és a hanggal közvetített érzelmek lehetséges funkciói

A videoannotálás során arckifejezéseket is annotáltunk. Az érzelmi állapotokra utaló címkékkel az arcon közvetített helyzetértékelést is rögzítjük: egy szorongó vagy gunyoros arckifejezésből nagyjából meghatározhatjuk a beszélő viszonyulását saját mondanivalójához vagy a partnerétől hallott kérdéshez/kommentárhoz. Azonban az annotátor közelítőleg sem állapíthat meg semmit a beszélői attitűdről, ha nincs tudatában, hogy a mimika – főleg a formális beszélgetésekben – függetlenedhet a megfigyelt személy tényleges hangulatától és beállítódásától, éppen valódi véleményének eltitkolása okán. Ezért az érzelmi vonatkozású videoannotációs szinteken („facial expression class", „emotion") adott időszakaszon mindig a korábbi állapothoz vagy a későbbihez viszonyítva adhatunk címkét (azaz az éppen annotálandó szakasznál nagyobb szakasz figyelembe vételével), ugyanis valaki akkor „happy", ha a későbbi vagy korábbi nyájas mosolyához képest boldog. Bár ez az eljárás erősen szubjektívnek tűnhet, de vegyük figyelembe, hogy mindenkinek van egy „alapmimikája", vagyis a helyzethez illesztett állandónak nevezhető arckifejezése, amelyhez mint alapállapothoz a felfokozottabb érzelmi állapotokból visszatér. Ez pedig ugyanúgy lehet egy közönyt sugárzó merev arckifejezés, mint egy folytonos udvarias mosoly. Az alapmimika és a tényleges érzelmi állapotot tükrözni akaró közötti szembenállás tagadhatatlanná lesz, ha összevetjük a „facial expression class" és „emotion" annotációs szinteket. Az „emotion" szinten, az egyén hangját hallva meggyőződhetünk róla, arckifejezése mennyire őszinte, illetve, amit nyugodt vagy derűs hangja nem árul el, az esetenként az arcára van írva. Példák:

(1) Fájlazonosító: `071_J_C2.mts`

Vizsgált időszakasz (perc, másodperc, századmásodperc): 03:13:20 – 03:14:75

Elhangzott szövegrész: „minden további nélkül megtanulom."

Megjegyzés: A megfigyelt személy arca („facial expression class") előbb nyugodt („natural"), aztán szomorú („sad"). Ugyanezen időben az érzelemkifejezés szintjén („emotion") feszült („tense") van jelölve.

(2) Fájlazonosító: `071_J_C2.mts`

Vizsgált időszakasz: 03:25:59 – 03:32:75

Elhangzott szövegrész: „…keretein belül is bizonyos technológiákkal."

Megjegyzés: A vizsgált időszakaszban az egyén eleve feszült, de az eztán következő idegességéhez („tense") képest ez még semleges érzelemként („natural") jelölhető az érzelemkifejezés („emotion") szintjén.

A 071-ből vett második példa nem tekintendő együttjárásnak, a viszonyító annotációra viszont jó péla lehet.

3.3. Együttjárások

Az interjúban elhangzottakat a videókon megerősítheti vagy relativizálhatja a taglejtéssel kifejezett kommunikatív funkció, amely az annotáció két szintjén jelölhető. Jelölhető mint utalás („deictic" level), és/vagy jelölhető az ún. emblematikus szinten („emblem"). A partner figyelmének felkeltése például megjelenhet valamilyen gesztussal, a partnerre (a kérdezőre) figyelés megjelenhet gesztus nélkül. Ebben az annotációtípusban csak egy címke van a figyelemre („attention"). Az unimodális (csak lehetséges kommunikatív funkciók jelölőit nyilvántartó) annotációban elfogadott, videoannotációban viszont paradoxnak hat, hogy egyetértő hangzó megnyilatkozást fejrázás kísérjen (a példákban az időtartam jelölése: perc:másodperc:ezredmásodperc).

(3) Fájlazonosító: `071_J_C2.mts`

Vizsgált időszakasz: 03:13:20 – 03:14:75

Elhangzott szövegrész:„minden további nélkül megtanulom."

Megjegyzés: A felsorolásban emblémának nevezett szinten az elhangzott szövegrész értelmében egyetértés („emblem: agree") jelölhető, ami viszont fejrázással jár együtt („headshift: shake").

(4) Fájlazonosító: `077_J_C2.mts`

Vizsgált időszakasz: 01:26:79 – 01:28:76

Elhangzott szövegrész: „hát szeretnék."

Megjegyzés: Az elgondolkodó/emlékező arckifejezés („facial expression class: recalling") a másfelé (nem előre) nézéssel („gaze: left down") jár együtt.

(5) Fájlazonosító: `077_J_C2.mts`

Vizsgált időszakasz: 01:22:79 – 01:26:76

Elhangzott szövegrész: „miért jelentkezett a hirdetésre?" (A kérdező mondja, a kérdezett ezalatt figyel.)

Megjegyzés: A figyelem („attention") az annotátor számára nem pusztán az előre irányuló merev tekintetből tevődik össze („emotion: natural"; „gaze: forwards"). Ha így lenne, nem kellene a figyelmet, amit észlel, az arra fenntartott szinten jelölni („emblem: attention").

(6) Fájlazonosító: `077_J_C2.mts`

Vizsgált időszakasz: 01:42:00 – 01:44:35

Elhangzott szövegrész: „… vagy távolállónak érez?" (A kérdező mondja, a kérdezett figyel.)

Megjegyzés: A 077_J_C2.mts videón (ahogy sok másikon is) megfigyelhető, hogy amíg a figyelem a kérdező felé fordul, az érzelemkifejezés az arcon felfüggesztődik. Ez az erős koncentráció jele.

(7) Fájlazonosító: `077_J_C2.mts`

Vizsgált időszakasz: 00:40:39 – 00:42:75

Elhangzott szövegrész: „uhum. és ilyen fodrászképző?" (A kérdező mondja, a kérdezett figyel.)

Megjegyzés: A vizsgált időszakaszban figyelem és feszült arckifejezés észlelhető.

(8) Fájlazonosító: `077_J_C2.mts`

Vizsgált időszakasz: 01:29:20 – 01:29:96

Elhangzott szövegrész: „… állást kapni."

Megjegyzés: Figyelem („attention") az emblémaszinten és meglepettség („surprise") az arcon.

(9) Fájlazonosító: `077_J_C2.mts`

Vizsgált időszakasz: 01:38:40 – 01:39:15

Elhangzott szövegrész: „… kérdés, …"

Megjegyzés: Ha a figyelem lankad, a kérdezett lefelé néz (jobbra le, balra le vagy középre le). Ebben a konkrét esetben a kérdezett balra lefelé néz („gaze: left-down"). 01:38:40-ig figyelem („attention") volt jelölhető, ettől az időponttól kezdve már nem, vagyis a figyelem megszűnt.

(10) Fájlazonosító: `077_J_C2.mts`

Vizsgált időszakasz: 01:45:60 – 01:47:56

Elhangzott szövegrész: A kérdezett semmit sem mond, de a fejét rázza („headshift: shake").

Megjegyzés: Itt szemlesütés figyelhető meg („gaze: left down"), valószínűleg kínos meglepetés hatására („facial expression: surprise").

(11) Fájlazonosító: `077_J_C2.mts`

Vizsgált időszakasz: 01:44:79 – 01:45:56

Elhangzott szövegrész: „hát"

Megjegyzés: Ugyanaz a szemlesütés látható, mint az előző példában, viszont az arckifejezés gondolkodó, emlékező, egy szóval elmélkedő („recalling").

Azt is megjegyezzük, hogy a példákban nem biztos, hogy a megadott arckifejezés nem tart tovább az időpontokkal leírt szakasznál, de bizonyos, hogy az együttjárás az így megjelölt időintervallum teljes hossza alatt fennállt.

A képi anyag, illetve hangos képanyag annotációja után most a csak a hanganyag annotálásával szerzett tapasztalatainkról számolunk be.

4. Az audioannotálás tapasztalatai. Automatikus prozódiai annotáció

Az audióanyag első annotálását kézzel végeztük. Ennek folyamán létrejött az elhangzott szöveg átírása, továbbá sor került bizonyos kommunikatív jelenségek (egyebek között beszélőváltás, megakadás, újraindítás, különböző érzelmek) annotálására. E kezdeti szakaszban mindezeket a kommunikatív jelenségeket olyan, időnként esetlegesnek tűnő időintervallumon belül határoztuk meg, amely sokszor, de távolról sem mindig a tagmondat valamely tentatív fogalmának felelt meg. Ezen intervallumon belül a jelenségek pontosabb időbeli meghatározására nem került sor. Míg ez az annotáció számos fontos kommunikatív jelenség jelölését lehetővé tette, az az igény, hogy mindezek strukturálisan és időben is pontosabban köthetőek legyenek egy formális nyelvi szerkezethez (és így adatbázisban az együttállásokat is figyelembe véve lekérdezhetőek legyenek), a későbbiekben szükségessé tett egyrészt egy formálisan megragadható szintaktikai annotációt, másrészt azt, hogy a prozódiai paramétereket egy speciális fonetikai szoftver alkalmazásával automatikusan annotáljuk. Az alábbiakban ez utóbbi annotáció részleteit ismertetjük.

4.1. Az automatikus prozódiai annotáció szerepe

A multimodális pragmatikai annotáció során feltárt kommunikációs események, illetve a szintaktikai annotációkban jelölt nyelvi egységek elemzési lehetőségei csak úgy válhatnak teljessé, ha gépileg feldolgozható formában rendelkezésre állnak az őket kísérő, azok potenciális markereit képező vizuális és akusztikus jegyek. A HuComTech korpusz videoannotációi a vizuális oldalról szolgáltatják ezeket az információkat (nem-verbális gesztusok, tekintetirány, különböző arckifejezések manuális és automatikus címkézése). Az automatikus prozódiai annotáció célja, hogy hasonló információk akusztikai oldalról is elérhetővé váljanak.[3]

A megvalósítás első lépését a beszédfolyam alapvető fizikai paramétereit képező adatok (mint amilyen az alapfrekvencia és az intenzitás) kinyerése és tá-

[3] Az automatikusság fokozott igénye ezen a területen nem csupán a hatékony megvalósítás érdekében áll fenn, hanem egyúttal bizonyos szupraszegmentális jelenségek (mint pl. a relatív beszédtempó vagy hangmagasság) manuális leírási nehézségeit, megbízhatatlanságát és szubjektivitását igyekszik áthidalni. A nyelvhasználat során feltehetőleg öntudatlanul, de operálunk ezekkel az információkkal, viszont a prozódiai annotáció esetében (ahol az annotátor szubjektív észleleteire reflektál) ugyanezen jelenségek sokkal kevésbé ragadhatók meg egyértelműen és objektív módon, mint például a fejmozgás iránya a videó annotációja során.

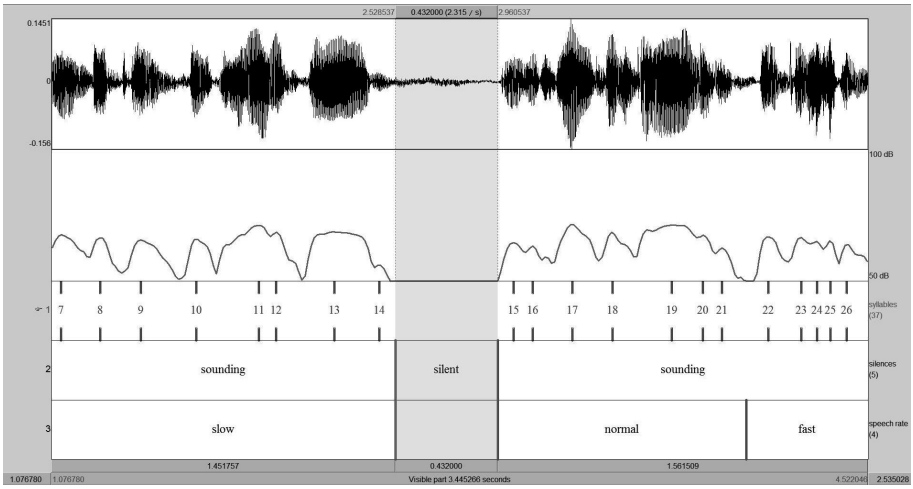
rolása jelenti, amely után közvetlenül is lehetőség nyílik a különböző pragmatikai és szintaktikai címkék mentén történő vizsgálatok elvégzésére. A szükséges adatok lekérdezésével statisztikai számításokat végezhetünk például arról, hogy egy adott típusú kommunikációs gesztus a többihez képest milyen átlagos hangmagassággal és intenzitással valósul meg a verbális közlés folyamán. A következő lépést az intenzitás- és frekvenciaadatok összetettebb feldolgozása jelenti, amelynek során további szupraszegmentális jelenségeket kívánunk lekérdezhető formában felcímkézni. Ezeknek a címkézési eljárásoknak képezi tárgyát a beszédtempó és a beszéddallam annotációja, amelynek lépéseit az alábbiakban egy kapcsolódó tanulmány (Szekrényes et al. 2011) alapján összegezzük.

4.2. A beszédtempó annotációja

A beszédtempó annotációjához elsősorban a beszédfolyam egy olyan automatikusan detektálható elemére van szükségünk, amelynek egy adott időegységre mért gyakorisága, sűrűsége megragadhatóvá teszi annak metrikus (időbeli) struktúráját. Jong és Wempe (2009) a beszédtempó vizsgálatához a szótagmagokat választották mérési objektumként, amelyeknek az automatikus detektálása a beszéd intenzitásának dinamikus kiugrásai alapján, az intenzitásgörbe csúcsértékeinek meghatározott küszöbértékek (csúcsok közötti minimális értékbeli különbség stb.) szerinti szűrése által történik. A beszéd sebességének ingadozása így az intenzitáscsúcsok közötti távolság változásain keresztül válik mérhetővé, ahol minden intenzitáscsúcs egy szótagmag helyét reprezentálja (2. ábra). A szótagmagok detektálása után előzetes kalkulációkat végezhetünk az adott beszélő egyedi beszédtempójáról, majd a beszélő normál (átlagos) beszédtempójához viszonyítva osztályozzuk az aktuális beszédtempót a felcímkézendő beszédszegmensek mentén. A címkézési eljárás egy lehetséges kimenetét a 2. ábra szemlélteti a Praat program (Boersma–Weenink 2005) annotációs felületén. A módszer tervezett implementálásának további részleteiről lásd még Szekrényes et al. (2011) tanulmányát.

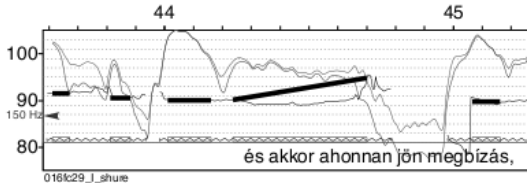
4.3. A beszéddallam annotációja

A prozódiai annotáció következő lépését az alapfrekvencia progressziójának (a dallammenet alakulásának) elemzése és címkézése jelenti. Az eljárás során a beszédfolyam meghatározott szegmentumaiban kimért F0 értékek leírta dallammenetet igyekszünk a progresszió fő irányvonalait megragadó stilizált formára



2. ábra. A beszédtempó annotációja

hozni (ezt szemléltetik a 3. ábra vastag vonalai), amelyhez meghatározott küszöbértékek használatával valamilyen egzakt tonális karaktert jelelő annotációs címke (emelkedő, ereszkedő, eső stb.) vagy címkekombináció rendelhető.



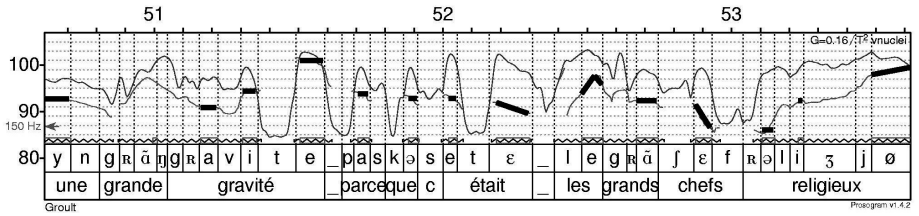
3. ábra. A beszéddallam stilizációja

Ennek a célnak a megvalósításához gyakorlati és elméleti útmutatásként Piet Mertens munkáját (Mertens 2004) terveztük felhasználni, aki tanulmányában számos fontos előzetes kikötést fogalmaz meg a prozódiai annotációval kapcsolatban:

1. Az annotációnak alapvetően az ember által érzékelhető intonációt kell reprezentálnia objektív és könnyen értelmezhető módon.
2. Az alapfrekvencia változását hosszabb beszédfolyamon keresztül is tükröznie kell a szélesebb tartományokra kiterjedő változások rögzítése érdekében.
3. A fizikai jelek időbeli szerveződését meg kell őriznie a szünetek, hezitációk, beszédtempó és a ritmus azonosíthatósága érdekében.
4. Az annotációnak automatikusnak vagy félautomatikusnak kell lennie.

5. Az annotáció elméletsemleges kell, hogy legyen, a széleskörű használhatóság érdekében.
6. Az annotáció lehetőleg időben illesztett fonetikai és szöveges átírást tartalmazzon az olvashatóság és szöveges keresés lehetőségének biztosítása érdekében.

A szerzők által kifejlesztett transzkripciós rendszer (l. 4. ábra) a vokális szótagmag alapfrekvenciájának stilizált kontúrját felhasználva automatikusan vagy félautomatikusan rendel prozódiai annotációt fonetikai transzkripcióhoz. Maga a stilizálás a fönti kikötésekkel ellentétben ugyan nem teljesen elméletsemleges, hiszen a tonális érzékelés pszichoakusztikai modelljére épül (l. Alessandro–Mertens 1995), viszont megőrzi az akusztikai jel temporális jellemzőit, és beépíti a szöveges, illetve a fonetikai átírást is, ahol ez utóbbi a vokális szótagmag azonosításában játszik szerepet.



4. ábra. A Mertens-féle transkripciós rendszer

A módszer algoritmizálása a Praat beszédfeldolgozó program beépített szkriptnyelvén történt. A transzkripciókat grafikus formában generáló Praat szkript a hozzá tartozó dokumentációval együtt Prosogram (v2.8) néven szabadon hozzáférhető,[4] így a jelen tanulmányban részletesen nem ismertetjük. Az eredményül kapott stilizációkat a különböző beszédszegmensek dallamkarakterének címkézéséhez kívánjuk felhasználni. Mivel a program grafikus formátumú kimenete az ehhez szükséges információk feldolgozását nem támogatja, az algoritmus olyan technikai jellegű átdolgozása szükséges, amelynek eredményeként a kimenet numerikus formában tartalmazza a stilizációk kezdő- és végpontjának időpillanatait és frekvenciaértékeit is.

A munka jelenleg előkészítő szakaszban van. Az automatikus prozódiai annotáció sikeres implementációja után lehetőség nyílik olyan lekérdezések összeállítására, amelyek a manuálisan annotált kommunikációs események prozódiai markereinek feltérképezését segíthetik elő. Ezeknek a feltárt prozódiai jellemzőknek a birtokában a későbbiekben lehetővé válik a vizsgált kommunikációs események (társalgási fordulók, témaváltások stb.) gépi detektálásának vagy pre-

[4] http://bach.arts.kuleuven.be/pmertens/prosogram/

dikciójának algoritmizálása, amely a számítógéptől ugyanezen prozódiai jellemzők felismerését követeli meg. A prozódiai annotáció során kifejlesztett eljárások ehhez szintén jól alkalmazhatóak lesznek.

5. Esettanulmány: a nem szándékolt ismétlések multimodális jegyei

Az előző pontokban ismertetett video- és prozódiai annotálás után egy rövid esettanulmányban mutatunk ízelítőt a HuComTech korpuszban rejlő lehetőségekből. A korpusz adataira támaszkodva azt a feltevést kívánjuk alátámasztani, hogy a nem szándékos ismétlések egy, az ismétlést követő tartalmas szó[5] memóriából történő előhívását könnyítik meg. Ezt a szót a beszélő nyomatékosítani kívánja, hogy ezzel is segítse a feldolgozást a hallgató számára. A beszélő a nyomatékosításhoz nemcsak akusztikai jelenségeket, hanem vizuális, nem verbális markereket is alkalmaz. Korpuszunk lehetővé teszi, hogy ezeket a nem verbális jeleket együtt vizsgáljuk az akusztikai jelekkel, így a nem szándékos ismétlés multimodális vizsgálata hozzájárulhat a kommunikáció igen összetett jelenségének jobb megértéséhez és az ismeretek számos nyelvészeti, nyelvtechnológiai és egyéb célú alkalmazásához.

Az ismétléseket úgy tekinthetjük, mint a spontán beszéd folyamatában, egy szó kivitelezésének a beszélő bizonytalanságából adódó, nem szándékos megismétlését (Gósy 2002). Az ismétlés történhet változtatással vagy változtatás nélkül (Fox et al. 1996, 230). Jelen esettanulmányban az egyszerű, változtatás nélküli ismétléseket tanulmányozzuk.

5.1. Az esettanulmány vizsgálatának előzményei

Az ismétléseket szintaktikai környezetükben, az azokat tartalmazó tagmondatokban vizsgáltuk. Az alábbiakban, az egyszerűség kedvéért, tagmondatnak nevezzük azt az egységet, amelyben az ismételt szó és az ezt követő tartalmas szó elhelyezkednek, függetlenül attól, hogy az adott tagmondat a spontán beszédbeli megnyilatkozásban gyakran jelentősen különbözhet attól, amit a leíró nyelvtan

[5] Kenesei (2000) tartalmas szavaknak nevezi a főnevek, igék, melléknevek, határozószók alkotta nyitott szóosztályokat, amelynek komponensei kölcsönzéssel, szóképzéssel stb. gyarapíthatók, míg ezzel szemben a funkciószavak (segédigék, kötőszavak, névelők) száma változatlan, állandó (*op.cit.* : 95).

ezen fogalma takarna.[6] A tagmondatokat a diskurzusbeli elhelyezkedésük szerint csoportosítottuk: forduló fenntartása, fordulóváltás (átadás/átvétel).

A HuComTech korpusz adatai alapján történt korábbi vizsgálatok (Abuczki 2011; Tóth 2011), melyek célja a fordulóátadás, illetve -átvétel során megjelenő tekintet- és gesztusmintázatok elemzése volt, azt találták, hogy fordulóátadáskor az interjúalany tekintetét beszédpartnerére szegezi, és a forduló átvételekor is hasonló tendencia figyelhető meg. Ahhoz, hogy a vizuális jegyek értelmezésekor kizárjuk a fordulóátadás és -átvétel ilyen hatását, a nem szándékos ismétléseknek csak azon eseteit vizsgáltuk, amelyek nem estek egybe fordulóváltással. Az ilyen esetek közül kizártuk azokat a tagmondatokat, amelyekben együttbeszélés fordult elő, ugyanis ennek pszicholingvisztikai mechanizmusa eltérhet az egyedülbeszélésétől (vö. Gósy 2008; Gyarmathy 2011).

A HuComTech annotált spontánbeszéd-korpuszból felhasznált 6 informális dialógus korpuszba rendezett, időben szinkronizált multimodális adatai lehetővé teszik, hogy segítségükkel az ember–ember interakció vizuális és auditív markereit együttesen elemezhessük, együttállásokat és összefüggéseket állapíthassunk meg verbális és nem verbális jegyek között. Ezáltal hozzájárulhatunk a nem szándékolt ismétlések multimodális tulajdonságainak a megismeréséhez, és ezek felismeréséhez akár egy ember–ember, akár egy ember–gép kommunikációra fókuszáló alkalmazás számára. Az itt ismertetendő eredményeink a vizsgált dialógusokon belül az interjúalanyok adataira támaszkodnak.

Az interjúalanyok 97 tagmondatában fellelhető 106 ismétlést, valamint az azokat közvetlenül követő tartalmas szavaknak a beszélő általi szándékolt előhívására utaló tekintetét és gesztusait vizsgáltuk abból a szempontból, hogy e kétféle mozzanat során az alanyok milyen vizuális markereket használtak.

5.2. Az egyes markerek együttállása: az interjúalanyok tekintetviselkedése az ismétlés során

A beszélők a szó ismételt kiejtése során 26,4%-ban (28 esetben) tekintettek a szemben ülő partnerre (továbbiakban röviden TP (= tekintet a partnerre)), és 73,6%-ban (78 esetben) egy ettől eltérő irányba (rövidítve TM (= tekintet más irányba)) irányították tekintetüket.[7] Ez az ismétlés alatti magas arányú tekintet-irány-változás feltehetően azért történt, mert eközben a megformálandó szón

[6] A spontán beszéd szintaktikai felosztásáról részletesebben a következő pontban lesz szó.

[7] Bár lekérdezhetőek korpuszunkból, az irányok részletezésére (bal-jobb, fel-le) itt nem törekedtünk, mert személytől függ, ki milyen irányba tekint a beszéd során.

vagy szintaktikai szerkezeten gondolkodtak, azaz a tekintet irányának módosítása összefüggésben volt a nem szándékolt ismétléssel. Kitűnt, hogy azokban az esetekben, ahol az interjúalany tekintete az ismétlés során az interjúvezető felé fordult, a beszéd menetében nem történt megtorpanás. Ebből arra következtetünk, hogy ekkor a soron következő szó/kifejezés előhívása és produkciója minden bizonnyal nem okozott különösebb kognitív nehézséget.

5.3. Az ismétlések utáni tartalmas szavak tekintetmintázatai

Az esettanulmányhoz felhasznált adatokat két csoportba sorolhatjuk annak alapján, hogy az ismételt szót követően megtörténik-e a tartalmas szó előhívása:

1. Megtörténik az előhívás: a beszélő ismétlését követő első tartalmas szó alatti tekintetmintázatok vizsgálhatók. Például:[8]

(12) hogy mennyire változatos a <a> <a> <a> kultúra

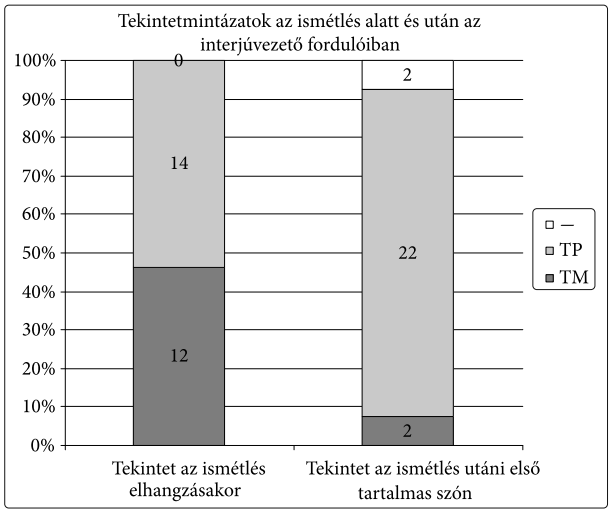
2. Nem történik meg az előhívás (Ø = nincsen ismétlést követő tartalmas szó, amelynek a tekintetmintázatát vizsgálhatnánk). Vagy azért, mert az ismétlést nem követi tartalmas szó, mert az interjúalany új szintaktikai egységet kezd, ezáltal befejezetlenül hagyva a megkezdett tagmondatot (13), a tagmondatok közötti határt a „|" jel jelzi), vagy mert tartalmas szó helyett újabb ismétlés következik (14). Ezekben az esetekben – a tartalmas szó előhívásának hiánya miatt – ezen mozzanat nem vizsgálható:

(13) szóval ez így <így> – | ah, ez így fáj.

(14) úgyhogy <úgyhogy> én <én> amikor így <így> láttam,

Az esetek 90%-ában (96 db tartalmas szónál) vizsgálható volt a tekintet iránya az ismétlést követően, így ezeket az első csoportba soroltuk. Az első csoportba tartozó mintákat is megvizsgáltuk a tekintet iránya szerint (TM/TP), és a következő eredményre jutottunk: a tartalmas szavak 25%-ában nem tekintettek a partnerre a beszélők (TM = 24 db), míg a vizsgált esetek 75%-ában a tekintet a partner felé irányult (TP = 72 db). Az 5. ábra tartalmazza az ismétlések kétféle tekintetmintázatát, valamint az első tartalmas szavak tekintetmintázatát, illetve annak hiányát darabszámok szerint. A második csoporthoz sorolható minták az esetek 10%-át (10 db) alkotják. Mivel ez utóbbi esetben a tartalmas szó elmarad, a lehetséges tekintetmintázatra és annak vizsgálatára sem került sor.

[8] Az itt használt jelek megegyeznek a HuComTech korpusz audioannotációja során használt szimbólumokkal, azaz: „%" = nyújtás (a megnyújtott hang előtt szerepel), „< >" = ismételt szó, „–" = befejezetlen tagmondat.



5. ábra. Összesítő grafikon a tekintetmintázatok eredményeiről

5.4. A tekintetviselkedés együttállásai

A leírt tekintetmintázatokat a kommunikáció multimodális jellege miatt érdemes együttjárásaik szerint is megvizsgálni. Vessük össze az ismétlések alatti és az ismétlések utáni tekintetirányokat. A rendelkezésre álló adatok alapján az ismétlés alatti és az ezt követő első tartalmas szó tekintetviselkedésének hat lehetséges kategóriája különíthető el:

a) TM + TP = ismétlés alatt a tekintet a partner irányától eltérő irányba mutat, míg a soron következő első tartalmas szó elhangzásakor a beszélő a partnerre tekint;

b) TP + TM = a beszélő az ismétlés alatt a partner irányába tekint, de a tartalmas szó alatt más irányba;

c) TP + TP = a beszélő végig az interjúvezető irányába tekint;

d) TM + TM = a beszélő sem az ismétlés elhangzásakor, sem az azt követő tartalmas szó alatt nem néz a partnerre;

e) TM + Ø = a beszélő tekintete nem a partnerre szegeződik az ismétlés alatt, viszont nem követi őt ugyanazon tagmondatban tartalmas szó, amelynek tekintetmintázatát elemezhetnénk;

f) TP + Ø = a beszélő a partnerre tekintve megismétli a szót, de az ismétlést tartalmas szó nem követi.

A leggyakrabban előforduló együttállás a TM + TP (az összes vizsgált jelenség 47%-a (45 eset)), vagyis, amikor a beszélő másfelé tekint az ismétlés alatt, de az első lényeges tartalmas szó kiejtésekor, mintegy megerősítésként, a partnerre tekint. Ez utalhat arra, hogy a szóban forgó ismétlések valamilyen információban gazdag vagy fontos szót előznek meg.

Az a) pontban található együttállás ellenkezőjére (TP + TM), vagyis amikor a beszélő az ismétlés elhangzása után tekintetét elkapja a hallgatóról, nem találtunk példát a vizsgált hat beszélőnél, ami két dolgot jelenthet. Egyrészt azt, hogy az ismétlés nem valamilyen hiba korrigálása, azaz nem egy kognitív önellenőrzési folyamat jelölője, hanem időhúzó tényező. Ebben az utóbbi esetben eredményünk alátámaszthatja Németh (2012) következtetését, amely szerint az ismétlés feladata az időhúzás. Másrészt várható az is, hogy ha az ismétlés során a beszélő nem bizonytalanodik el, tehát tekintete a partnerre szegeződik, akkor a soron következő ismétlés sem fog produkciós nehézséget eredményezni, amely miatt a beszélő hirtelen elkapná tekintetét az interjúvezetőről.

A TP + TP (27 eset, 28%), illetve TM + TM (24 eset, 25%) jelenségek, mint láthatjuk, közel azonos arányban fordultak elő. Az előbbi esetben a hallgató számára nem okoz problémát a beszéd követése és szemantikai tartalmának értelmezése, mivel a beszélő tekintetével figyelemmel követi a passzív fél reakcióit. Az utóbbi eset viszont kísértetiesen hasonlít a telefonon keresztül zajló spontán beszélgetésekre, amely szituációkban a hallgató nem látja az aktuálisan beszélő fél tekintetét, az ismétlés mégsem számít megakasztó tényezőnek.

Az utolsó csoportra (f) nem volt példa, tehát a tartalmas szó előhívása nélküli nem szándékolt ismétléskor a beszélő mind a 10 esetben más irányba nézett (TM + Ø).

5.5. Az ismétlések viszonya további multimodális markerekkel

A 96 tekintet-együttállást megvizsgáltuk a gesztusmintázatok szempontjából is. Arra voltunk kíváncsiak, hogy a személyközi kommunikációban milyen más modalitások játszanak még közre – a tekintet mellett – az elhangzott információadó tartalmas szavak hangsúlyos jellegének érzékeltetésében. Először a 24 esetet adó TM + TM tekintetmintázat-pár során észlelhető gesztusok vizsgálatainak eredményeit közöljük. Azokban az esetekben, amikor az interjúalany tekintete az ismétléstől kezdve az első tartalmas szóval bezárólag a partnertől eltérő irányba irányul, a kézi gesztusokat és a fejmozgást figyelve a következő kapcsolatokat detektáltuk: a kézi gesztusok és a fejmozgás (oldalra/előre irányuló fejbiccentés, bólogatás, fejrázás) önállóan, és együttesen is kísérhetik az ismétlést és az első tar-

talmas szót. Összesen 18 kézi gesztust és 9 fejmozgást azonosítottunk. Fejmozgást önállóan 2 esetben sikerült észlelnünk, kézi gesztikulációt pedig 11 esetben. Vagyis 7 esetben a két jelenség együttesen kísérte az ismétlést és az első tartalmas szót, és 5 esetben nem volt vizuálisan értékelhető mozgás (6. ábra).

6. ábra. TM+TM esetben megnyilvánuló gesztusmintázatok. Az elhangzott ismétlések: „de <de> amúgy tényleg a tanár az legalább"; „hogy <hogy> igen Petit el szeretném vinni egy sörre"; „viszont mi a <a> <a> lány legjobb barátnőjével csináltunk egy új baráti kört"; „meg <meg> <meg> van az emberben egy ilyen igény arra"

Feltehetjük ugyanakkor, hogy a gesztusok nem csak a fent elemzett esetben segítik a feldolgozást a hallgató részéről, hanem akkor is, amikor a beszélő tekintetének iránya a partner felé irányul (7. és 8. ábra). Ezek alátámasztásához azonban további mérési adatok és vizsgálatok szükségesek.

Az itt bemutatott eredményeket a nem szándékos ismétlések nem verbális jelei alapján kaptuk. Természetesen a multimodalitásnak része a beszédhangokkal való manipulálás is, beleértve a beszéddallam, az intenzitás és a tempó (benne a szünet) eszközeinek a használatát. Jól észlelhetően a prozódia ilyen jellegű használata is jelen van vizsgált anyagunkban, pontos adatokat azonban csak további mérések alapján tudunk közölni. Az ismétlések multimodális kifejezésének itt bemutatott eredményei ugyanakkor jól alátámasztják Hunyadi (2011a) azon állítását, hogy kézzelfogható kapcsolat van a verbális és a nem verbális jelenségek között.

7. ábra. Gesztusmintázatok, amikor a beszélő a tartalmas szó produkciója során a partnerre tekint. Az elhangzott tagmondatok a képek sorrendjében: „hogy <hogy> kitaláltam ezt"; „hát <hát> az egyszerű volt"; „de <de> látom az ilyen jeleket"; „azok a <azok> az amerikaiak"; „hát <hát> előkapták a pillangókést"

8. ábra. Nincs kézi gesztikuláció, csak oldalirányú fejbillentés („ez <ez> a balesetem volt")

Ez a rövid esettanulmány tehát arról tanúskodik, hogy a HuComTech korpusz hat interjúalanyának ismétlései és az ezt követő tartalmas szavak kiejtése alatti tekintetviselkedés és gesztikuláció egymással összefügg. Az ismétléseket követő tartalmas szavak memóriából történő előhívásához, illetve nyomatékosításához a beszélők olyan vizuális elemeket használnak, mint a tekintet irányával való manipulálás, kézi gesztikuláció és fejmozgások. Az esettanulmányban vizsgált verbális kifejezések produkciója és a hozzá tartozó mozdulatok közötti viszony azt

sugallja, hogy lehetséges összefüggés az ismétlések bizonyos típusait követő egyes gesztusok megléte és a nyelvi produkció mögötti kognitív működés között.

6. A HuComTech korpusz szintaktikai annotációs szintje

A következőkben a szintaktikai annotációs szint bemutatására, a spontán beszéd nyelvi megnyilatkozásainak mondatokra és tagmondatokra való felosztására kerül sor.

6.1. A szintaktikai annotációs szint célkitűzése és szabályrendszerének kialakítása

A szintaktikai annotációs szint létrehozásának kettős célja van: egyrészt létre kívánunk hozni egy olyan eszközt, melynek segítségével adott szempontok alapján rendszerszerűen le lehet írni a magyar beszélt nyelv nyelvtanát, másrészt – mivel ez az szint egyike a HuComTech multimodális korpusz annotálási szintjeinek – ezáltal hozzá kívánunk járulni ahhoz, hogy kutathatóvá váljék a beszélt nyelv szintaxisának a kommunikáció egyéb moduljaival való összefüggése is. A prozódiával való összefüggésének vizsgálata hozzájárulhat a beszéd teljesebb automatikus felismeréséhez, a prozódiai és a gesztusokra vonatkozó információkkal együtt pedig az ember–ember, valamint az ember–gép kommunikáció jobb megértéséhez juthatunk.

Az annotációs alapelvek kidolgozása során az empirikus adatok gyűjtése és osztályozása együtt járt az összegyűjtött és osztályozott adatok egyfajta preteoretikus rendszerben való összegzésével. Az annotációs szabályrendszer tehát induktív és deduktív módszer együttes alkalmazásával jött létre. Megközelítésünk preteoretikus volta egyrészt tükrözi a különböző leírások és elméletek közötti konszenzust, ugyanakkor nem szándékszik olyan elméleti elkötelezettséget tenni, ami az eredményül kapott annotáció későbbi széles körű felhasználását korlátozhatná.

6.2. A szintaktikai szabályrendszer elemzési alapegységei (az annotáció szintaxisának alapjai)

Elemzésünk lényegét tekintve strukturális és nem funkcionális. A szintaktikai jelölési szabályrendszer keretei között központi fogalomnak a tagmondatot tekintjük. A tagmondat alapvető kritériumaként a predikatív viszonyt tesszük fel. Az

ennél szélesebb értelmű mondatot önmagában nem, csak a tagmondatok kapcsolódásain keresztül határozzuk meg. Ennek értelmében egy mondat ott ér véget (és potenciálisan ott kezdődik egy újabb mondat), ahol további tagmondatot már nem tudunk strukturálisan csatlakoztatni. A mondat belső szerkezeti összefüggései között fontosnak tartjuk a hierarchia és a szerkezeti hiány azonosítását és jelölését. A szerkezeti hierarchia esetében a tagmondatok közötti alá-, fölé- és mellérendelést jelöljük. A hiány fogalmának bevezetését a spontán nyelvi megnyilatkozások természete kívánja meg. Az egyes kapcsolódásokban azon szerkezeti elemek hiányát vizsgáljuk, amelyek az adott szerkezet építését befolyásolják. A szintaktikai szerkezetek ilyen szempontú besorolása lehetővé teszi, hogy világosan látható legyen, adott esetekben a grammatikalitás milyen fokon teljesült explicit módon.

Nem célunk a mondatoknak funkcionális elemzését adni, különösen azért, mert a spontán beszédben történő nyelvi megnyilatkozás funkciója – nem kis mértékben éppen a hiányok miatt – gyakran áttevődik nyelven kívüli eszközökre vagy akár jelöletlen is marad, így a megnyilatkozás funkcióinak rendszerszerű megismerése hangsúlyozottan az interpretáció feladata lenne. Ezért – mivel ezt számos esetben a struktúra nem jelöli – a tagmondatok közötti hierarchikus viszonyok jelölésén túli részletesebb elemzésre, így pl. az alá- és mellérendelő mondattípusok megnevezésére nem vállalkozunk.

Ami a hiányzó mondatelemeket illeti, a következő megfontolásokkal élünk. Csak azoknak az elemeknek a felszíni hiányzását tüntetjük fel szerkezeti hiányként is, amelyek vonzatok, azaz elhagyhatatlan bővítmények (vö. Komlósy 1992; Keszler 2000), tehát a grammatikai struktúra sérülése nélkül nem hagyhatók el azon nyelvi egység mellől, amelyhez tartoznak. Az alanyt viszont nem tekintjük vonzatnak, elmaradása mégis hiányként van jelölve annak ellenére is, hogy a személyragból következtethetünk rá.

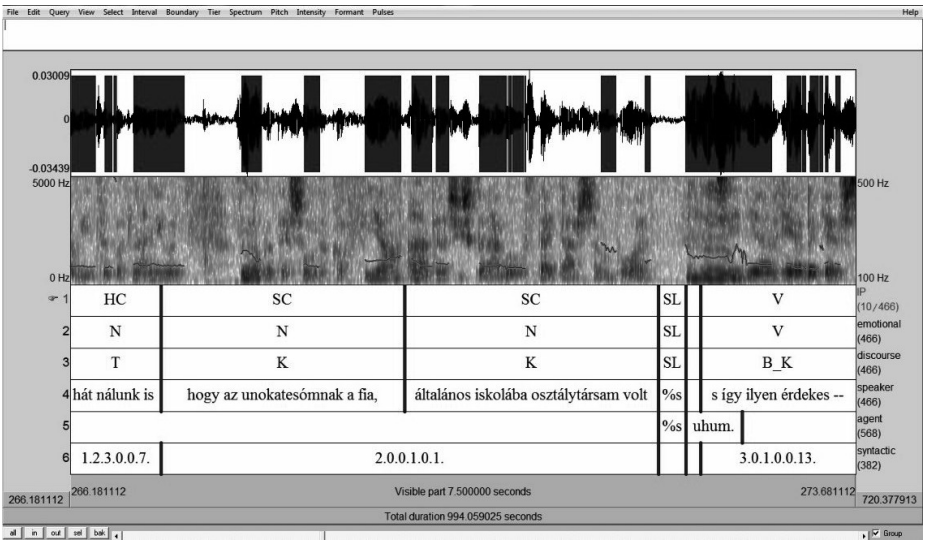
6.3. Az annotációs szabályrendszer módszere és a kódrendszer

Módszerünk egyrészt a kategorizálás, másrészt a formalizálás. Ábrázoljuk az egyes szintaktikai egységek sorrendiségét, mégpedig úgy, hogy – mivel szigorúan az elhangzott beszédben dokumentált tényekből indulunk ki – szabályrendszerünk megengedi az írott nyelvre vonatkozó leíró grammatika szempontjából nem megfelelő sorrend létrejöttét, létezését. Így megengedi a leíró nyelvtani szempontból „helyes" és „helytelen" mondatokat, és nem jelöli azok ilyen szempontú megkülönböztetését. Alapvető célja, hogy azonosítsa a strukturálisan meghatározható szintaktikai viszonyokat a mondatszerkezetek felépítésében (valójában a szavak egymáshoz való kapcsolódásaiban) és a tagmondatok egymásutániságában.

A kódrendszer a szintagmatikus összefüggéseket ábrázolja, a mondatrészi kategóriák egyszerű függőségi viszonyait, de megmutatja a lehetőséget arra is, hogy hogyan reprezentálhatjuk magát a kontextust. Az annotációs szabályrendszer hiánykategóriájának használatával csak a leíró nyelvtan szintaxisa szempontjából értelmezünk valóságos hiányokat és így „kivételt képező", netalán „helytelen" mondatszerkezeteket. Ezzel szemben a multimodális szinteken történő szimultán annotálás valójában éppen azt teszi lehetővé, hogy megfigyelhessünk, észrevegyünk és azonosítsunk egy adott modalitásból hiányzó elemet mint annak egy másik modalitásban és ugyanazon időpillanatában történő megvalósulását: így például egy szintaktikai szinten befejezetlen mondat lezárását nyomon követhetjük a hozzá kapcsolódó kézmozdulatok, mimika és egyéb jelek figyelembe vételével.

A szintaktikai kapcsolódások kódolásához számozást alkalmazunk. A több számjegyű számozás a tagmondatok közötti sorrendiséget, a tagmondatok közötti viszonyt és a fellépő szintaktikai hiány jellegét fejezi ki. Az első számjegy a mondat kezdetét jelöli, és az ezt követő számjegyek az egymást követő tagmondatok egymás közötti, valamint saját belső szintaktikai viszonyaira – beleértve a hiányokat is – utalnak. Ez a számozás ott fejeződik be, ahol további szerkezeti kapcsolatokat nem találunk. Így ez a pont valójában egybeesik azzal, amit egy hagyományos értelemben vett mondat végének neveznénk (és ami az ezt követő nyelvi anyagot egy újabb mondat kezdeteként jelöl meg). Így számunkra egy mondat határait nem az interpretáció és nem a prozódia határozza meg, hanem a – bármilyen hiányos – szintaktikai viszonyok.

A 9. ábrán láthatjuk a kódolási rendszert, a multimodális annotáció hatodik, legalsó szintjén. A kódrendszerben az első számjegy tehát az adott mondaton belül az egymást követő tagmondatok sorszámát jelenti. A második számjegy azt jelöli, hogy az adott tagmondathoz tartozik-e alárendelés, és ha igen, akkor ez a hányadik számú tagmondat. Ha ilyen alárendelés nincs, akkor ez 0 értékkel van jelölve. A harmadik számjegy a tagmondathoz tartozó mellérendelő tagmondat(ok) sorszámát jelöli. Ha ilyen nincs, akkor ott a 0 érték szerepel. A negyedik számjegy azt mutatja meg, hogy az adott tagmondat hányadik számú tagmondatnak az alárendeltje. Ennek hiányában itt is a 0 érték jelenik meg. (Az alárendelt kapcsolatban álló tagmondatok kölcsönös meghatározottsága lehetővé teszi, hogy e viszony jelölését függetlenné tegyük a tagmondatok felszíni sorrendjétől.) Az ötödik számjegy a grammatikai kapcsolat hiányát jelöli, rámutat a beágyazás és a beékelés jelenségeire. A hatodik számjegy a hiány kategóriáit hozza felszínre (főmondat, előtte vagy utána álló mellérendelő tagmondat, utalószó, kötőszó,



9. ábra. A szintaktikai annotációs szint kódolása

grammatikai, illetve logikai alany, állítmány, tárgy [tágyas ige esetén], határozó, jelző, ige), illetve jelöljük azt is, ha nem hiányzik semmi az adott tagmondatból, vagy ha a mondat befejezetlen, illetve ha irreleváns a hiány kategóriájának felvetése.

Jelölési konvencióinknak megfelelően az egymást követő, különböző elemzési szempontokra vonatkozó számok, jegyek közé pontot teszünk. Ha egy elemzési szemponthoz több számérték is tartozik, akkor az azokra utaló számjegyeket vesszővel választjuk el.

Tudatában vagyunk annak, hogy az itt ismertetett annotációs rendszer nem ad választ a legrészletesebb viszonyokra irányuló szintaktikai kérdésekre, és különösen nem elégíti ki a funkcionális nyelvleírás mentén felvetődő számos igényt. Mindez a választott megközelítésünk következménye. A szándékunk csupán annyi volt, hogy olyan preteoretikus annotálást adjunk, amelyből sokféle, egymástól különböző elméleti megközelítés is haszonnal kiindulhat, és főként azt várjuk, hogy a spontán beszéd szintaxisára is kiterjesztett annotált multimodális adatbázisunkkal a nyelvtechnológia, különösen a szintaxist valamilyen formában figyelembe venni szándékozó beszédfelismerés és -szintézis számára nyújthatunk közvetlenül is felhasználható anyagot.

7. A HuComTech korpusz pragmatikai szempontból: unimodális annotáció

A HuComTech korpusz szintaktikai annotációjának ismertetése után annak következő szintje, a pragmatikai szempontú annotáció kerül bemutatásra. A 7. pont az unimodális, funkcionális megközelítésű pragmatikai annotáció, a 8. pont pedig a multimodális pragmatikai annotáció mögött meghúzódó elméleti megfontolásokat, valamint az annotáció szintjeit és címkéit mutatja be. A HuComTech-projekt arra is lehetőséget nyújt, hogy a különösen a fizikai jelek felismerését és szintézisét előtérbe helyező video- és audioannotáció mellett vállalkozzunk egy olyan annotációra is, ami egy kommunikatív esemény pragmatikai viszonyait tárja fel. Ennek során ugyancsak figyelembe veszünk audio- és videojeleket, azonban ezekre az előzőktől eltérően tekintünk. Nem magukat a jeleket keressük és – ha megtaláltuk – annotáljuk, hanem az esemény pragmatikai vonatkozásait szem előtt tartva pragmatikai jellemzőket keresünk, és ezekhez rendeljük magukat a jeleket. Azaz ez az annotálás az előzővel ellentétben kifejezetten interpretatív, így a kétféle annotálás szükségszerűen feltételezi és kiegészíti egymást. Ez történhet unimodálisan éppúgy, mint multimodálisan. Mivel a kétféle megközelítés különböző elméleti elvárásokon alapszik, korpuszunkon mindkétféle megközelítést megvalósítjuk.

Az itt következőkben az unimodális funkcionális–pragmatikai annotációra térünk ki, és annak eszközét, alszintjeit és címkéit mutatjuk be. Ezt megalapozandó, röviden ismertetjük a séma kidolgozása mögött húzódó elméleti megfontolásokat és gyakorlati célkitűzéseket.

7.1. Elméleti megfontolások

Pragmatikai szempontból éppúgy, mint a technológia követelményeinek szem előtt tartása miatt céljaink közé tartozik a kommunikatív esemény bizonyos, unimodálisan is jól megragadható mozzanatainak kinyerése a HuComTech korpuszból, majd pedig a felismert jegyek, markerek alapján a társalgás menetével kapcsolatos predikciók megtétele. A technológiai szempont itt különösen jelentős: míg egy hétköznapi kommunikatív esemény során az lehet a benyomásunk, hogy az esemény mozzanatait holisztikusan, azaz az adott pillatanban elérhető összes (verbális és nem verbális) modalitás együttes feldolgozásával érzékeljük és értelmezzük, a technológiai implementáció megkívánja, hogy ezen összetett adatfolyamot jól kezelhető diszkrét elemeire bontsuk. Maga a vállalkozás azonban pragmatikaelméleti szempontból is fontos lehet, hiszen a kommunikáció mul-

timodálisan igencsak összetett eseményét így felbontva az értelmezés mögöttes kognitív folyamataira is fény derülhet.[9]

Több olyan alkalmazott nyelvészeti, társalgáselemzési megalapozottságú tanulmány született már az interperszonális kommunikáció invariáns struktúrájának és szekvenciális elrendezésének, rendezőelveinek megragadásával kapcsolatban, amely a kézenfekvő verbalitásra helyezve a hangsúlyt, multimodálisan értelmezte a kommunikációt (Sacks et al. 1974; Sacks 1995; Németh T. 1996; Schegloff 2006; Abuczki 2011), ugyanakkor igen kevés olyan kutatás folyt és tanulmány született, amely különválasztva a modalitásokat, unimodális megközelítést alkalmazva, csupán egyetlen modalitásból érkező információkra szorítkozna (Esposito–Esposito 2010). Kutatócsoportunk úgy véli, hogy a társalgás komputációs megragadásának és a dialógusrendszerek modellezésének érdekében érdemes modalitásonként külön-külön explicitté tenni az egyes kommunikatív jelenségek gépi eszközökkel is detektálható felszíni jegyeit.

Az általunk a multimodális annotáció mellett, attól különböző eljárásként alkalmazott unimodális pragmatikai annotációtól azt várjuk, hogy lehetőséget biztosítson arra, hogy az ember–ember és az ember–gép kommunikációt egyaránt új megvilágításba helyezzük. Ennek megfelelően a jelenleg folyó annotálás során új megközelítést alkalmazva különválasztjuk az információs csatornákat (percepciósan bármilyen nehéznek is tűnik mindez). Célunk a kommunikációs esemény bizonyos aspektusainak megragadása pusztán vizuális, vagy pusztán akusztikus input alapján. Az unimodális annotáció mögött meghúzódó feltevésünk az, hogy a kommunikációs eseménynek vannak olyan összetevői, amelyek pusztán vizuális eszközökkel is megragadhatóak, és egyetlen modalitás markerei is kifejezhetnek bizonyos kommunikatív funkciókat, ráadásul akár olyanokat is, amelyek bizonyos mértékben különböznek azoktól, amelyeket ugyanezen modalitás más modalitásokkal való kombinációjában kifejez.

Az annotáció során bár markereket annotálunk, közvetlenül nem egy-egy tekintetmintát vagy fejmozdulatot stb. keresünk, hanem kommunikatív funkciókat, és ezekhez azonosítjuk az őket megvalósító markereket (Hunyadi 2011a). Azaz ezen funkciókat valamilyen markerek segítségével azonosítjuk, de a kiindulás mégis a vélt kommunikációs esemény megragadása, ellentétben az eddigi videoannotációk fő vonulatával, ahol lényegében előre meghatározott fizikai markerek (pl. arckifejezések, kézmozdulatok, testtartás) jelenlétét annotáltuk, függet-

[9] Ahogy McNeill fogalmaz: A gesztusok „a beszélt nyelvvel időben, jelentésükben és funkciójukban oly szorosan összefonódnak, hogy a beszélt nyelvi megnyilatkozásokat és a gesztusokat akár ugyanazon mögöttes mentális folyamat különböző oldalainak is tekinthetnénk" (McNeill 1995, 1).

lenül azok funkciójától. Nem gondoljuk azt, hogy unimodálisan teljes biztonsággal meg tudjuk határozni ezeket a funkciókat, ezeket a multimodális verifikálás felül is bírálhatja. Megközelítésünk a kommunikáció formális alapszerkezetének megismerésére irányul, úgy, hogy ezáltal egy tetszőleges kommunikációs esemény formális modellen alapuló technológiai létrehozása lehetővé váljék (vö. Hunyadi 2012). Ahhoz, hogy a nem verbális modalitások formális leírása egységes modellben legyen kezelhető, előbb külön-külön kell az egyes modalitásokat megvizsgálni. A Hunyadi-modell (2011b) lehetővé teszi a verbális és a nem verbális kommunikáció modalitásonként különválasztott absztrakt és felszíni jegyeinek együttes technológiai reprezentálását, ezzel elősegítve a több kommunikációs csatorna által közvetített információ egyidejű feldolgozását. Unimodális annotációnk tehát a fenti gondolatmenetet követve a kommunikáció absztrakt és felszíni jegyeinek rendszerbe foglalásához, ezen belül különösen a beszélőváltás és az egyetértés/nem-egyetértés gépileg is detektálható jegyeinek szisztematikus leírásához kíván hozzájárulni.

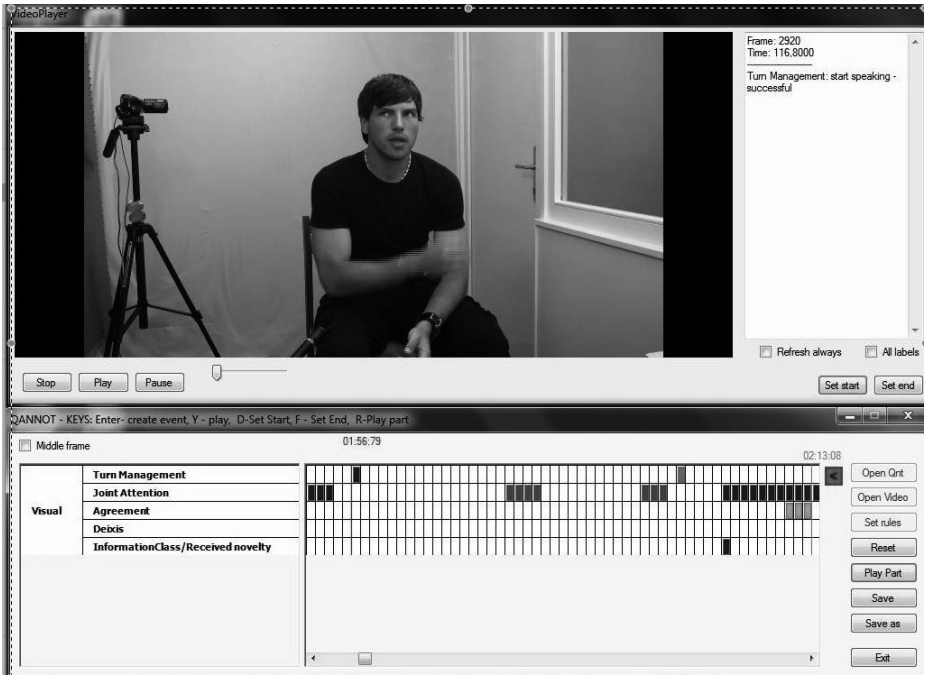
7.2. Az annotáció folyamata

Az annotáció folyamata gyakorlatilag a szemlélt kommunikációs esemény időbeli határokkal *(timestamps)* való ellátása. Ezen határokon belül egy legördülő menüből kiválaszthatjuk a mozdulathoz vagy a gesztusként értelmezhető mozdulatsorhoz illeszkedő címkét, és jelöljük az adott esemény időbeli terjedelmét (l. 10. ábra). A különböző típusú eseményekhez rendelt színkódolt sávok egymással időben szinkronizáltak, így szinkrón természetű (nem szekvenciális), ún. együttállásra vonatkozó címkelekérdezéseket is lehetővé tesznek.

Unimodális annotációs rendszerünk előnye, hogy olyan nyelv- és kultúrafüggetlen, univerzális kategóriákkal dolgozik, mint például a beszélés kezdete és vége által határolt társalgási forduló, vagy az unimodálisan is jól megragadható egyetértés/nem egyetértés gyakori kommunikatív aktusa. Mindezek a kategóriák a beszélők korától, társadalmi státuszától és a szituációtól függetlenül is invariáns részét képezik mindenféle társalgásnak, ami lehetővé teszi unimodális sémánk bármely nyelvű spontán beszéden vagy multimodális korpuszon való alkalmazását.

Az unimodális annotáció történhet vagy csak vizuális, vagy csak audió input alapján is. Itt az előbbiről szólunk részletesebben.

Vizuális-unimodális annotációnk első áttekintése megerősíti, hogy pusztán vizuális input (gesztikuláció, szájmozgás) alapján is meg tudjuk állapítani, hogy ki az aktuális beszélő fél az interakcióban. A megnyilatkozások időtartamának



10. ábra. Az unimodális annotáció felhasználói felülete a Qannot programban

megoszlása alapján azt is meg tudjuk állapítani, hogy mennyire kiegyensúlyozott az interakció a résztvevő felek között. Továbbá, csupán egyetlen modalitás alapján a másik félre fordított figyelem és a beszélgetésbe vonódás mértékét is meg tudjuk állapítani. Sőt, a nem verbális viselkedés vizuális jegyeiből kitűnik az is, amikor a beszélőnek csak az a szándéka, hogy megkezdje a mondandóját, de nem tudja elindítani azt, mert félbeszakítják, vagy az, amikor valamilyen közelebbről nehezen azonosítható okból nem folytatja, esetleg újrakezdi a beszélést.

A fentieknek megfelelően az unimodális annotáció a következő szinteken történik (az angol elnevezések az annotáció valóságos címkéinek felelnek meg):[10]

1. a fordulókezelés szintje *(Turn Management Class),* amelyen belül megkülönböztetjük a beszélés kezdetét a beszélés végét, a közbevágást és a beszédkezdés szándékát;
2. a figyelem szintje *(Attention Class),* amelyen belül megkülönböztetjük a közös figyelembe vonódást és a figyelemfelkeltést;
3. az egyetértés szintje *(Agreement Class),* amely kétféle attribútumot, pozitívat vagy negatívat vehet fel, és azon belül is megkülönbözteti a teljes egyetértést, a részleges egyetértést,

[10] A HuComTech-projekt kutatásának egésze angol nyelven dokumentált, így annak annotációs sémája is nemzetközi, angol nyelvű terminológiát követ.

az egyetértés alapértelmezett esetét, a bizonytalanságot; a nem-egyetértés alapértelmezett esetét, a társalgás elvágásának/blokkolásának szándékát és az érdektelenséget/közönyt;

4. a deixis szintje *(Deixis Class),* amelyben olyan deiktikus viselkedés kerül rögzítésre, amelyre a videoannotáció megfelelő szintje nem tér ki;
5. az információ szintje *(Information Class),* amelyben új információ észlelését rögzítjük.

A HuComTech-kutatócsoport által alkalmazott angol nyelvű unimodális annotációs terminológia rövid összegzése a következő (l. 1. táblázat):

1. táblázat. Az unimodális annotációi szintjeinek és címkéinek áttekintő táblázata

Unimodális annotáció osztályai ***(classes)***		**Unimodális annotáció címkéi** ***(attribútumai)***
Turn Management Class (fordulókezelés/beszélőváltás)		start speaking successfully (sikeres beszédkezdés, beszélőváltás)
		breaking in (közbevágás)
		intending to start speaking (beszédkezdés szándékának kifejezése, de nem sikeres beszédkezdés
Attention Class (figyelem)		calling attention (figyelemfelhívás)
		paying attention (figyelem kifejezése)
Agreement Class (egyetértés)	+agreement (egyetértés)	default case of agreement (egyetértés alapértelmezett esete)
		full agreement (teljes egyetértés)
		partial agreement (részleges egyetértés)
		uncertainty (bizonytalanság)
	–agreement (nem egyetértés)	default case of disagreement (nem egyetértés alapértelmezett esete)
		blocking (társalgás elvágása)
		uninterested (közönyös)
Deixis Class (deixis)		other (egyéb, a videoannotáció szintjén nem jelölt deixis)
Information Class (információ-struktúra)		received novelty (észlelt új információ)

Nemcsak kommunikatív, hanem komputációs szempontból (Bunt–Black 2000) is érdemes a fordulót (*turn;* l. unimodális sémánk első szintjét) a kommuniká-

ció alapegységének tekinteni, hiszen vizuálisan és akusztikusan is jól körülhatárolható, többségében univerzális jelenségek kísérik a forduló átadását, így a beszélőváltás számítógépes eszközökkel is jól detektálható. A fordulókezelés a társalgásban a küldő szerep jogának a beszélők közötti elosztását jelenti. A fordulókezelés és a beszélőváltás gépileg is detektálható jegyeinek (nyelvfüggetlen és nyelvfüggő verbális, vizuális és nem verbális akusztikus jegyeinek) explicit felfedése és szisztematikus rendszerbe (pl. döntési fába) foglalása elsődleges céljaink közé tartozik. Egyik kérdésünk az, hogy a beszélgetőpartnerek hogyan osztják ki maguk között a szó átvételének jogát, illetve hogyan, milyen vizuális és akusztikus jelek (markerek) alapján ismerik fel a beszélőváltásra alkalmas pillanat elérkezését. Ez a kérdés akkor is eldönthető, ha a Qannot videokép alatti hangerőszabályozóján elnémítjuk a felvételt (vagyis eldönthető unimodálisan is), ugyanis látható, mikor kezdi el a beszélő a beszélést és mikor hagyja abba (ráadásul kitűnik, hogy a beszélés kezdetének vizuális markere gyakran hamarabb jelenik meg az akusztikus markernél). Annotációink lekérdezésével többek között azt is meg tudjuk állapítani, hogy átlagosan milyen hosszúságú észlelt szünet jelenti a forduló átadását, és ezzel milyen nem verbális jelenségek járnak együtt. Unimodális annotációnk eredményeinek segítségével további kihívásaink közé tartozik olyan jól körülhatárolható és gépi eszközökkel is megragadható kommunikatív viselkedés automatikus felismerése, mint például az egyetértés és a nem egyetértés kommunikatív aktusa. Eredményeinkkel remélhetőleg hozzájárulhatunk a beszédtechnológia és a dialógusrendszer-modellezés egyik feladatának, a társalgási fordulóvég, más szóval a lehetséges beszélőváltási pont predikciójának megoldásához is, amely predikció alapját kell hogy képezze mindenféle természetes menetű, gördülékeny, időben szinkronizált kérdés–válasz, vagy bármely egyéb kommunikációs szekvenciát követő ember–gép kommunikációnak.

7.3. Várható eredmények

Az unimodális annotáció befejezésével és kiértékelésével a következő elméleti és empirikus eredményeket várjuk elérni:

1. a kommunikációs esemény szerkezetének pontosabb feltárása;
2. a fordulókezelés és beszélőváltás jellemzőinek pontosabb, modalitásonkénti megragadása;
3. az interakciót irányító, diskurzust szervező verbális akusztikus jelenségek, a nem verbális akusztikus és a vizuális viselkedés további részleteinek és azok összefüggéseinek a megismerése;
4. hozzájárulás egy beszélőváltást előrejelző program betanításához.

A következő pontban a különféle információs csatornákból érkező információkat együttesen figyelembe vevő pragmatikai annotáció, a multimodális pragmatikai annotáció bemutatására kerül sor.

8. A HuComTech korpusz pragmatikai szempontból: multimodális annotáció

8.1. A multimodális pragmatikai annotáció célja

A HuComTech korpusz multimodális pragmatikai annotációjának célja kettős. Az elmélethez kötődő cél felfedni a hétköznapi személyközi kommunikáció mögöttes, hierarchikus és szekvenciális szerkezeti sajátosságait, amelyek a kommunikációs események strukturálásában alapvető szerepet játszanak. Ezt a célt úgy tudjuk megvalósítani, hogy a kommunikatív viselkedésekben rejlő verbális akusztikus, nem verbális akusztikus, valamint vizuális jegyeket, Hunyadi (2011a; 2012) terminológiájában markereket, kommunikatív funkciójuk szerint azonosítjuk, valamint az azonosított markereket korreláltatjuk egymással (illetve a többi annotációs szint releváns címkéivel).

8.2. Az annotációs rendszer

A multimodális pragmatikai annotáció alapját a kommunikatív aktusok képezik. A kommunikatív aktusok multimodális illokúciós aktusokként értelmezendők, mivel a verbális közlesek mellett a gesztusokat és a nem verbális akusztikus információkat is figyelembe vesszük az annotáció során. A társalgás szerkezetében a fordulók a legkarakteresebb egységek (és a fordulók kommunikatív aktusokból állnak), valamint a kommunikatív aktusok képesek a beszélő hétköznapi vágyainak és szándékainak kifejezésére is, ezért kézenfekvő volt az annotációs rendszer kommunikatív aktusokra történő alapozása. Tipológiánk kidolgozásához a Bach- és Harnish-féle rendszert választottuk ki (Bach 2006). E megkülönböztetés melletti érveinket részletesen tárgyaljuk Abuczki (2011)-ben. A kommunikatív aktusok típusai:

- konstatívok *(constatives)* = ítélkezők: válaszadás, megerősítés, informálás, predikció, visszaemlékezés
- direktívák *(directives)* = végrehajtók: kérés, parancs, javaslattétel
- komisszívok *(commissives)* = elkötelezők: beleegyezés (pl. egy fogadásba), fölajánlás, ígéret
- viselkedők *(acknowledgements):* üdvözlés, búcsúzás, elfogadás (pl. meghívásé)
- nem azonosítható *(none)*

A négy kommunikatívaktus-típus közül a konstatívok olyan aktusokat foglalnak magukban, amelyek a beszélőnek egy propozicionális tartalomhoz fűződő hiedelmét fejezik ki. Mégpedig úgy, hogy a beszélő mindeközben szándékozza azt is, hogy az aktus propozicionális tartalmát feldolgozza és elhiggye a hallgató is (Abuczki 2011). A direktívek olyan aktusokat tartalmaznak, amelyeknek propozicionális tartalma a hallgató egy elvárt/preferált jövőbeli cselekedetére vonatkozik, s amelyek kifejezik a beszélő azon szándékát, hogy a hallgató a szóban forgó aktus hatására tegye meg a cselekedetet (uo.). A komisszívok olyan aktusok, amelyek a beszélő azon szándékát fejezik ki, amellyel elkötelezi magát egy jövőbeli aktus megtételére (uo.). Végezetül a viselkedők olyan aktusok, amelyek a beszélő valamilyen affektív, érzelmi, attitűdbeli viszonyulását fejezik ki a hallgató iránt (uo.). Léteznek azonban olyan esetek is, amikor egy fordulóban nem azonosíthatók a kommunikatív aktusok imént tárgyalt típusai. Ez az eset akkor áll fenn, amikor az adott forduló nem tartalmaz olyan információt, amely fogódzóként szolgálhatna a fent ismertetett négy típusba való besorolásához. Ezeket a részeket a „none" (nem azonosítható) címkével jelöljük az annotáció során.

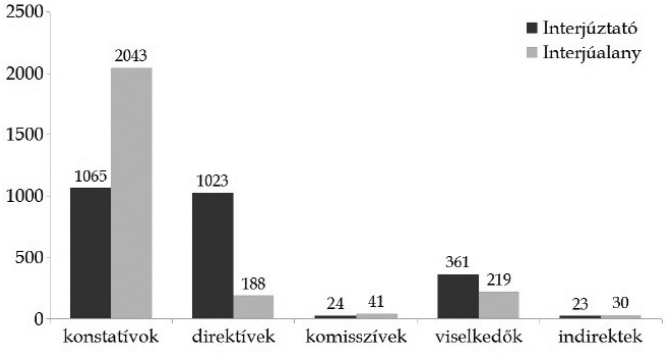
A kommunikatív aktusok mellett az úgynevezett támogató aktusokat is annotáljuk. Ezek az aktusok kiegészítik, támogatják a velük egységben szereplő kommunikatív aktust. A tematikus kontroll szintjén azt vizsgáljuk, hogy a társalgás egyes fordulói milyen módon illeszkednek a társalgás egészébe. A fordulók efféle globálisabb vizsgálata megmutatja, hogy a társalgás során az egyes témák miképpen szerveződnek egységekbe, hogyan történik az egyes társalgási témák motivált egymásba fűzése, illetve a motiválatlanság. A társalgási témák motivált egymásba fűzése rávilágít a társalgásbeli együttműködés mintázataira is.

A pragmatikai annotáció utolsó szintjén a társalgás univerzumába kerülő új lexikai információkat jelöltük. Erre azért volt szükség, hogy a későbbiekben megvizsgálhassuk azon hipotézisünket, amely szerint az új információ bevezetése élénkebb, erőteljesebb gesztikulációval jár együtt.

8.3. Előzetes eredmények

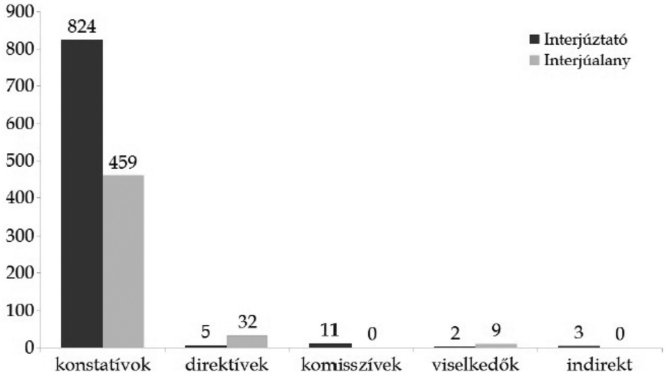
Noha a HuComTech korpusz multimodális pragmatikai annotációja jelenleg is zajlik, előzetes eredményeink jól mutatják, hogy a kommunikatív aktusok, a társalgási fordulók, valamint a fordulókból álló szomszédsági párok közötti viszonyok rendszerszerűek. 35 formális és informális felvétel annotációja alapján láthatjuk, hogy az interjúvezető és az interjúalany szerepeiknek megfelelő kommunikatívaktus-típusokat hoztak létre (11. ábra): az interjúvezető a szcenáriónak megfelelően kb. fele-fele arányban produkál direktív és konstatív aktusokat, míg

az interjúalanyok elsősorban válaszolnak a direktív aktusokra, így az ő konstatív aktusaik száma kiemelkedően magas.



11. ábra. A különböző kommunikatívaktus-típusok előfordulásainak száma

Előzetes címkelekérdezéseink azt is mutatják, hogy a támogató aktusok közül a visszajelzések *(backchannelek)* alapvetően a konstatív aktusokkal járnak együtt (12. ábra).

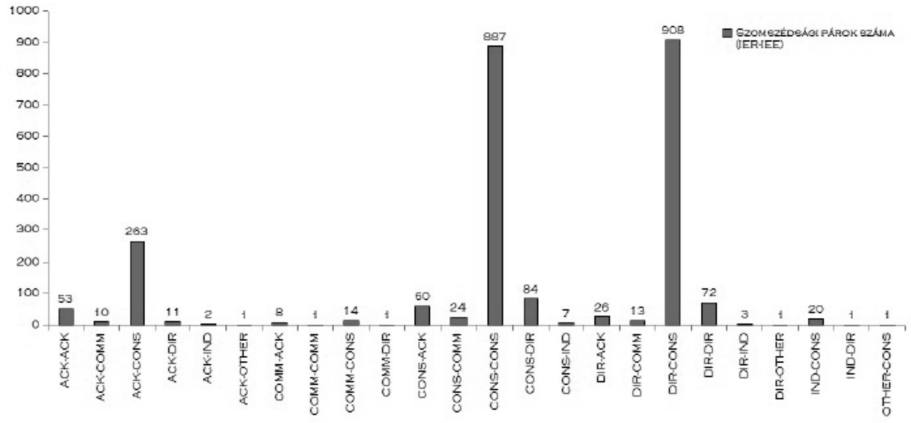


12. ábra. A különböző kommunikatívaktus-típusokra adott visszajelzések száma

Ennek oka minden bizonnyal az, hogy a konstatív aktusok esetében megjelenik egy propozicionális tartalom, amelynek megértéséről a hallgató biztosítja a beszélőt. Fontos azonban látni, hogy a visszajelzések funkciói ennél sokkal szerteágazóbbak, így további kutatások alapját kell képezniük. A 12. ábrán azt is láthatjuk, hogy alapvetően az interjút készítő személy a gyakoribb visszajelző fél a társalgásban. Ennek oka egyrészről az lehet, hogy a felvételek forgatókönyvében

az interjúalany gyakrabban kerül a társalgás középpontjába (aktív kommunikátori szerepben), mint az interjúkészítő fél.

Megvizsgáltuk az egyes szomszédsági párokat alkotó fordulók kommunikatív aktusait is (13. ábra). Az egyes fordulók párokba rendeződésének két kiemelkedő mintázatát figyelhetjük meg: a konstatív–konstatív, valamint a direktív–konstatív együttállást. Ez a kétféle mintázat alapvetően strukturálja a korpusz diskurzusszerkezetét, s jól tükrözi a társalgás mögött meghúzódó forgatókönyvet: az interjúkészítő feladata az információ kérése, az interjúalany feladata az információ adása (ez történik a direktív–konstatív együttállás során). A konstatív–konstatív együttállás valószínűsíthetően egy-egy társalgási téma kidolgozásában játszik fontos szerepet, ám a korpusz jelenlegi feldolgozottsága mellett ilyen típusú lekérdezést még nem tudunk végrehajtani.



13. ábra. A kommunikatív aktusok szomszédsági párba rendeződése

8.4. Kitekintés

A HuComTech korpusz multimodális pragmatikai annotációs rendszerének több előnye is van más annotációs rendszerekhez képest:

1. A rendszer nyelvfüggetlen, univerzális kategóriákkal dolgozik. Mind a kommunikatív aktusok típusai, mind a támogató aktusok, mind a tematikus kontroll tulajdonságai univerzális jellemzői a társalgásnak.

2. Az annotáció során együttesen, multimodálisan vesszük figyelembe a vizuális, a nem verbális akusztikus, valamint a verbális akusztikus információkat. Mivel annotációnk kevés lexikai információn alapul, lehetővé válhat a későbbiekben az információk automatikus kinyerése.

3. A címkézés lehetővé teszi a későbbiekben, hogy egy-egy típust önmagában hívjunk le, és szükség eseten tovább finomítsunk. Ez a megoldás gazdaságos.

4. A fordulók mint strukturális elemek és a kommunikatív aktusok típusai mint funkcionális elemek együttes szerepeltetése lehetővé teszi, hogy a fordulókból kibontakozó szomszédsági párokhoz is tudjunk megfelelő kommunikatívaktus-típusokat rendelni. Ez közelebb vihet bennünket olyan predikciók megtételéhez, amelyeknek a segítségével anticipálni tudjuk a következő fordulót a társalgásban.

Ennek a rendszernek a hatékonyságát jól jelzik előzetes eredményeink, amelyek empirikus adatokat szolgáltatnak a társalgásbeli szerepek jellemzőinek fölfejtéséhez, így hozzájárulva a forgatókönyvvel kapcsolatos predikciók megtételéhez.

9. Összefoglalás

A HuComTech korpusz létrehozását eredetileg egy elméleti cél elérése motiválta: az, hogy létrehozzuk az ember–gép kommunikáció olyan technológiai modelljét, amely alapvetően épít az ember–ember kommunikáció lényeges és e feladat szempontjából releváns jellemzőire. A feladat ilyen megközelítését a bevezetőben indokoltuk.

A 2. pontban javaslatot tettünk az ember–gép kommunikáció újszerű elméleti–technológiai modelljére. A szemléletében generatív, főbb jellemzőiben moduláris, szekvenciális és multimodális modell olyan primitívek halmazát tételezi fel, amelyek, vagy amelyeknek műveletekkel létrehozott képződményei valamilyen módon felszíni alakzatot is öltenek ún. markerek formájában. A modell lényeges feladata ugyanakkor, hogy olyan keretet biztosítson, amely nemcsak a kommunikáció elméleti leírását adja, hanem lehetővé teszi a technológiai megvalósítást is. Ehhez szükség van arra, hogy a markerek tulajdonságait a technológia számára elérhető módon leképezzük.

Az annotáció feladata, hogy általa a elméleti modell építőköveit, a markereket feltárjuk. Így az annotáció során markereket (vagy bizonyos szintjein a markerek előfordulásának funkcionális–pragmatikai interpretációit) keresünk és azonosítunk. Az annotáció azon túlmenően, hogy alátámasztja és elősegíti a kommunikáció felépítésének a megismerését, alapjául szolgál a technológiai megvalósításnak is. Ezt a modell technológiai interfésze biztosítja azáltal, hogy

a modell elméleti komponensének markerekben megjelenő kimenetét a modell technológiai komponensében paraméterek értékeinek felelteti meg. A modell fontos jellemzője, hogy kétirányú, azaz egyaránt szolgálja a szintézist (egy kommunikatív esemény technológiai megvalósítását) és az analízist (ezen esemény interpretációját, „megértését"). Lehetővé teszi e két, ellentétes irányú folyamat egyidejű kezelését, ami által alkalmassá válik egy ember–gép kommunikáció kétirányú folyamatának egységes kezelésére.

A tanulmány további pontjaiban az annotálás különböző lényeges aspektusait mutattuk be, ismertetve az annotálás folyamatát és az adatbázis lekérdezése alapján már elérhető bizonyos eredményeit.

A 3. pontban ismertettük a videoannotálás elveit és az annotálás során megjelölt főbb kommunikációs jegyeket. Bemutattuk, milyen mértékben befolyásolhatja az egyes jegyek (markerek) multimodális együttállása a kommunikatív esemény interpretációját.

A 4. pontban az akusztikus információ kódolásáról szóltunk. Kitértünk a manuális kódolás egyes eredményeire, és az általa felvetett, a további feldolgozásra hatással levő kérdéseit is indokoltuk, valamint részleteztük az automatikus prozódiai annotálás munkálatait. Az így kinyert prozódiai jellemzők birtokában lehetővé válik a vizsgált kommunikációs események gépi detektálásának vagy akár predikciójának az algoritmizálása, ami fontos lépés lehet nyelvtechnológiai alkalmazások továbbfejlesztésében.

A 5. pont esettanulmánya azt mutatta be, hogyan alkalmazható a multimodális annotáció a kommunikációs események jobb megértésére. Nem szándékos ismétlések és az erre következő tartalmas szavak kiejtése során megfigyeltük a csatlakozó tekintetviselkedést és gesztikulációt. Bemutattuk, hogy a beszélők az ismétléseket követő tartalmas szavak memóriából történő előhíváshoz, illetve nyomatékosításához jellemzően olyan vizuális elemeket használnak, mint a tekintet irányával való manipulálás, kézi gesztikuláció és fejmozgások. Az esettanulmány során láthatóvá vált egyes gesztusok és verbális kifejezések produkciója közötti összefüggés. Ezek az ismeretek ugyancsak jól hasznosíthatóak lehetnek a kommunikáció bizonyos mozzanatainak gépi felismerésében és értelmezésében.

A 6. pont a beszélt nyelv szintaxisának rendszerszerű lejegyzését célul kitűző újszerű szintaktikai annotálás elveit mutatta be számos, az írott nyelv szintaxisa által nehezen kezelhető eset bemutatásával. Ettől az annotálástól azt várjuk, hogy egyrészt jobban megértsük a beszélt nyelv szintaktikai szerveződését, másrészt hozzájáruljunk a szintaxisra támaszkodó nyelvtechnológiai alkalmazások továbbfejlesztéséhez.

A 7. és a 8. pontban a pragmatikai annotálás két párhuzamos, egymástól független rendszerét mutattuk be. Az unimodális annotálás (7. pont) célja, hogy

pragmatikailag értelmezhető kommunikatív tulajdonságokat ismerjünk fel csupán egyetlen, mégpedig a vizuális csatorna jelei alapján. Első eredményeink azt mutatják, hogy – ellentétben a természetesnek tűnő elvárással – az ilyen unimodális megközelítés során is lehetővé válik bizonyos kommunikatív funkciók felismerése, ami különösen fontos lehet egy-egy esemény modalitásonkénti technológiai megvalósítása számára. A multimodális pragmatikai annotáció (8. pont) is jelentős újdonságot mutat. Előnyként értékelhetjük más annotációs rendszerekhez képest, hogy nyelvfüggetlen, univerzális kategóriákkal dolgozik, és kevés lexikai információn alapul, amelyek a későbbiekben tovább finomíthatók. A kommunikatív aktusok típusai és a funkcionális elemek együttes szerepeltetésével lehetővé válik a társalgás következő fordulójának az anticipálása.

Összegzésképpen tehát megállapíthatjuk, hogy az annotált korpusz alapján létrehozott – és rövidesen közvetlenül is elérhető – adatbázis célja a kommunikáció rendszerszerű megismerésén túlmenően olyan eszköz nyújtása a nyelvtechnológusok és más szakemberek számára, amely lehetővé teszi a multimodális emberi viselkedés jobb és rendszerszerű megértését, valamint olyan alkalmazások létrehozását, amelyek az ember–gép kommunikáció szerteágazó területein – a humán felhasználónak az eddigieknél érezhetően emberközelibb környezetet biztosítva – saját hatékonyságát jelentősen növelheti.

Irodalom

Abuczki Ágnes 2011. A multimodális interakció szekvenciális elemzése. In: Németh T. (2011, 119–144).

Alessandro, Christophe d' – Piet Mertens 1995. Automatic pitch contour stylization using a model of tonal perception. Computer Speech and Language 9: 257–288.

Bach, Kent 2006. Speech acts and pragmatics. In: Michael Devitt – Richard Hanley (szerk.): Blackwell guide to philosophy of language. Malden MA & Oxford: Blackwell. 147–156.

Bernsen, Niels Ole – Laila Dybkjær 2005. Natural and multimodal interactivity engineering – Directions and needs. In: Kuppevelt et al. (2005, 1–22).

Bódog Alexa (szerk.) 2011. Az ember–gép kommunikáció technológiájának elméleti alapjai. IKUT zárókötet. Debrecen: Debreceni Egyetemi Kiadó.

Boersma, Paul – David Weenink 2005. Praat: Doing phonetics by computer. (Version 5.1.43)

Bunt, Harry – William Black 2000. The ABC of computational pragmatics. In: Harry Bunt – William Black (szerk.): Abduction, belief and context in dialogue: Studies in computational pragmatics. Amsterdam & Philadelphia: John Benjamins. 1–46.

Chomsky, Noam 1965. Aspects of the theory of syntax. Cambridge MA: MIT Press.

Chomsky, Noam 1977. Essays on form and interpretation. New York: North Holland.

Chomsky, Noam 1986. Knowledge of language: Its nature, origin and use. New York: Praeger.

Cole, Ronald A. – Joseph Mariani – Hans Uszkoreit – Annie Zaenen – Victor Zue (szerk.) 1997. Survey of the state of the art in human language technology. Cambridge: Cambridge University Press.

Dix, Alan J. – Janet E. Finlay – Gregory D. Abowd – Russell Beale 2003. Human–computer interaction. 3rd edition. Englewood Cliffs, NJ: Prentice Hall.

Esposito, Anna – Antonietta Maria Esposito 2010. On speech and gestures synchrony. In: Anna Esposito – Alessandro Vinciarelli – Klára Vicsi – Catherine Pelachaud – Anton Nijholt (szerk.): Lecture notes in computer science. Berlin: Springer. 252–272.

Flanagan, James L. 1997. Overview – Multimodality. In: Cole et al. (1997, 329–342).

Fox, Barbara – Makoto Hayashi – Robert Jasperson 1996. Resources and repair: A cross-linguistic study of syntax and repair. In: Elinor Ochs – Emanuel A. Schegloff – Sandra A. Thompson (szerk.): Interaction and grammar. Cambridge: Cambridge University Press. 185–237.

Földesi András 2011. Unimodális funkcionális annotáció a HuComTech multimodális korpuszban. In: Bódog (2011, 40–46).

Gósy Mária 2002. A megakadásjelenségek eredete a spontán beszéd tervezési folyamatában. Magyar Nyelvőr 126: 192–204.

Gósy Mária 2008. A zaj hatása a beszédre. Beszédkutatás 2008: 5–21.

Grice, H. Paul 1957. Meaning. Philosophical Review 67: 377–388.

Grice, H. Paul 1975. Logic and conversation. In: Peter Cole – Jerry L. Morgan (szerk.): Syntax and semantics, vol. 3: Speech acts. New York: Academic Press. 41–58.

Gyarmathy Dorottya 2011. A multimodális interakció szekvenciális elemzése. In: Németh T. (2011, 119–144).

Hunyadi László 2011a. A multimodális ember–ép kommunikáció technológiái – elméleti modellezés és alkalmazás a beszédfeldolgozásban. In: Németh T. (2011b, 15–42).

Hunyadi László 2011b. Az ember–gép kommunikáció elméleti-technológiai modellje. Háttér és alapkérdések. In: Bódog (2011, 6–12).

Hunyadi, László 2012. Multimodal human–computer interaction technologies – Theoretical modeling and application in speech processing. Argumentum 7: 240–260.

Jakobson, Roman 1969. Nyelvészet és poétika. In: Roman Jakobson (szerk.): Hang – jel – vers. Budapest: Gondolat Kiadó. 211–258.

Jong, Nivja H. de – Tor Wempe 2009. Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods 41: 385–390.

Kenesei István 2000. Szavak, szófajok, toldalékok. In: Kiefer Ferenc (szerk.): Strukturális magyar nyelvtan 3. Morfológia. Budapest: Akadémiai Kiadó. 75–136.

Keszler Borbála (szerk.) 2000. Magyar grammatika. Budapest: Nemzeti Tankönyvkiadó.

Komlósy András 1992. Régensek és vonzatok. In: Kiefer Ferenc (szerk.): Strukturális magyar nyelvtan 1. Mondattan. Budapest: Akadémiai Kiadó. 299–527.

Kuppevelt, Jan C. J. van – Laila Dybkjær – Niels Ole Bernsen (szerk.) 2005. Advances in natural multimodal dialogue systems (Text, Speech and Language Technology 30). Dordrecht: Springer.

Lewis, David K. 1969. Convention. Cambridge MA: MIT Press.

Mariani, Joseph 1997. Multimodality. In: Cole et al. (1997, 329–370).

McNeill, David 1995. Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press.

Mertens, Piet 2004. The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In: Bernard Bel – Isabelle Marlien (szerk.): Proceedings of the 2nd International Conferemce of Speech Prosody, Nara, 23–26 March 2004.

Németh Zsuzsanna 2012. A javítási műveletek interakciós funkciói: ismétlés és csere a magyarban. Beszédkutatás 2012: 154–167.

Németh T. Enikő 1996. A szóbeli diskurzusok megnyilatkozáspéldányokra tagolása (Nyelvtudományi Értekezések 142). Budapest: Akadémiai Kiadó.

Németh T. Enikő 2011a. A humán kommunikáció modelljei és az ember–gép kommunikáció. In: Németh T. (2011b, 43–62).

Németh T. Enikő (szerk.) 2011b. Ember–gép kapcsolat. A multimodális ember–gép kommunikáció modellezésének alapjai. Budapest: Tinta Könyvkiadó.

Oviatt, Sharon L. 2003. Multimodal interfaces. In: Julie A. Jacko – Andrew Sears (szerk.): The human–computer interaction handbook: Fundamentals, evolving technologies and emerging applications. Mahwah, NJ: Lawrence Erlbaum. 286–304.

Pápay Kinga – Szeghalmy Szilvia – Szekrényes István 2012. HuComTech multimodal database annotation. Argumentum 7: 330–347.

Sacks, Harvey 1995. Lectures on conversation. Cambridge MA & Oxford: Blackwell.

Sacks, Harvey – Emanuel A. Schegloff – Gail Jefferson 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50: 696–735.

Schegloff, Emanuel A. 2006. Sequence organization in interaction: A primer in conversation analysis. Cambridge: Cambridge University Press.

Shannon, Claude – Warren Weaver 1949. The mathematical theory of communication. Urbana, IL: University of Illionis Press.

Sperber, Dan – Deirdre Wilson 1986/1995. Relevance: Communication and cognition. Cambridge, MA & Oxford: Blackwell.

Szekrényes István – Csipkés László – Oravecz Csaba 2011. A HuComTech-korpusz és -adatbázis számítógépes feldolgozási lehetőségei. Automatikus prozódiai annotáció. In: Tanács Attila – Vincze Veronika (szerk.): A VIII. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 190–198.

Thórisson, Kristinn R. 2008. Modeling multimodal communication as a complex system. In: Ipke Wachsmuth – Manuela Lenzen – Günther Knoblich (szerk.): Springer lecture series in computer science: Modeling communication with robots and virtual humans. New York: Springer. 143–168.

Tóth Csilla 2011. Tekintetmintázatok és funkcióik a HuComTech-projekt szimulált állásinterjúiban. In: Németh T. (2011b, 101–118).

Wahlster, Wolfgang 1991. User and discourse models for multimodal communication. In: Joseph W. Sullivan – Sherman W. Tyler (szerk.): Intelligent user interfaces. New York: ACM Press. 45–67.

Wahlster, Wolfgang (szerk.) 2006. SmartKom: Foundations of multimodal dialogue systems. New York: Springer.

A theoretical–technological model of human–computer interaction and its implications for language technology

Abstract: With its goal to contribute to a more humanlike human-machine interaction, the HuComTech project aims at studying and describing those aspects of human-human communication that are supposed to be of primary relevance for our interaction with machines. Whereas language is undoubtedly the most important prerequisite of communication, it is by far not the only one: gestures, gaze patterns, head and body movements as well as ways of speech production equally contribute to a successful communication being essentially multimodal. The paper gives an overview of this interdisciplinary approach to communication presenting a novel two-way, generative model of communication at the intersection of theory and technology and offering the first observations and results based on 60 hours of audio-video recordings. In particular, it describes principles of the video annotation (and observations on facial expressions and their alignment with other multimodal features), principles of the audio annotation (and observations on speech tempo, intonation, repetitions), principles of the annotation of the syntax of speech as well as of the uni- and multimodal pragmatic annotation. The paper also includes a case study of observed gaze patterns and their communicative functions.

Keywords: human–computer interaction, multimodality, multimodal corpus, annonation, prosody, syntax, pragmatics

Kísérletek beszédfelismerők akusztikus modelljének nyelvek közötti átvitelére

Tóth László

MTA-SZTE Mesterséges Intelligencia Kutatócsoport, Szeged
tothl@inf.u-szeged.hu

A gépi beszédfelismerők akusztikai komponense hagyományosan a beszédhangokat modellezi. Emiatt a modellek nyelvek közötti közvetlen átvitele csak akkor lehetséges, ha a két nyelv hangkészlete legalább részben átfed. Ezzel szemben elvileg nyelvfüggetlen modellezést ígér az a megközelítés, amely beszédhangok helyett fonológiai megkülönböztető jegyek detektálására törekszik a beszédfolyamban. Ebben a cikkben arra teszünk kísérletet, hogy egy angol nyelvre betanított jegydetektáló kimenetét használjuk fel a magyar beszédhangok felismerésében. Az összehasonlítási alapként szolgáló rendszerek egyikében szintén angol nyelvű modellekből indulunk ki, de jegyek helyett beszédhangok modelljeit igyekszünk a magyar hangkészlethez igazítani, valamint készítettünk egy tisztán magyar nyelvű adatokon tanított rendszert is. A felismerési eredmények nem igazolják azt a várakozásunkat, hogy a jegydetektáló rutin nyelvfüggetlensége miatt jobb, de legalábbis nem rosszabb eredményt képes nyújtani, mint a tisztán magyar modell. Ennek lehetséges okait vitatjuk meg a cikk záró részében.

Kulcsszavak: gépi beszédfelismerés, megkülönböztető jegyek, multilingvális beszédfelismerés, keresztnyelvi beszédfelismerés, mesterséges neuronhálók

1. Bevezetés

A modern gépi beszédfelismerő rendszerek alapvetően két fő részre bonthatók, egy akusztikai és egy nyelvi komponensre. Előbbi feladata a beszédjelben bizonyos absztrakt percepciós osztályokat megtalálni és azonosítani, azaz az akusztikus jel és bizonyos fonetikai-fonológiai címkék kapcsolatát modellezni. A nyelvi komponens pedig a javasolt címkesorozatokat támogatja vagy elveti annak függvényében, hogy azok a vizsgált nyelvben valószínűsíthető szavakat-szósorozatokat alkotnak-e vagy sem. Jelenleg mind az akusztikai, mind a nyelvi modellek matematikai, gépi tanulási elveken működnek, azaz viszonylag kevés nyelvészeti szakértelmet igényelnek, de – főleg a nyelvi komponens esetén – elég valószínűnek látszik, hogy a rendszerek további javításához nyelvspecifikus (morfológiai, szintaktikai, szemantikai) tudás bevitelére lesz szükség. A gépi tanuláson alapuló komponensek sem nyelvfüggetlenek, mivel statisztikai alapon működnek, ami azt jelenti, hogy hatalmas mennyiségű, az adott nyelvből származó tanító adaton

kell őket betanítani. Emiatt hatalmas tanító korpuszokat kell készíteni – jelenleg minden egyes nyelvre külön-külön. Ez a rendkívül műveletigényes fázis a nyelvi modellezés szintjén elkerülhetetlennek tűnik; azonban az akusztikai-fonetikai szinten nem hamvába holt ötlet közös – sőt, akár univerzális – építőelemeket keresni. Ezeket kihasználva a világnyelvekre elkészített akusztikus modellek teljesen, vagy legalább részben átvihetőek lennének a kevésbé kutatott nyelvekre, így egy „új" nyelvre nem kellene a beszédkorpusz építését a nulláról kezdeni, vagyis a fejlesztést gyorsabbá és olcsóbbá lehetne tenni. Jelen cikkünkben két, az akusztikus modellek nyelvek közötti „átültetését" célzó módszert hasonlítunk össze. A témában további irodalmat keresőknek Schultz és Kirchhoff (2006) áttekintő kötetét ajánljuk kiindulópontnak.

Jelenleg a beszédfelismerők akusztikai komponense szokásosan a beszédhangokat modellezi. Ebből következően a modellek nyelvek közötti közvetlen átvitelére akkor van esély, ha a két nyelv hangkészlete legalább részben átfed. Ilyenkor az adott hangok modelljei használhatók a másik nyelvű felismerőben is, vagy például a két nyelven külön-külön gyűjtött tanító példák egyesíthetők, mindkét nyelv modellépítési fázisát támogatva ezzel. Ezen az ötleten alapulva próbál „nyelvfüggetlen" beszédfelismerést elérni a SPICE nemzetközi projekt, amely a weben keresztül gyűjt interaktív módon hanganyagot lényegében az IPA fonetikai táblázat teljes hangkészletéhez (Schultz et al. 2007).

A beszédhangok modellezésénél általánosabbnak ígérkezik az a megközelítés, amely a beszédhangok helyett a fonológiai megkülönböztető jegyek jelenlétének detektálására törekszik a beszédfolyamban. A jegyek detektálásán alapuló felismerés ötlete időről időre újra felbukkan a beszédtechnológiában, és több elméleti előnyt is kínál a beszédhang-alapú megközelítéshez képest, de igazából sosem sikerült teret nyernie. Az egyik feltételezett előny, hogy mivel a megkülönböztető jegyek jóval univerzálisabbak és kevesebben vannak, mint a beszédhangok, így – elvileg – ezekre építve jóval könnyebb és hatékonyabb nyelvfüggetlen akusztikus modellt készíteni.

Írásunkban angol nyelvre betanított rendszerekből készítünk magyar nyelvű akusztikus modellt, ahol az eredeti, angol felismerő az egyik esetben beszédhangok, a másik esetben megkülönböztető jegyek felismerésére van betanítva. A 2–4. pontban röviden ismertetjük a beszédhang-alapú, illetve a jegyalapú modellezés működését, összevetjük azok előnyeit és hátrányait. Az 5–6. pontban bemutatjuk az angol nyelvű modellekből kialakított magyar nyelvű rendszer működését és hatékonyságát. A kísérletekben a jegyalapú modell nem bizonyul jobbnak, mint a beszédhang-alapú modell. Ezért a diszkusszióban részletesen elemezzük a két rendszer tévesztési viselkedését. Hangcsoportonként megvizsgáljuk, hogy mely hangokra működött jobban az egyik rendszer, melyekre a másik,

és hogy ez a mintázat korrelál-e a két nyelv hangkészletének hasonlóságaival és különbözőségeivel. Megvitatjuk továbbá a jegyalapú rendszer kudarcának lehetséges okait is.

2. Beszédhang-alapú akusztikus modellezés

A beszédfelismerési technológia az akusztikum modellezésében a hagyományos lineáris szemléletből indul ki, azaz feltételezi, hogy minden egyes azonosítandó címkéhez tartozik egy időben viszonylag egyértelműen behatárolható jelszakasz (szegmentum). A szegmentumok azonosításában nincs élesen szétválasztott fonetikai és fonológiai szint, azaz a címkézéshez alapvetően az adott nyelv fonémakészletéből szokás kiindulni, de gyakran allofónok is külön címkét kapnak, valamint arra is akad példa, hogy fonémákat is egybevonnak (azt feltételezve, hogy nyelvi szinten könnyebb lesz őket szétválasztani, mint akusztikai szinten). Így a kiindulási címkekészlet 40–50 körüli elemből szokott állni, melyekre az egyszerűség kedvéért „beszédhang" címkeként fogunk hivatkozni. Ezután statisztikai elven működő gépi tanulási algoritmusokat (általában ún. Gauss-keverékmodellt, ritkábban mesterséges neuronhálót, mindkettőről l. Bishop 2006) tanítunk be arra, hogy a beszédjel egy adott pillanatáról megmondják, hogy milyen beszédhanghoz (címkéhez) tartozhat, és milyen valószínűséggel. Az egyes pillanatokhoz rendelt címke-valószínűségeket az ún. rejtett Markov-modell (Huang et al. 2001) segítségével fűzzük össze. Ez a matematikai eszköz segít megtalálni a jel legvalószínűbb szegmentálását, és az egyes szegmentumok legvalószínűbb címkéjét. A címkékből szintén lineáris módon, azaz címkesorozatok megadásával definiálunk szavakat (ezekből áll az ún. kiejtési szótár).

Elég csak néhány beszédjelet közelebbről megvizsgálnunk ahhoz, hogy rádöbbenjünk, a linearitás elve mennyire tarthatatlan. Artikulációs szerveink működésének nyilvánvaló fizikai korlátai miatt a szomszédos beszédhangok még a leggondosabb artikuláció esetén is részben összemosódnak. A koartikulációt[1] a beszédtechnológia úgy próbálja kezelni, hogy a szegmentumokat tovább bontja, szokásosan három szakaszra: a középső szakasz hivatott leírni a hang viszonylag stabil (izolált ejtéshez közelítő) ejtési fázisát, a két szélső pedig a szomszédos hangokba való hangátmeneti fázisokat. Továbbá, mivel az egyes hangok akusztikai képe erőteljesen függ a szomszédos hang mibenlététől, ezért az időtengely

[1] Az akusztikai modellezés szintjén alapvetően a fonetikai koartikulációt fogjuk koartikuláción érteni. A fonológiai koartikulációt (hiátustöltés, hasonulás, fonémakiesés) (Gósy 2004) a kiejtési szótár definiálásakor szokás figyelembe venni.

mellett a címkekészletet is tovább bontjuk: a három ejtési fázis más-más elnevezést kaphat a vizsgált és a szomszédos hang függvényében. Ezt a technológiát trifón-modellezésnek nevezzük, az így előálló, finomított címkehalmaz elemeit pedig szenonoknak (Huang el al. 2001). A trifón-technikával a rejtett-Markov-modelles felismerés hatékonysága jelentős mértékben növelhető. Hátránya, hogy a címkék megnövekedett száma miatt a szükséges betanító korpusz méretét is növelni kell, hogy minden lényeges, hangkörnyezettől függő ejtésvariánsra kellő mennyiségű előfordulás jusson. A mai rendszerekben a szenonok száma jellemzően öt- és tízezer között mozog, betanításukhoz pedig minimálisan több 10 órányi időtartamú korpusz kell, de inkább a 100 órás nagyságrendet szokás szükségesnek ítélni (egy magyarhoz hasonló „kis" nyelv esetére l. pl. Alumäe 2005).

3. A beszédhang-alapú modellezés gyengeségei

Vegyük észre, hogy a fent ismertetett felbontási trükkel csak a felismerési alapegységeken finomítottunk, de a linearitási feltevést nem vetettük el: a felismerőrendszer továbbra is feltételezi, hogy az egyes szavak bizonyos egységek előre meghatározott sorozataként állnak elő, és hogy az egyes egységekhez egy-egy meghatározott jelszakasz tartozik. A beszédfelismerési kísérletek szerint gondosan artikulált, például olvasott beszéd esetén a koartikuláció ilyen egyszerű, alapvetően csak a szomszédos hangokat figyelembe vevő kezelése elégségesnek nevezhető: a piacon kapható diktálórendszerek – a beszélő hangjához való adaptáció után – gyakorlatilag is használható felismerési pontosságot tudnak elérni. Mi magunk magyar nyelvű, hangoskönyveken végzett kísérleteinkben a beszédhangok sorozatát nyelvi támogatás nélkül is 86%-os pontossággal tudtuk felismerni, és a felismerő fonetikai szintű kimenete szabad szemmel is olvashatónak bizonyult (Tóth 2009). Általános tapasztalat azonban, hogy spontán beszéd esetén a felismerők hatásfoka drasztikusan leromlik. A jelenség megértése céljából végeztek olyan vizsgálatokat, ahol egy megbeszélésen felvett hanganyagot utólag újraolvastattak a résztvevőkkel. Az olvasott és a spontán felvételeken mért felismerési hiba között közel kétszeres faktort kaptak (Weintraub et al. 1996). Magyar nyelvre Mihajlik és társai próbálkoztak spontán és tervezett beszéd (hírműsorok) ugyanazon technológiával való felismerésével (Mihajlik et al. 2009). Habár az eredmények nem precízen összemérhetők, hiszen a két feladat közt a beszédmódon kívül más eltérések is voltak, az általuk kapott bő kétszeres hibatényező is jól érzékelteti, hogy milyen jelentős hatékonyságromlás lép fel spontán beszéd esetén. A felismerési hibák részletes elemzése nyomán több kutató is arra a következtetésre jutott, hogy a beszédhang-alapú ejtésmodellezés nem biztos, hogy

alkalmas a spontán beszéd leírására – még a finomított, trifónokat használó megoldással sem (Ostendorf 1999). Az alapvető problémát az okozza, hogy spontán beszédben gyakran lépnek fel olyan redukciós jelenségek, amelyek több hangon is átívelnek, így nem modellezhetők két szomszédos hang közötti átmeneti fázisok segítségével. Vegyük például azt az esetet, amikor a *tonhal* szó ejtésekor az [n] kiesik – legalábbis olyan értelemben, hogy nem rendelhető hozzá szegmentum. Azonban nem tűnik el teljesen nyomtalanul, mivel a megelőző magánhangzót nazalizálja (Siptár–Törkenczy 2000). Ilyenkor az sem jó megoldás, ha a kiejtési szótárban a szó leírásakor meghagyjuk az [n]-t: a felismerő keresni fogja a hozzá tartozó jelszakaszt, pedig ilyen nincs. De az sem jó, ha kihagyjuk: ekkor a rendszer nem fogja tudni, hogy nem normál ejtésű [o]-t kell keresnie, így a nazalizált változatot nem fogja szeretni. További finomításokkal persze lehet kezelni ezeket a problémákat – például minden „trükkös" jelenségre külön-külön modell bevezetésével –, de egyáltalán nem biztos, hogy ez a célravezető megoldás. Sokak szerint inkább radikálisan meg kell változtatni a teljes akusztikus modellezést. Egyesek nagyobb egységek, pl. a szótag felé lépnének: ezt azok az (angol nyelvre végzett) mérések motiválják, melyek szerint a szótagon belüli pozíció befolyásolja a koartikuláció működését (Greenberg 1999). A népesebb irányzat szerint viszont a beszédhangnál kisebb egységekkel kellene dolgozni – amire a nyelvészet a megkülönböztető jegyeket kínálja megoldásként.

4. Megkülönböztető jegyeken alapuló modellezés

A megkülönböztető jegyes fonológia szerint a fonéma megkülönböztető jegyek (hangtulajdonságok) együttesen jelentkező halmaza (Hall 2001; Péter 2001). Jakobson, Fant és Halle (1952) eredeti munkájában mindössze 12 jeggyel írta le a fonémákat (Parsons 1986), melyet később 16-ra bővítettek (Kassai 1998). A megkülönböztető jegyes ábrázolás legkorábbi magyar nyelvre vonatkoztatott kidolgozása Szépe György nevéhez köthető (Szépe 1969). Fontosnak tarjuk kiemelni, hogy az eredeti szerzők akusztikai vizsgálatokra építve definiálták a jegyeket, azaz a többségükhöz azt is megadták, hogy mely spektrális jellegzetességekből lehet őket azonosítani. Későbbi munkák azonban artikulációs irányba vitték el a jegyeket, amivel beszédfelismerési szempontból az a probléma, hogy ezek beszédjelből való detektálhatósága sok esetben egyáltalán nem nyilvánvaló. A jegyek száma is jelentősen megugrott, és sokszor eltérő szerzők eltérő jegykészletekkel dolgoznak. Az eredeti elmélet „felpuhult" olyan értelemben is, hogy a kiinduló felvetés szerint a jegyek binárisak, azaz csak két lehetséges értékük van (egy adott pillanatban vagy jelen vannak, vagy sem). Később azonban sokan megengedtek egy

harmadik, „irreleváns" értéket is, vagy azt, hogy bizonyos jegyek különböző mértékben (fokozatokban) is jelentkezhessenek.

A beszédfelismerésben a megkülönböztető jegyeken alapuló akusztikai modellezés ötlete évtizedek óta jelen van (l. pl. Bocchieri–Wilpon 1993; Espy-Wilson 1994; Hansen 1997; Kirchhoff et al. 2000; Stüker et al. 2003; Metze 2005), de nem tud betörni a kutatás fősodrába. Támogatói fő érvként azt szokták fölhozni, hogy a jegyek segítségével könnyebben lehetne kezelni a koartikulációs jelenségeket. Ehhez csak azt kell megengedni, hogy az egyes jegyek ne teljesen szinkronban kapcsoljanak be és ki a szegmentumok határain, hanem átterjedhessenek a szomszédos jelszakaszokra. Ily módon a különféle lehetséges átmeneti ejtési fázisokat sokkal kompaktabban le lehet írni a jegyek kombinációjaként, mint ha ezek mindegyikére külön címkét aggatunk, majd külön gépi tanulási modellt tanítunk be rájuk. Ráadásul a jegyek segítségével nem csak az átmeneteket lehet természetes módon kezelni, hanem az olyan, lineáris módon nem felírható eseteket is, mint az előző pontban felhozott szegmentum-kieséses példa. Ennek jegyes felírásakor nem kellene bajlódni a nazalizált magánhangzó külön modellezésével, hiszen a szokásos magánhangzókat leíró jegyek mellett egyszerűen csak a nazalizáció jegyének „bekapcsolásával" le lehetne kódolni a problémás szegmentumot. További fontos érvként szokták említeni azt is, hogy a jegyek elvileg nyelvfüggetlenek, Jakobson például azt állította, hogy 16 jeggyel leírható a világ nyelveinek fonémaállománya (Kassai 1998). Ezért az utóbbi években a multilingvális beszédtechnológiai kutatások terjedésével a jegyalapú modellezés lehetősége újra fókuszba került, és alapvetően emiatt kísérletezünk vele mi is.

Mint említettük, a jegyeken alapuló felismerési technológiával csak kevesen próbálkoznak, és rendszereik inkább csak kísérleti jellegűek. Emiatt egységesen elfogadott elv sincsen arra nézve, hogy hogyan kellene jegyekre épülő felismerőt készíteni. Abban eléggé egyetért az irodalom, hogy a legtöbb jegy jelenléte jó hatékonysággal felismerhető gépi tanulási algoritmusokkal. Innentől azonban sokféle megoldás kínálkozik az akusztikus modell felépítésére. A legegyszerűbb, ha az eredeti nyelvészeti definíciót követve a fonémák jelenlétét a jegyek egyidejű jelenlétével definiáljuk. Technikailag ez a megoldás azért egyszerű, mivel ily módon a jegyek beépíthetők a hagyományos, fonéma-jellegű egységekkel dolgozó felismerőkbe (mi is ezt fogjuk tenni, technológiai részletek a következő pontban). Vegyük észre azonban, hogy így tulajdonképpen visszajutunk a korábbi, lineáris modellhez, azaz a jegyalapú felírásnak éppen azt a tulajdonságát – a hangátmenetek és egyéb összeolvadások könnyű felírása – nem aknázzuk ki, ami az előnyét jelentené. Cikkünkben viszont alapvetően nem a koartikulációs jelenségek kezelése miatt próbálkozunk a jegyalapú modellezéssel, hanem a feltételezett nyelvfüggetlenséget szeretnénk megvizsgálni. Ezért csupán a teljesség kedvéért említ-

jük meg, hogy a jegyek optimális felhasználásához valószínűleg a teljes akusztikai modellezést újra kell majd gondolni. Egyrészt a jelenlegi, linearitásban gondolkozó módszerek nem képesek az önálló szegmentumként nem jelentkező, de a ráutaló jegyekből mégis megfejthető hangok kikövetkeztetésére (l. nazalizációs példa). Másrészt sok pszicholingvisztikai vizsgálat utal arra, hogy az emberi beszédértésben a mentális lexikon aktiválásához nem feltétlenül szükséges teljes beszédhangok azonosítása, hanem bizonyos akusztikai kulcsok megfelelő kombinációban való jelenléte is elegendő (Marslen-Wilson–Warren 1994). Mindkét tényező azt sejteti, hogy feltehetőleg a beszédfelismerők szokásos, beszédhangok lineáris sorozataként megadott „kiejtési szótárát" is ki kell majd dobni, és más megoldásra cserélni. Jelenleg azonban még az emberi beszédpercepció működéséről is viták folynak, és az azt imitálni szándékozó számítógépes modellek is meglehetősen gyerekcipőben járnak (ezekről jó áttekintést ad – a mérnöki megoldásokkal párhuzamba állítva – Scharenborg et al. 2005).

5. Artikulációs jegyek felismerése mesterséges neuronhálókkal

Vizsgálatainkban angol nyelvre betanított jegydetektáló algoritmusok segítségével próbálunk majd magyar nyelvű akusztikus modellt építeni. Mindezt az motiválja – a jegyek feltételezett nyelvfüggetlensége mellett –, hogy rendelkezésünkre áll egy olyan jegyfelismerő rendszer, amelyet kétezer (!) órányi angol nyelvű beszéden tanítottak be. Összehasonlításképp, a magyar felismerőt az MTBA adatbázison fogjuk tanítani, amely csupán 7 óra hanganyagot tartalmaz (Vicsi et al. 2002). Mint korábban említettük, a beszédfelismerő rendszerek megfelelő betanításához több tíz, de inkább száz órányi hanganyag szükséges, ezért joggal feltételezhetjük, hogy a két nyelv közti eltérések ellenére a jóval alaposabban betanított angol rendszer hatékonyabban fog működni a magyar nyelvű felismerésben, mint a magyar tanítású rendszer – főleg, ha a jegyfelismerés nyelvfüggetlensége is teljesül.

A felhasznált jegyfelismerő rendszer nemzetközi összefogással készült a Johns Hopkins Egyetem 2006-os nyári workshopján (Frankel et al. 2007). Ennek során mesterséges neuronhálókat tanítottak be az egyes jegyek kinyerésére 2000 órányi angol nyelvű, telefonon keresztül rögzített beszéden. A neuronhálókat úgy állították be, hogy kimeneteiken az egyes jegyek – pontosabban az egyes lehetséges jegyértékek – adott pillanatban való jelenlétének erősségét (valószínűségét) adják vissza egy 0 és 1 közé eső érték formájában. A rendszerben nyolc jegyet

definiáltak, melyek jórészt többértékűek; a jegyek mindegyikére egy-egy önálló neuronhálót tanítottak be. A jegyeket és a lehetséges értékeiket az 1. táblázatban mutatjuk be. A jegyek és értékeik megválasztása sok esetben önkényesnek tűnik, a magánhangzókat például egyrészt egyáltalán nem bontották jegyekre (a „magánhangzó" jegy lehetséges értékei konkrét fonetikai címkék), másrészt mégis, mert a „nyelvállás" és a „nyelv vízszintes helyzete" jegyek csak a magánhangzók kedvéért vannak felvéve (más hangoknál „semleges" értéket kapnak). Sajnos az ismertető cikkből nem derül ki, hogy a jegyek és értékeik ilyen megválasztását mi motiválta, viszont a cikk végén kapunk egy táblázatot, melyből legalább azt megtudhatjuk, hogy az egyes fonetikai címkéknek milyen jegyértékeket feleltetnek meg a szerzők (Frankel et al. 2007). A jegyek felismerésére betanított neuronhálók szabadon letölthetők az edinburghi egyetem vonatkozó weboldaláról (Frankel 2006).

Artikulációs jegy	A jegy lehetséges értékei
képzés helye	bilabiális, labiodentális, dentális, alveoláris, posztalveoláris, veláris, glottális, rotikus, laterális, semleges
képzés módja	magánhangzó, approximáns, legyintőhang, réshang, zárhang
nazalitás	+, –
fonáció	zöngés, zöngétlen, hehezetes
ajakkerekítés	+, –
magánhangzó	aa, ae, ah, ao, aw1, aw2, ax, ay1, ay2, eh, er, ey1, ey2, ih, iy, ow1, ow2, oy1, oy2, uh, uw, nem magánhangzó
nyelvállás	legfelső, felső, középmagas, középső, alsó-középső, alsó, semleges
nyelv vízszintes helyzete	hátsó, mediális-hátsó, mediális, mediális-elülső, elülső, semleges

1. táblázat. A felismerőben használt artikulációs jegyek és értékkészletük

Kísérleteinkben arra voltunk kíváncsiak, hogy a jegyalapú modellek vajon valóban jobban átvihetők-e a magyar nyelvre, mint a beszédhang-alapúak. Ennek kiderítéséhez szükségünk volt egy viszonyítási alapra, vagyis egy beszédhangokon betanított angol nyelvű modellre. Ehhez az Edinburghi Egyetem kutatói bocsátottak rendelkezésünkre egy további neuronhálót, amelyet 46 beszédhang-jellegű címke felismerésére tanítottak be egy másik projekt során, több száz órányi hanganyagon (sajnos nem ugyanazon, mint amelyet a jegydetektáló hálók esetén használtak).

6. A felismerőrendszer és a felismerési eredmények

A beszédfelismerési technológia matematikai részletekbe menő ismertetésétől megkíméljük az olvasót. Az alkalmazott ún. tandem technikáról részletesen olvashat Hermansky és munkatársainál (2000), az általunk használt konkrét paraméterértékekről és egyéb beállításokról pedig egy korábbi konferenciacikkünkben (Tóth et al. 2008). A megértéshez annyit kell tudni, hogy a tandem rendszer két lépésben alkalmaz gépi tanulást (innen a „tandem" név). Az első lépésben egy mesterséges neuronháló a beszéd minden egyes pillanatához megállapítja, hogy egy előre meghatározott jellemzőkészlet egyes tagjai mennyire vannak jelen (tipikusan 0 és 1 közötti értékeket visszaadva). Ebben a lépésben tulajdonképpen bármiféle olyan absztrakt „jegyet" megpróbálhatunk kinyerni a neuronhálóval, amelyet hasznosnak ítélünk az egyes beszédhangok megkülönböztetése céljából, teljesen mindegy, hogy azt nyelvészeti, mérnöki, vagy milyen okból véljük információhordozónak.

A második lépésben egy hagyományos, rejtett Markov-modelles felismerőt futtatunk a neuronháló által „kibányászott" jellemzőkön. Mint a 2. pontban elmondtuk, ez a lépés beszédhangoknak (vagy azok finomított változatainak, a trifónoknak) megfelelő szegmentumok sorozataként próbálja dekódolni a bemenetét, vagyis esetünkben a neuronhálótól kapott jellemzőértékeket. Kísérleteinkben ezt a második rendszert mindig ugyanúgy fogjuk tanítani: az MTBA magyar nyelvű adatbázis 52 fonetikai címkéjének felismerésére fogjuk „kérni", amihez az adatbázis 4/5 részét bocsátjuk a rendelkezésére a tanuláshoz, a fennmaradó 1/5 részen pedig tesztelünk. Amit variálni fogunk, az az első lépésben, a neuronháló által végzett jellemzőkinyerés. Mivel a második, beszédhang-felismerő rész mindig ugyanaz lesz, ezért a felismerési eredmények eltérése azt fogja megmutatni, hogy mely neuronhálós jellemzőkészlet a legjobb, azaz melyik segíti leginkább a magyar beszédhangok felismerését. Háromféle konfigurációt fogunk összevetni. Az egyikben az **artikulációs jegyekre, angol nyelvre** betanított hálót fogjuk használni. A második esetben az **angol nyelvű beszédhangok** jelenlétét detektáló háló szolgáltatja a bemenetet. Végül a harmadik kísérletben az MTBA-t használva készítünk egy **magyar beszédhangokat** detektáló neuronhálót, és azt használjuk inputként. A felismerési eredmények összehasonlításának célja tehát annak vizsgálata, hogy mely jellemzőkből könnyebb magyar beszédhangokat felismerni: nyelvfüggetlen (?) artikulációs jegyek alapján, angol beszédhangok alapján, vagy magyar beszédhangok alapján. Elsőre azt gondolhatja az olvasó, hogy természetesen a tisztán magyar nyelvű rendszer lesz a legjobb, a két nyelv hangkészletének eltéréseiből adódóan. Emlékeztetnünk kell azonban arra, hogy az angol rendszerek (mind a hang-, mind a jegyalapú) jóval nagyobb adatbázison tanul-

tak, így feltehetően pontosabbak, mint a magyar. Ezért nem lehetetlen az az elsőre abszurdnak ható lehetőség, hogy mivel az angol beszédhangokat jobban fogja felismerni a vonatkozó neuronháló, így a második lépésben a magyar hangkészletet is jobban sikerül „kikombinálni" az általa adott kimenet alapján, még a két nyelv hangtani eltéréseinek ellenére is. Másképp fogalmazva, a nyelvek közötti különbségből adódó hátrányt kompenzálhatja a nagyobb mennyiségű tanítóanyagból adódó előny. Hogy melyik nyom többet a latban, azt nehéz előre megjósolni. Azt viszont egyértelműen kijelenthetjük, hogy a legjobb eredményeket a jegyalapú neuronhálóra épülő rendszertől várhatjuk, hiszen egyrészt az volt a legnagyobb mennyiségű adaton tanítva, másrészt az elvileg nyelvfüggetlen, így a tanítási és tesztelési nyelv eltérésének sem szabad visszavetnie a teljesítményét.

A 2. táblázatban a három rendszer által elért beszédhang-felismerési pontosságot hasonlíthatjuk össze.[2]

Bemeneti jellemzők	Felismerési pontosság
artikulációs jegyek	57,93%
angol beszédhangok	59,02%
magyar beszédhangok	62,08%

2. táblázat. A három rendszer által elért felismerési pontosság

7. Elemzés

A kapott eredmények több szempontból is csalódást keltőek. Először is, egyik angol nyelvről átültetett modell sem érte el a tisztán magyar tanítású modell teljesítményét. Így tehát nem teljesült az a reményünk, hogy a nagy mennyiségű adaton tanított angol modellekből kiindulva megúszhatjuk, hogy magyarra is hasonló hatalmas korpuszokat kelljen összegyűjtenünk. Még nagyobb csalódás, hogy a jegyalapú rendszer nemhogy a magyar modell teljesítményét nem haladta meg, de még a beszédhang-alapú, angolról átültetett modellnél is rosszabbul teljesített. Azaz nemcsak hogy nem viselkedett nyelvfüggetlenként, hanem még a beszédhangoknál is nyelvfüggőbbnek bizonyult.

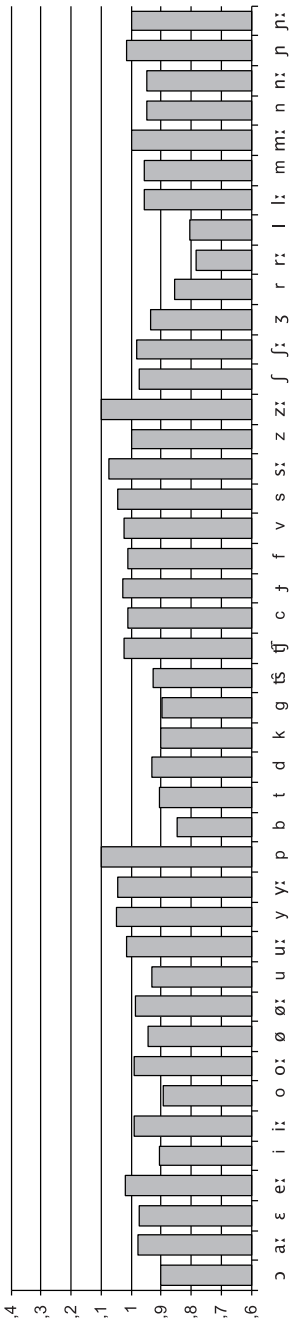
[2] Más cikkeinkben (Tóth et al. 2008; Tóth 2009) az itt közölteknél jobb eredményeket fog találni az olvasó, mivel ott további trükköket is bevetünk (például spektrális jellemzőket is keverünk az inputba, illetve fón-bigramokat használunk, amelyek egyfajta fonotaktikai modellként foghatók fel). Jelen írásban azonban nem a minél pontosabb felismerés, hanem a jegyalapú modellezés, illetve a nyelvek között modellátvitel hasznosságának felmérése a cél.

A nagyobb rálátás érdekében a felismerési eredményeket az egyes beszédhangokra külön-külön lebontottuk, és így is összevetettük, ez látható az 1. és 2. ábrán (l. 322. o.). A könnyebb összehasonlítás kedvéért a következő ábrákon mindig két rendszer felismerési pontosságának **hányadosát** fogjuk megjeleníteni. Az 1. ábrán például a beszédhang-alapú angol modell és a magyar modell beszédhangonként vett pontosságát vetjük össze, melynek értelmezése tehát a következő: 1 alá eső értékek esetén a magyar modell teljesített jobban, 1 fölé eső értékek esetén az angol; és természetesen minél alacsonyabb (magasabb) egy oszlop, annál nagyobb volt a különbség a magyar (angol) rendszer javára.

A grafikonok elemzése előtt meg kell magyaráznunk, hogy bizonyos mássalhangzók hosszú változata miért nem szerepel. Az [f], [v] és [ʒ] hangok esetén ennek technikai oka van: e hangok hosszú ejtésben olyan kevésszer fordulnak elő az adatbázisban, hogy önálló statisztikai modellezésüknek nem lett volna értelme. Ide kapcsolódóan meg kell jegyeznünk azt is, hogy a rejtett Markov modellezési technika csak marginálisan veszi figyelembe a szegmentumok hosszát – hiszen a legtöbb nyelvben nincs megkülönböztető szerepe –, ezért a magyar nyelvre tanított rendszerek meglehetősen rosszul teljesítenek a rövid-hosszú párok elkülönítésében, és így sokan eleve egybevonva kezelik ezeket (Mihajlik et al. 2007).

Szintén hiányoznak a hosszú variánsok a felpattanó zárhangok és az affrikáták esetén (a [c, ɟ] hangokat az utóbbiakhoz soroltuk). E hangoknál a zárrészt és a zörejrészt külön modelleztük, és mivel az ejtés hossza csak az előbbi képzési mozzanatot befolyásolja, így a nyújtás jelét csak a zárrészre alkalmaztuk, a zörejeket pedig hossztól függetlenül egységesen kezeltük. Például a [t] hang a mi kódolásunkban [- ˈt], a hosszú párja pedig [-ː ˈt]. A táblázatban szereplő felismerési értékek tehát valójában csak a [ˈt]-vel jelölt zörejfázisra vonatkoznak.

Rátérve az eredményekre, első lépésben a magyar és az angol beszédhang-alapú modelleket vetettük össze. Arra voltunk kíváncsiak, hogy vajon van-e összefüggés az egyes hangokra kapott eredmények és az adott hang „idegensége" között. Ésszerűnek tűnt ugyanis a feltevés, hogy az angolban hiányzó, vagy legalábbis nagyon másként ejtett magyar hangokon fog a legrosszabbul teljesíteni az angolból átültetett modell. Az 1. ábra azonban csak részben támasztja ezt alá. A magánhangzók esetén például azt vártuk volna, hogy az [y] (és méginkább az [yː]) okozza majd a legnagyobb gondot az angolból áttett rendszernek – ezzel szemben éppen e két hangon teljesített a legjobban. A két rendszer tévesztési mátrixát közelebbről megvizsgálva ennek okaként azt találtuk, hogy az angol modell jóval sikeresebben különíti el az [y] hosszú-rövid párjait, mint a magyar. Mint említettük, a hosszokat direkt módon egyik rendszer sem modellezi, így érdemi magyarázatot erre a viselkedésre nem tudunk adni.



1. ábra. A beszédhang-alapú angol modell hangonkénti felismerési pontossága a magyar modellhez viszonyítva

2. ábra. A jegyalapú modell hangonkénti felismerési pontossága a magyar modellhez viszonyítva

A felpattanó zárhangok esetén az angol rendszer sokkal rosszabb volt a magyarnál – kivéve a [p] hangot, melyre viszont látványosan jobban viselkedett. Ennek megértéséhez tüzetesebben megvizsgáltuk a tévesztési mátrixok zárhangokra vonatkozó blokkját, és nagy meglepetésünkre azt találtuk, hogy a fő problémát nem az egymás közti tévesztések okozzák. Ugyanis amennyiben eltalálta a rendszer, hogy felpattanó zárhangról van szó, akkor e csoporton belül már 86%-os pontossággal jól azonosított (az angol rendszernél ez 87%!), amit rendkívül jó eredménynek tartunk. A hibák inkább abból eredtek, hogy a zár- és zörejfázisokat külön modelleztük. A külön egységként kezelt zörej azonban többnyire annyira rövid (más hangok átlagos hosszához képest), hogy a rendszer hajlamos „törölni", azaz a hozzá tartozó szegmentumot hozzácsapni a rákövetkező hanghoz, vagy a megelőző zárfázishoz. Elemzésünk szerint ez már a magyar rendszerben is súlyos teljesítménycsökkenést okoz, de az angolból származtatott modellben még erősebb fokban jelentkezik. Ehhez az is hozzájárulhat, hogy az angol neuronháló tanításakor feltehetőleg nem bontották külön a zár- és zörejrészt (de erre nézve nincs biztos információnk).

Az affrikátáknál rendre 1 körüli értékeket kaptunk, azaz a két rendszer közel ugyanolyan jól teljesített, egyedül a [ʒ] esetén van nagyobb eltérés az angol modell hátrányára. Ennek oka, hogy ez a modell jóval többször azonosítja e hangot [s]-ként, mint a magyar, amit indokolhat a kérdéses hang relatív ritkasága az angolban. Ilyen alapon viszont a [c, ɟ] hangpárra kellett volna a legalacsonyabb értékeket kapnunk, de nem így történt. A [c] esetén végzett részletes vizsgálat azt mutatta, hogy bár ésszerű módon jóval többször soroltatik [t]-nek az angol modell által (hatszor, szemben a magyar 1-évl), a magyar modell viszont más hangok irányába téveszt jóval többet, ami összességében kompenzálja az angol modell [t–c] bizonytalanságát.

A réshangok esetén ismét elég egyenletesen fej-fej melletti a két rendszer teljesítménye, egyedül a [ʒ]-nél teljesít feltűnően rosszabbul az angolból átültetett modell, ami ismét összhangban látszik lenni az adott hang angolbeli ritkaságával. A nazálisokra szintén eléggé szinkronban van a két felismerő, bár itt hipotézisünk szerint az [ɲ] esetén rosszabb viselkedést vártunk volna az angol modelltől.

A felismerési pontosság a likvidák esetén mutatja a legnagyobb eltérést, mégpedig a magyar modell javára. Ezek a hangok eleve nehezen felismerhetőek, az [l] főleg a magánhangzókhoz való hasonlósága, az [r] pedig sokféle lehetséges megjelenése (perdületek száma) miatt. Az [l]-t az angol rendszer jóval többször azonosította magánhangzóként, [r]-ként vagy [v]-ként, mint a magyar. Az [r] esetén is a magánhangzóként, illetve a [v]-ként, [l]-ként való felismerések száma nőtt meg az angol rendszerben. Továbbá, feltehetően a bizonytalan azonosításuk miatt, ezek a hangok feltűnően sokszor estek a korábban ismertetett törlési jelenség áldozatául.

Végezetül, elhagytuk a grafikonról a [j] és [h] hangokat, ezek a két rendszerben közel egyformán viselkedtek.

A jegyalapú rendszerre is kiszámoltuk a magyar rendszerhez viszonyított felismerési mutatókat, a kapott értékeket a 2. ábra mutatja. Az 1. ábrával összevetve azt láthatjuk, hogy a két táblázat oszlopértékei meglehetősen szinkronban mozognak, azaz a jegyalapú rendszer nem csak nem lett nyelvfüggetlenebb, mint a beszédhang-alapú, de azzal egészen analóg módon is viselkedik. Sőt, érdekes módon pont az angol számára idegen vagy ritka hangoknál – pl. [yː, c, ɟ, ʒ, ɲ] – mutat rosszabb teljesítményt, ami megerősíti azt a korábbi benyomásunkat, hogy mintha még nyelvfüggőbb is lenne, mint a beszédhang-modellek átvitelével készített rendszer.

Mi lehet a kudarc oka? Hibás lenne az alapfeltevés, miszerint a jegyek segítségével nyelvfüggetlen akusztikus modellt készíthetünk? Véleményünk szerint az eredményekből nem szabad határozott következtetést levonni a kiindulási koncepcióra nézve, mivel a kivitelezésnek sajnos rengeteg gyenge pontja van. Rögtön az első ilyen a jegykészlet megválasztása. Érdeklődésünkre a cikk szerzői megerősítették azt a gyanúnkat, miszerint a jegykészlet és az egyes jegyek lehetséges értékeinek megválasztásakor csak az angol nyelv hangjainak leírhatóságára koncentráltak, további nyelvek lefedése egyáltalán nem volt szempont (Karen Livescu személyes közlése). A legkirívóbb példa természetesen az angol magánhangzók egy az egyben való bevétele, mint jegyeké, de erre utal például a képzés helyénél a palatális érték hiánya is. További – immár beszédtechnológiai – probléma a jegyek detektálását végző neuronháló betanítása. Erre a gépi tanuló algoritmusra hárul az akusztikai sokszínűség kezelése, azaz a jegyérték helyes felismerése az ejtésvariációtól függetlenül. A jelenlegi algoritmusoktól sajnos nem várható el, hogy egy másik nyelvi, így a tanítás során nem látott ejtésvariánsra is általánosítani tudjanak, hiszen még adott nyelven belül is nehezen boldogulnak. Azt pedig még kevésbé lehet elvárni, hogy olyan jegyeket vagy jegyértékeket is felismerjenek, amelyek a tanító nyelvben nem, vagy nagyon ritkán fordulnak elő (pl. angolban a palatális képzési hely). Ebből az következik, hogy a jegyalapú felismerőket is több nyelven kellene tanítani a nyelvfüggetlen működés eléréséhez. Elsőre úgy tűnhet, hogy ezzel el is vesztettük azt az előnyt, ami miatt a jegyalapú modellezéssel próbálkozni kezdtünk. Gondoljunk azonban bele, hogy a jegyek segítségével sokkal kompaktabban lehet leírni a nyelveket, mint beszédhangokkal, ezért valószínűsíthető, hogy egy jegyalapú nyelvfüggetlen rendszer betanításához jóval kevesebb nyelvű és/vagy nyelvenként jóval kisebb minta is elég lenne.

A betanítás problematikájához tartozik az is, hogy tudomásunk szerint jelenleg nem létezik nagyobb méretű, jegyek szintjén címkézett beszédkorpusz (az angolra sem). Így a jegydetektáló neuronhálók betanításakor azt a stratégiát al-

kalmazták, hogy a beszédhang-szintű címkékből következtettek vissza a jegyekre egy táblázat alapján. Így a rendszer nem tudta figyelembe venni azt a jelenséget, amikor egyes jegyek kisebb-nagyobb mértékben áthúzódnak a szomszédos szegmentumra. Ez okozhatott címkézési hibákat, így ronthatta a detektorok pontosságát (eleve a beszédhang-címkék is automatikus eljárással, így bizonyos mértékben hibákkal terhelten készültek, hiszen 2000 órányi hanganyagot végighallgatni is hónapokig tartana, nemhogy felcímkézni!). A kézi ellenőrzés mellett algoritmikai trükkökkel is lehetne ezen az egyszerű betanítási címkézésen finomítani.

Végezetül, mint korábban említettük, a jegyalapú detektorok igazán hatékony kihasználásához túl kellene lépni a lineáris szemléleten. A legtöbb esetben azonban a kisebb ellenállás irányában haladva nem készítenek speciális, a jegyek ügyes felhasználására kihegyezett dekódolót, hanem egyszerűen egy hagyományos felismerőbe táplálják be a jegyeket a szokásos spektrális jellemzők helyett. Esetünkben is ez történt, azaz a jegyeket beszédhang-címkékké konvertáltuk egy újabb gépi tanulási lépéssel. Nagyon valószínű, hogy jobb eredményeket kaptunk volna egy olyan módszerrel, amely ki tudja használni a jegyek speciális tulajdonságait.

Azt is meg kell még említenünk, mint a következtetések levonásában zavart keltő tényezőt, hogy a gépi tanulási módszerek sokszor az intuíciónak ellentmondó eredményeket adnak, például előfordulhat, hogy jobbnak vélt jellemzőkön rosszabb hatékonysággal működnek. Ilyenkor sokszor nem az alapkoncepcióval van a baj, hanem a tanuló algoritmus paramétereivel, modellválasztásával, optimumkeresési módszerével, vagy egyéb technológiai tényezővel.

Hogy cikkünk végkicsengése mégse legyen teljesen negatív: az angol modellek adaptációjával végzett kísérletezés során végül sikerült olyan megoldást találnunk, amelyben az angol modell segített pontosabb magyar modellt kifejleszteni (Tóth et al. 2008). Ez azonban pusztán technikai okoknak volt köszönhető, és nem a nyelvészeti szaktudás bevonásának. Azt reméljük, hogy az utóbbi is hasznosnak bizonyul előbb-utóbb.

8. Összefoglalás

Írásunkban az artikulációs megkülönböztető jegyek gépi beszédfelismerésben való felhasználására tettünk kísérletet. Bár fő motivációnk a jegyek nyelvfüggetlensége volt, bemutattuk azt is, hogy a jegyalapú modellezés számos szempontból a koartikuláció kezelését is meg tudná könnyíteni. Kísérleteinkben egy 2000 órányi angol nyelvű felvételen betanított, ingyenesen közkinccsé tett jegydetektáló rendszert próbáltunk magyar nyelvű felvételek felismerésében alkalmazni.

Az eredmények sajnos nem igazolták, hogy a rendszer nyelvfüggetlen lenne, és így nagyobb tanítókorpusz gyűjtése nélkül lehetővé tenné más nyelvű felismerők készítését. A hibák részletes elemzéséből arra a következtetésre jutottunk, hogy maga a kivitelezés számos ponton támadható, így a negatív eredmény nem cáfolja a koncepció használhatóságát, pusztán azt mutatja, hogy a kényelmes út, azaz a hagyományos eszköztár feladatra való ráerőszakolása nem vezet célhoz. Úgy véljük, hogy a jó eredmények eléréséhez a jegydetektorok betanításán is finomítani kell, és speciálisan a jegyalapú reprezentációhoz kidolgozott technológiára is szükség lesz.

Irodalom

Alumäe, Tanel 2005. Phonological and morphological modeling in large vocabulary continuous Estonian speech recognition system. In: Margit Langemets – Priit Penjam (szerk.): Proceedings of the Second Baltic Conference on Human Language Technologies. Tallin: Tallin University of Technology. 89–94.

Bishop, Christopher 2006. Pattern recognition and machine learning. New York: Springer.

Bocchieri, Enrico – Jay Wilpon 1993. Discriminative feature selection for speech recognition. Computer Speech and Language 7: 229–246.

Espy-Wilson, Carol 1994. A feature-based approach to speech recognition. Journal of the Acoustical Society of America 96: 65–72.

Frankel, Joe. 2006. Articulatory feature multi-layer perceptrons. Kézirat. http://www.cstr.ed.ac.uk/research/projects/featureMLPs/ (2012.03.09.).

Frankel, Joe – Matthew Magimai-Doss – Simon King – Karen Livescu – Özgür Çetin 2007. Articulatory feature classifiers trained on 2000 hours of telephone speech. In: Proceedings of Interspeech 2007. Antwerp: ISCA. 2485–2488.

Gósy Mária 2004. Fonetika, a beszéd tudománya. Budapest: Osiris Kiadó.

Greenberg, Steven 1999. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. Speech Communication 29: 159–176.

Hall, Tracy Alan (szerk.) 2001. Distinctive feature theory. Berlin & New York: Mouton de Gruyter.

Hansen, Anya Varnich 1997. Acoustic parameters optimized for recognition of phonetic features. In: Proceedings of Eurospeech 1997. Rhodes: ISCA. 397–400.

Hermansky, Hynek – Dan Ellis – Shihab Shamma 2000. Tandem connectionist feature extraction for conventional HMM systems. In: Proceedings of ICASSP 2000. Istanbul: IEEE. 1635–1638.

Huang, Xuedong – Alex Acero – Hsiao-Wuen Hon 2001. Spoken language processing. New Jersey: Prentice Hall.

Jakobson, Roman – Gunnar Fant – Morris Halle 1952. Preliminaries to speech analysis: The distinctive features and their correlates. Cambridge MA: MIT Press.

Kassai Ilona 1998. Fonetika. Budapest: Nemzeti Tankönyvkiadó.

Kirchhoff, Katrin – Gernot Fink – Gerhard Sagerer 2000. Combining acoustic and articulatory feature information for robust speech recognition. Speech Communication 37: 303–319.

Marslen-Wilson, William – Paul Warren 1994. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. Psychological Review 101: 653–675.

Metze, Florian 2005. Articulatory features for conversational speech recognition. Doctoral dissertation, Universität Fridericiana zu Karlsruhe.

Mihajlik, Péter – Balázs Tarján – Zoltán Tüske – Tibor Fegyó 2009. Investigation of morph-based speech recognition improvements across speech genres. In: Proceedings of Interspeech 2009. Brighton: ISCA. 2687–2690.

Ostendorf, Mari 1999. Moving beyond the 'beads-on-a-string' model of speech. In: Proceedings of ASRU 1999. Keystone: IEEE. 79–84.

Parsons, Thomas 1986. Voice and speech processing. New York: McGraw-Hill.

Péter Mihály 2001. Strukturális fonológia. In: Siptár Péter (szerk.): Szabálytalan fonológia. Budapest: Tinta Könyvkiadó. 9–36.

Scharenborg, Odette – Dennis Norris – Luis ten Bosch – James McQueen 2005. How should a speech recognizer work? Cognitive Science: A Multidisciplinary Journal 29: 867–918.

Schultz, Tanja – Alan Black – Sameer Badaskar – Matthew Hornyak – John Kominek 2007. SPICE: Web-based tools for rapid language adaptation in speech processing systems. In: Proceedings of Interspeech 2007. Antwerp: ISCA. 2125–2128.

Schultz, Tanja – Katrin Kirchhoff (szerk.) 2006. Multilingual speech processing. Amsterdam: Elsevier.

Siptár, Péter – Miklós Törkenczy 2000. The phonology of Hungarian. Oxford: Oxford University Press.

Stüker, Sebastian – Florian Metze – Tanja Schultz – Alex Waibel 2003. Integrating multilingual articulatory features into speech recognition. In: Proceedings of Eurospeech 2003. Geneva: ISCA. 1033–1036.

Szépe György 1969. Az alsóbb nyelvi szintek leírása. Általános Nyelvészeti Tanulmányok 6: 359–466.

Tóth László 2009. Beszédfelismerési kísérletek hangoskönyvekkel. In: Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): A VI. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 206–216.

Tóth, László – Joe Frankel – Gábor Gosztolya – Simon King 2008. Cross-lingual portability of MLP-based tandem features – A case study for English and Hungarian. In: Proceedings of Interspeech 2008. Brisbane: ISCA. 2695–2698.

Vicsi Klára – Tóth László – Kocsor András – Gordos Géza – Csirik János 2002. MTBA – magyar nyelvű telefonbeszéd-adatbázis. Híradástechnika 57: 35–43.

Weintraub, Mitch – Kelsey Taussig – Kate Hunicke-Smith – Amy Snodgrass 1996. Effect of speaking style on LVCSR performance. In: Proceedings of ICSLP 1996. Philadelphia: ISCA. 16–19.

Attempts at cross-lingual porting of the acoustic model of speech recognizers

Abstract: The acoustic component of automatic speech recognizers conventionally works with phones as the unit of modelling. Because of this, transferring acoustic models between languages is possible only when the phone sets of the two languages at least partly overlap. In contrast, a modelling method based on the detection of phonological distinctive features would hold the promise of language-independent modelling. In this paper we attempt to utilize a feature detector trained on English data in the recognition of Hungarian speech signals. For comparison, we construct a similar cross-lingual system, but with the detector component being trained on English phones instead of distinctive features, while in a third set-up a purely Hungarian system is created. The experimental results do not meet the expectations that the system built on distinctive features produces a better – or, at least, not worse – cross-lingual recognizer, due to the language-independent behaviour of the feature detector. We discuss the possible reasons of this in the final part of the paper.

Keywords: automatic speech recognition, distinctive features, multi-lingual speech recognition, cross-lingual speech recognition, artificial neural nets

Multifunkcionális beszélt nyelvi adatbázis – BEA*

Gósy Mária

Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Budapest
gosy.maria@nytud.mta.hu

A tanulmány bemutatja a BEszélt nyelvi Adatbázis (BEA) fejlesztésének sajátosságait, jellemző adatait, eredményeit és kutathatóságát. Ez az első olyan magyar nyelvű adatbázis, amely nagy mennyiségű hangzó anyagot tartalmaz (jelenleg több mint 230 óra), adatközlőinek száma 265 (20 évestől 85 évesig), a felvételi körülményei pedig állandóak (azonos felvételi technika, csendesített szoba). Minden beszélővel különböző típusú narratívákat, interpretált beszédet, társalgást, felolvasást és mondatismétléseket rögzítenek. A felvételek különböző szintű átirata, illetve annotálása folyamatban van. Ezek a paraméterek már önmagukban nemzetközileg is a legjelentősebbek közé sorolják a BEA-t.

Kulcsszavak: narratíva, társalgás, felolvasás, adatbázis, annotálás

1. Bevezetés

A számítógépes technológia segítségével létrehozható nagyméretű beszédadatbázisokat a fonetika harmadik forradalmának nevezték a hangszínképelemzés és a számítógépes beszédanalizáló szoftverek után egy 2011-es fonetikai workshop[1] nyitó gondolataként a Pennsylvaniai Egyetemen. Napjainkban már nagy méretű írott és hangzó nyelvi adatbázisok állnak rendelkezésre különböző nyelveken, és ez azt jelenti, hogy a kutatók olyan kérdésekre is választ kaphatnak, amelyekre korábban – megfelelő nyelvi anyag hiányában – nem volt mód. A szöveg filológiai megközelítése ma már nem csupán az írott, hanem a beszélt szöveget is jelenti. A nyelvészet számos területén fokozódik az igény a valós nyelvhasználat megismerésére, tanulmányozására. A hatalmas adatmennyiség feldolgozásában a szabályalapú módszereket a statisztikai módszerek váltják fel, ami szemléletbeli változással is jár.

* A szerző köszönetét fejezi ki a Fonetikai Osztály minden munkatársának az adatbázis fejlesztésében végzett munkájukért. A BEA fejlesztése a 78315. sz. OTKA és a MONTANA projekt támogatásával készül.

[1] New tools and methods for very-large-scale phonetics research: http://www.ling.upenn.edu.phonetics/workshop/

A korszerű beszédadatbázisok rögzített felvételei különféle szempontok szerint strukturáltak és lekérdezhetők. Többségük csak audiofelvétel, de vannak videós rögzítések is (pl. CUAVE: Patterson et al. 2002; Popescu-Belis et al. 2009). Majdnem mindegyik adatbázishoz szövegfájlok is tartoznak, amelyek a rögzített beszédanyag különböző szintű átírásának anyagai. Ezek az adatbázisok többféle módon osztályozhatók, céljuk, tartalmuk, írott változatuk, felvételi körülményeik stb. függvényében (pl. Clark–Fox Tree 2002). A beszédadatbázisok egy része csak olvasott szövegeket, más részük spontán beszédanyagot tartalmaz, és vannak olyanok, amelyekben mindkét típusú beszéd megtalálható. A felolvasások legtöbbször könyvrészletek, rádióhírek, szólisták stb. meghangosításai. A spontán szöveganyagokat laboratóriumban, telefonon át vagy terepen rögzítették, illetve a médiából válogatták; párbeszédek, társalgások, narratívák, élethelyzetek (azok modellálása), játékszituációk, illetve térképmódszerrel rögzített beszéd (pl. Anderson et al. 1991; Hennebert et al. 2000; Ruhi 2011).

A teljesség igénye nélkül röviden bemutatunk néhány jelentős beszédadatbázist. A legnagyobbak egyike a British National Corpus,[2] amely 100 millió szóból álló gyűjtemény (Burnard–Aston 1998); beszélt és írott szövegeket egyaránt tartalmaz. A London–Lund korpusz 50 dialógusból és mindössze 170 000 szóból áll (Svartvik 1990). A CallHome[3] elnevezésű, (eredetileg) amerikai angol korpuszban 120 párbeszéd található, átlagosan 30 percesek, amelyekben családi beszélgetések zajlanak telefonon keresztül. Az elsők között fejlesztették a Kielkorpuszt,[4] amely német spontán beszédanyagokból áll (Simpson et al. 1997). Főként skót angol beszédet tartalmaz az a 62 beszélős, 18 órányi adatbázis (HCRC Map Task), amely a térképmódszert használja (Anderson et al. 1991). Az ausztrál pizzarendeléses korpusz 3 óra 54 perc hosszúságú, egy év alatt 162 megrendelést rögzítettek (Hutchinson–Pereira 2001). A Stanford Egyetem (USA) Switchboard elnevezésű beszédkorpusza (Godfrey et al. 1992; Calhoun et al. 2010) 2400 párbeszédet tartalmaz 543 beszélőtől (az anyagban számos amerikai dialektus van reprezentálva). A személyfüggetlen beszédfelismerők betanítására használatos a TIMIT,[5] amelyhez 630 beszélő 10–10 mondatot olvasott fel (Keating et al. 1994). Beszédtechnológiai céllal fejlesztették a Verbmobil adatbázist (Bael et al. 2007). A BAS *(Bayerisches Archiv für Sprachsignale)*[6] német adatbázis több (al)korpuszból áll, például utcanevek felolvasásából, avagy a taxi diszpécserszolgálatának be-

[2] Részletesen: http://www.natcorp.ox.ac.uk

[3] Részletesen: http://tinyurl.com/busa370

[4] Részletesen: http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html

[5] Részletesen: http://tinyurl.com/at6bzo

[6] Részletesen: http://www.phonetik.uni-muenchen.de/forschung

szélgetéseiből. Tartalmaz továbbá narratívákat, gyermek- és kamasznyelvi, illetve nyelvjárási felvételeket is. Nagy adatbázis a CSJ *(Corpus of Spontaneous Japanese),* amelyben 661 órányi beszédanyag található 1395 beszélőtől (olvasott és különböző spontán beszédanyagok), a szavak száma 7,2 millió (Maekawa 2003). Európa hét nyelvét reprezentáló adatbázis például az EUROM és a BABEL, amelyek célja a beszédakusztikával, fonetikával, digitális jelfeldolgozással, illetve nyelvészettel foglalkozó szakemberek munkájának segítése, különböző olvasott szövegek felvételeivel (Chan et al. 1995; Vicsi 2001). A 22 language Telephone Speech Corpus elnevezésű gyűjtemény jelenleg 50 191 fájlt tartalmaz (magyar anyag is van benne), a beszélők számát nyelvenként 300-ra tervezik (Cole et al. 1995). A fonológiai variációk kutatásának igénye hívta életre a Buckeye amerikai angol társalgási korpuszt (Pitt et al. 2005), amely 307 ezer szót tartalmaz 40 beszélőtől.

Magyar nyelvű beszédgyűjteményt ismereteink szerint elsőként Balassa József hozott létre a 20. század elején, az anyag azonban sajnos megsemmisült (vö. KKA 1994). Az 1940-es években Hegedűs Lajos fonetikus kezdeményezésére indult el a nyelvjárási hangfelvételek készítése azzal a céllal, hogy az ország különböző helyein rögzítsenek beszédet, mesemondást, ráolvasásokat stb.; ezt az anyagot az MTA Nyelvtudományi Intézete archiváltatta (a kilencvenes évek végén), ezáltal korszerű adathordozókon kutatásra alkalmassá vált (Gósy et al. 2011). A hetvenes évek elején Szende (1973) a spontán beszéd gyakorisági tényezőinek elemzéséhez négyféle spontánbeszéd-korpuszt használt fel. Osztálytermi felvételek és más, különféle spontán beszédanyagok készültek (átiratokkal) a hetvenes évek második felében (Keszler 1983). A Budapesti Szociolingvisztikai Interjú (BUSZI) a 20. század nyolcvanas éveinek végén 250 beszélővel magnetofonra felvett, egyenként 2–3 órás interjút tartalmaz (Kontra 1988; Váradi 2003). Az anyag számítógépes lejegyzése és kódolása is megtörtént.[7] A BABEL az első magyar beszédadatbázis, amely nemzetközi szabvány alapján készült, 60 bemondóval felolvastatott szövegeket tartalmaz (Vicsi–Vig 1998). Az MTBA magyar telefonbeszéd-adatbázis vezetékes és mobiltelefonról rögzített beszédkorpusz, amely a magyar beszédtechnológiai kutatások és fejlesztések támogatására készült, 500 adatközlő felolvasásait tartalmazza (Vicsi et al. 2002; Vicsi 2010).

A beszédadatbázisok a hangzó anyag mellett általában annak valamilyen szintű írásos változatát is tartalmazzák. Az átiratok a felhasználási területtől függően lehetnek helyesíráson alapuló lejegyzések, fonémák szerinti átiratok, fonetikai transzkripciók, jelölhetik az intonációt és egyéb szupraszegmentumokat stb. Az egyénileg kialakított sémák mellett univerzális, illetve adaptálható szoftverek is rendelkezésre állnak (pl. a Praat: Boersma–Weenink 2005, avagy a ToBI:

[7] Részletesen: http://www.nytud.hu/buszi/bsi.htm

Beckman et al. 2007). Teljes rendszert kínál az EXMARaLDA (Extensible Markup Language for Discourse Annotation: Schmidt 2009), amelyet kifejezetten a beszélt nyelv annotálására (más szóval címkézésére) fejlesztettek ki.

A beszélt nyelvi szövegek lejegyzési sajátosságai, a lejegyzés részletezettsége, formája, kritériumai különfélék, igazodnak az adott célhoz, illetve felhasználáshoz (pl. Grønnum 2009; Maekawa 2003). Az annotálások alapvető nehézsége abban rejlik, hogy rendszerint egy-két személy (fonetikus, nyelvész) végzi a lejegyzéseket, ezért azok eredménye kisebb-nagyobb mértékben ugyan, de mindig a lejegyzők szubjektív észleletét tükrözi (vö. Hunston 2002). Az adott hanganyag az egyén percepciós mechanizmusának szűrőjén megy keresztül, aki egyes jelenségekre jobban felfigyel, másokra kevésbé, továbbá sosem zárható ki a téves észlelés, a félreértés lehetősége (pl. Tóth–Kocsor 2003; Neuberger 2009). Noha a lejegyzés előre kialakított szempontok és követelmények szerint történik, amelyeket a lejegyzők legjobb tudásuk szerint betartanak, az átírások során azonban felmerülhetnek olyan (formai) problémák, amelyekre az előírások nem feltétlenül terjednek ki. Ezek hozzájárulnak a szubjektivitás növekedéséhez. Az egyik felhasználó számára megfelelő annotálás a másik számára nem lesz feltétlenül és teljesen elfogadható. Az írásos rögzítés mégis nagy segítség a kutató számára, mert ha ellenőrzésre szorul is az annotált szöveg, egyfajta kiindulásként mégis rendelkezésre áll. Nagy az igény az automatikus annotálásokat megvalósító szoftverekre, ezek használatát azonban a kézi ellenőrzésnek és az esetleges javításnak mindig ki kell egészítenie (pl. Olaszy–Bartalis 2008).

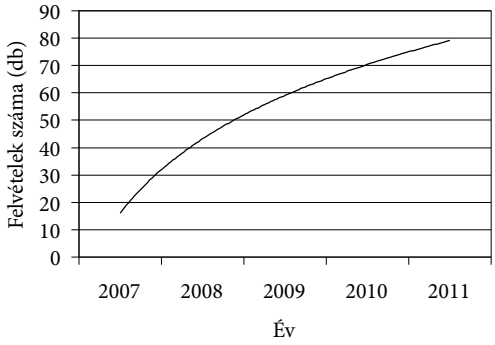
A lejegyzés nagyon időigényes munka. Egy percnyi elhangzó szöveg lejegyzéséhez – a szöveg típusától, az adatközlő beszédsajátosságaitól és a lejegyző gyakorlottságától (továbbá aktuális fiziológiás állapotától) függően – átlagosan tíz percre van szükség (az idő nagymértékben változhat az átírás komplexitásától függően). Ebbe az időtartamba beletartozik a kijelölt szövegrész első meghallgatása, az ismételt meghallgatások, az adott (hangzó) szakasz írásos változatának elkészítése, majd annak újbóli meghallgatás utáni ellenőrzése és esetleges javítása (Neuberger 2009).

A jelen tanulmány célja az MTA Nyelvtudományi Intézetének Fonetikai Osztályán készülő BEszélt nyelvi Adatbázis (BEA) fejlesztésének, eredményeinek és kutathatóságának bemutatása. Ez az első sok beszélővel rögzített, nagy mennyiségű hangzó anyagot és különböző szintű átiratukat, illetve annotálásukat tartalmazó adatbázis, amelynek a felvételi körülményei (stúdiókörülmények) állandóak. Ezek a paraméterek már önmagukban is nemzetközileg a legjelentősebb beszédadatbázisok közé sorolják a BEA-t.

2. A BEA adatbázis fejlesztése

Láttuk, hogy vannak magyar anyagot is tartalmazó korpuszok a világban, illetve magyar fejlesztésű gyűjtemények, ez utóbbiak azonban főként olvasott beszédet tartalmaznak. A szoros értelemben vett fonetikai, a tágabb értelemben vett nyelvészeti elemzések, azaz a spontán beszéd több aspektusú vizsgálata, továbbá a beszédtechnológiai feladatok igénye olyan multifunkcionális adatbázis fejlesztését tette szükségessé, amely elméleti és alkalmazott kutatások anyagául egyaránt szolgálhat. A már meglévő korpuszok és adatbázisok tapasztalatai alapján az MTA Nyelvtudományi Intézetének Fonetikai Osztályán 2007-ben indult meg a BEA fejlesztése. A távlati cél 500 személy beszédének rögzítése, amelyben igyekszünk arányosítani a nők és a férfiak részvételét, az életkor szerinti megoszlást, valamint az iskolázottságot. Az adatbázis tartalmának (protokolljának) a megtervezésekor tekintetbe vettük a fent felsorolt kutatási területek igényeit, folyamatosan a legkorszerűbb felvételi technikát alkalmazzuk, és bizonyos mértékig érvényesítünk (noha ez nem cél) szociológiai tényezőket. A fejlesztés tervezésével egy időben megkezdődött a hangzó anyag lejegyzési stratégiáinak, valamint a lekérdezhetőség módozatainak a kidolgozása. A próbafelvételeket követően 2007 októberétől elindult az adatbázis fejlesztése. Az 1. ábra az adatközlők számának növekedését szemlélteti éves bontásban.

A BEA teljes rögzített anyaga jelenleg 230 óra 11 perc 26 másodperc, ami körülbelül 3 200 000 szót jelent. A legrövidebb felvétel 24 perc 27 másodperces, a leghosszabb 2 óra 24 perc 47 másodperces; az átlag 52 perc. Két felvétel (0,7%) volt hosszabb 2 óránál; a teljes anyag 17,9%-a 1 és 2 óra közötti; 11%-a 50 perc és 1 óra közötti; 32%-a 40 és 50 perc közötti időtartamban realizálódott, mindössze 3,1%-a rövidebb fél óránál. Az adatközlők száma 265.



1. ábra. A BEA adatbázis felvételeinek száma évenkénti bontásban 2007-től

Rövidesen elkészül a BEA honlapja (2. ábra), amelynek célja az adatbázis megismertetésén és bemutatásán kívül az, hogy lehetőséget biztosítson a kutatóknak a munkájukhoz szükséges beszédanyagok kiválasztásához. A honlapra feltett táblázat tartalmazza a felvételek rendszerét, az adatközlők egyes jellemzőit (l. lejjebb), a felvétel időtartamát, a témákat. A kutató ennek alapján kiválasztja a szükséges fájlokat, és elküldi az igényeit a fejlesztőknek. Ezen a módon hozzáférést kaphat a beszédanyagokhoz.

2. ábra. A BEA (készülő) honlapja

2.1. A BEA adatbázis felvételi protokollja

Az adatbázis döntően spontán beszédanyagokat tartalmaz, azonban az összehasonlíthatóság érdekében mondatismétléseket és felolvasásokat is rögzítenek benne. A protokoll 6 részből áll, amelyek a következő megnevezésekkel jellemezhetők: narratíva, véleménykifejtés, tartalomösszegzés, társalgás, mondatismétlés, felolvasás. Minden adatközlővel többféle típusú spontán beszéd rögzítésére kerül sor. 1. A narratívák az adatközlő életéről, családjáról, munkájáról, hobbijáról szóló, többé-kevésbé összefüggő monologikus szövegek. 2. A véleménykifejtés (amely nagyobbrészt szintén narratíva) egy az interjúkészítő által megadott, éppen aktuális témának a véleményezése. Néhány a témák közül: jogosítványszerzés, zéró tolerancia a gépkocsivezetőkkel szemben, tervezett áremelkedések, házassági szerződés, klímaváltozás, tanárok elleni erőszak, budapesti közlekedés, otthonszülés, internetes és hagyományos könyvtár, állatvédelmi törvény, mobiltelefon kisgyerekeknek, olvasási szokások, hitelfelhalmozás, dohányzási tilalom, csipszadó. Az interjúkészítő itt is arra törekszik, hogy az adatközlő minél

hosszabban beszéljen összefüggően, ez a kommunikációs helyzet azonban megköveteli, hogy néha ő is megszólaljon, elmondja a saját véleményét, így esetenként dialógus jellegű beszédrészek is létrejönnek. (Az interjúkészítő szándékosan mindig az adatközlővel ellentétes véleményt igyekszik képviselni.) 3. A tartalomösszegzés voltaképpen irányított spontán beszéd. Az adatközlő felvételről meghallgat egy szöveget, és ezt követően rögtön a saját szavaival el kell mondania annak tartalmát. Az egyik szöveg egy rövid tudománynépszerűsítő cikk (174 szavas; 1 perc 37 mp tartamú), a másik egy anekdotaszerű történet (270 szavas; 2 perc 5 mp tartamú), rögzítésük átlagos női beszélővel történt. 4. A társalgás során az adatközlőn és az interjúkészítőn kívül egy további személy vesz részt a beszélgetésben. A téma változó, az élet mindennapjaihoz kapcsolódik; ugyanazon adatközlő esetében mindig különbözik a véleménykifejtés témájától. A társalgás témáiból: szilveszter, esküvői élmények, álláskeresés, drogtermelés a lakásban, húsvét, repülőszerencsétlenség, dohányzás, házasság vagy együttélés, érettségi, nyaralás, karácsonyi készülődés, gázkrízis Európában, iskolai erőszak, állattartás lakásban, a válság hatása a kultúrára, metróépítés, a könnyű drogok legalizálása, színházi élet, a diákok jogai, nők és karrier, gyermekvállalás, biciklizés mint közlekedési forma, koncertek, a diploma értéke stb. A véleménykifejtés és a társalgás témáit az interjúkészítő választja ki a beszélő életkorának, foglalkozásának, érdeklődésének megfelelően (ebben a narratíva iránymutató). 5. A mondatismétlés anyaga 25 egyszerű vagy összetett mondat (pl. *A farsangi bálban mindenkinek szép jelmeze volt*). A mondatot az interjúvezető olvassa fel az adatközlőnek, akinek egyszeri meghallgatás után azonnal meg kell azt ismételnie. (Ha sikertelen az ismétlési kísérlet, többször is elhangozhat a mondat.) 6. A protokoll szerint az adatközlő kétféle szöveget olvas fel. Az egyik a huszonöt, korábban ismételt mondat, a másik egy tudománynépszerűsítő cikk felolvasása. (A felvétel mindig ezzel zárul.)

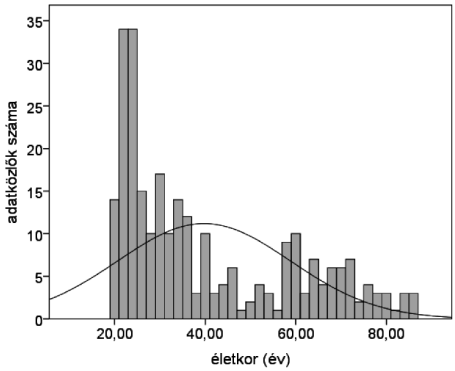
2.2. A beszédrögzítés körülményei

A felvételek mindig azonos helyen és technikai körülmények között készülnek, a Fonetikai Osztály saját tervezésű, zajszigetelt szobájában. A szoba méretei (a hangszigetelő réteget nem számítva): 340 × 210 × 300 cm. A hangcsillapítás mértéke a külső környezethez képest 50 Hz-en 35 dB, 250 Hz fölött pedig ≥ 65 dB. A szoba belső terének fala – az utózengés elkerülése érdekében – hangtörő felülettel van kialakítva. A folyosói nyílászáró két, egymástól 30 cm távolságra lévő ajtó, amelyek külön-külön nyithatók. Mind a külső, mind a belső ajtó jó hangcsillapítású. A belső ajtó a Magyar Rádió által is használt, különlegesen kiképzett

zajszigeteléssel van ellátva. A szobában felvett hanganyagok jel/zaj viszonyának értékei szerint ez a zajszigetelt szoba (ún. csendesített szoba) alkalmas jó minőségű hangfelvételek készítésére. A felvevő mikrofon típusa AT4040. A rögzítés digitális, közvetlenül a számítógépre történik a GoldWave hangeditáló szoftverrel, 44,1 kHz-es mintavételezéssel. Tárolás: 16 bit, 86 kbyte/s, monó. A rögzített felvételek jelenleg öszesen 71 GB-ot tesznek ki; DVD-ken és hat külső HDD-n is archiválva vannak. Az interjúkészítő a felvételek 95%-ában ugyanaz a személy volt (fiatal nő). A társalgások harmadik személye fiatal nő vagy férfi (kollégák).

2.3. Adatközlők

Az adatközlők száma jelenleg 265; egynyelvű, budapesti felnőttek, hallásuk életkoruknak megfelelően ép. Jelenleg 157 női és 108 férfi beszélő anyaga áll rendelkezésre. Életkoruk 20 és 85 év közötti (3. ábra). A fejlesztés során (mint már említettük) törekszünk az egyes életkoroknak a jelenleginél arányosabb reprezentálására.

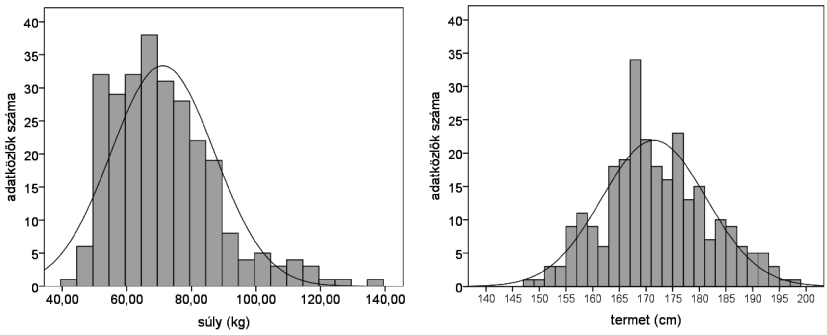


3. ábra. A BEA adatközlőinek megoszlása életkor szerint

A felvételt követően megtörténik az adatközlők kódolása (betű- és számsorok formájában), amely rendszert alkot: bea076n042 (= az adatbázis 76. felvétele, női beszélő, közülük a 42. adatközlő) vagy bea125f51 (= az adatbázis 125. felvétele, férfi beszélő, közülük az 51. adatközlő). A kódolással egyidejűleg anonimizáljuk a felvételeket, amelyek a kódok alapján kereshetők az adott személy azonosítható-sága nélkül. (A fejlesztés során a fejlesztők az MTA Nyelvtudományi Intézetének „Humán vizsgálatokon alapuló nyelvészeti kutatások etikai szabályozása" előírásait minden tekintetben szigorúan betartják. Az adatközlő a beleegyező nyilatkozatot a felvételt követően írja alá, így módja van meghallgatni a felvételt, illetve esetleges törlést kérni a felvett anyagban az aláírás előtt.)

Minden egyes felvétel esetében dokumentáljuk az adatközlő életkorát, iskolai végzettségét, foglalkozását, termetét, súlyát, esetleges beszédhibáját, azt, hogy dohányzik-e, valamint a spontán beszéd témáit. A jelenlegi adatközlők közül 47-en dohányoznak, 4-en csak dohányoztak (235-en nem dohányoznak). 10 adatközlő általános iskolai végzettséggel, 109 érettségivel rendelkezik, 1 fő technikumot végzett, 147 adatközlőnek felsőfokú végzettsége van. A foglalkozások rendkívül sokfélék, a teljesség igénye nélkül: védőnő, mérnök, tanár, takarítónő, gyógypedagógus, autószerelő, fűtő, színész, adminisztrátor, mentős, egyetemi hallgató, médiamunkatárs, bérelszámoló, énekes, háztartásbeli, orgonaépítő, köztisztviselő, szabó, orvos, informatikus, raktáros, munkanélküli, gondnok, közgazdász, grafikus, úszómester, forrasztó, kézbesítő, muzeológus, pap, kertépítő, pókerjátékos, forgatókönyvíró, cserépkályha- és kandallókészítő, nővér, játékfejlesztő, ingatlanszakértő stb.

A beszélő termete és súlya tágabb-szorosabb összefüggésben lehet a beszédével (ún. alkati harmónia, vö. Gósy 1999). Bizonyos (alkalmazott) kutatásokban és gyakorlati alkalmazásokban fontos tényező a súly és a termet megbecsülhetősége (4. ábra).



4. ábra. A BEA adatközlőinek megoszlása a súlyuk (balra) és a termetük (jobbra) függvényében

3. A BEA lejegyzése, annotálása

A BEA anyagainak lejegyzése több szinten történik. Ez lehetőséget ad arra, hogy a kutató a számára legmegfelelőbbet válassza ki, és azt használja fel a munkájában. A különféle szintű lejegyzések a fokozatosságot is biztosítják, az áttekintő megismeréstől a részletes annotációig. Jelenleg az alábbi átiratok szolgálják a nyelvészeti és beszédtechnológiai kutatásokat.

3.1. Központozás nélküli, ún. elsődleges átírás helyesírásban. Az átírás során a lejegyzők a Microsoft Office Word programot használják (.doc formátum). Az adatközlők jelölése mindig ugyanaz, A (adatközlő), T1 (interjúkészítő és az egyik társalgó partner), T2 (másik társalgó partner). Az átírási szabályok szerint (Gósy 2008; Gyarmathy–Neuberger 2011) nagybetűvel csak a tulajdonnevek vannak írva, a későbbi feldolgozás szempontjából fontosnak ítélt jelenségek, mint például a megakadások (félkövérítve), a fiziológiai és egyéb hangadások, például nevetés (! jellel), avagy az egyszerre beszélések (zárójelek között) tükröztetve vannak. Az átírás minden, nem normatívnak ítélt alakot félkövérrel jelöl; ha a közlés nem tartalmazza a javítást, a lejegyző [ ]-ben megadja a helyes szóalakot, például: ***érzezzük*** [érezzük] *magunkat*). Az egyes megakadásjelenségek típusát azonos módon jelölik; például a nyújtásokat az adott beszédhang betűjelének kettőzésével, a hezitálásokat (kitöltött szüneteket) betűhármasokkal (pl. ***öööö, mmm***), az észlelt néma szüneteket pedig a □ jellel. A lejegyzési útmutató kitér a köznyelvben használatos, de nem szótári alakban előforduló szavak (pl. *aszongya*, *asszem*, *nemtom*), az idegen szavak, rövidítések, betűszók és mozaikszók, illetőleg a lejegyző számára értelmezhetetlen szóalakok (** jelek között) lejegyzésének szabályaira (5., 6. és 7. ábra). Az átiratok tartalmazzák a teljes felvétel és a protokoll egyes részeinek időtartamát is.

ott egy pohár sört meg lehet inni és akkor hogyha az ember nem csinál □ ! **öööö** ugye balesetet akkor nem számít □ tehát hogyh**aa** csak úgy az embert szondáztatják é**ss** van benne egy pohár sör akkor semmi baj nincs □ ! hát ha persze csinál egy balesetet és volt benne még egy pohár sör is hát az nem az ne**mm** jó ! □ az nem **nem** ad hozzá a dologhoz □ ! hane**mm öööö** az inkább akkor súlyosbító körülmény ! d**ee** ! ! d **ee** * **mo*** **ma** Magyarországon **az** azt figyeltem meg hogy ! hogy akik **öööö** mondjuk így vezetgetnek **öööö** □ **öööö** egy-egy pohár alkohollal azok nem nagyon tudják megállni az egy **s** [sört] egy pohár sört hanem akkor betesznek mellé még két unikumot [unicumot] meg ! három pohár**öööö izé mmm** mit tudom én **mmm** □ királyvizet és akkor ! **akkor** az már nagyon erős ! □ tehát én azt **azt** veszem észre hogy aki ! □ **öööö** csak egy picikét iszogat **öööö** a**zz** a**zz** előbb-utóbb már többet is megenged magának és akkor □ ! **ilya** [ilyen] ez a magyar **öööö** □ **öööö** vezetési mentalitás ugye hogy hát **uu**gyan már nem számít hát

5. ábra. Spontánbeszéd-részlet elsődleges átírásban

Az elsődleges átírásnak egyaránt vannak előnyei és hátrányai. Előnynek tekinthető, hogy egy-egy fájlban áll rendelkezésre a teljes protokoll átirata (adatközlőnként), és így egy felvétel leírásában jól kereshetők szavak, szóhatárok, nonverbális

A: a mindenit **é é és** naponta jár Szegedről vagy
T2: **hát öööö** egy héten kétszer
A: egy héten kétszer tehát
(T1: félállásban van)
(A: van itt Pesten is valami)
T1: nem félállásban van
A: **jaa** félállás miatt értem az más
T1: pölö múlt héten **fe** [fent] ragadt még Szegedig se jutottál el vagy hogyan volt
T2: aha még Szegedig sem jutottam el
A: szóval nem pont Szeged
(A: meg kell hogy mondjam én eléggé)
(T2: igen mert én Szeged *?*)
A: elég jól ismerem ám Szeged környékét

6. ábra. Társalgásrészlet elsődleges átírásban

a kanadai botanikusok ! □ **öööö** kutatták é**ss öö** bebizonyították ! hogy □ a**nn**övények között is**öööö** **fel ff**elfedezhető**aa**! testvéri kapcsolat! ezt **öööö** □ ! úgy tudták**öööö**bizonyítani ! hogy **öö** □ különböző cserepekb**ee**ültették**aa** □ !**a**nem**nem öööö**testvéri ! **öööö**növényeket ! és amikor ! □ a növények **öööö** érezték **érezték** hogy **hogy** nem testvérek **m**mellé vannak ültetve ! akkor □ nagyobb gyökereket eresztettek ! **öööö**ezáltal **aa** tápanyagfelvételük is **öööö** nagyobb volt **szoal** [szóval] hogy **próbb** próbáltak**akkor** !**öööö** □ de most én nem tudtam □ jól !odafigyelni ! **öööö azelő** [azelőtt] **szo** [szóval] **ogy** [hogy] **ar** arra gondolok hogy akkor hogyha testvérek mellett **mérzik** érzik magukat akkor ugye nem**nem**növesztenek ! **öööö** □ tökéletes vagy nem **aka** □ növesztenek akkora gyökeret hogy ezáltal a testvéreiknek is **öööö** szabad ! **öö** tápanyagfelvételt biztosítsanak □

7. ábra. Részlet egy tartalomösszegzésből elsődleges átírásban

jelenségek stb. Hátránya, hogy az átirat nehezen hozható összhangba a hangzó változattal, időbe telik és gyakorlást igényel a szinkronizálásuk. A BEA adatbázis mintegy 63%-ának elsődleges átírása megtörtént.

3.2. Annotálás. Az átírásnak ez a formája a beszédszövegeknek, illetve az azokban megjelenő egyéb információknak egyfajta rögzítését jelenti oly módon, hogy a valamilyen formában leírt szöveg és a hangzás egyidejűleg megjeleníthető. A Praat[8] és a Transcriber[9] szoftverek ezt teszik lehetővé (előnyük, hogy mind-

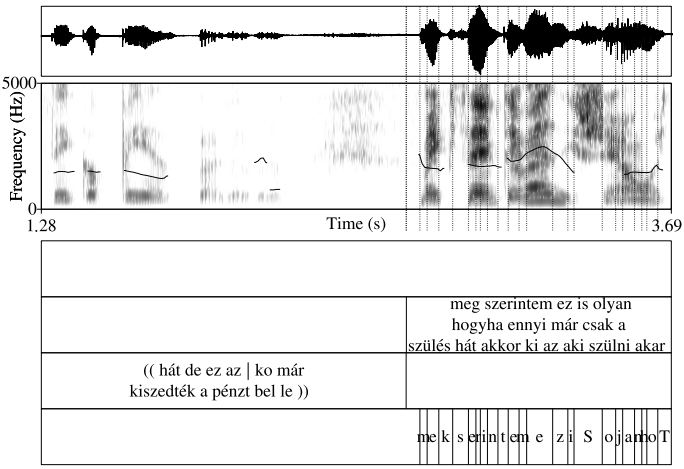
[8] http://www.fon.hum.uva.nl/praat/

[9] http://trans.sourceforge.net/en/presentation.php

kettő ingyenesen hozzáférhető). A Praat komplex akusztikai jelfeldolgozó, amely lehetőséget nyújt az annotálásra is (Boersma 2001). Beszédszövegek szegmentálására, címkézésére és lejegyzésére fejlesztették ki a Transcriber programot. Mindkettő felhasználóbarát grafikus felülettel rendelkezik, és többféle platformon (Windows, Unix) alkalmazható (a módszerről: Allwood et al. 2003; Weisser 2003; Markó–Bóna 2006; Gyarmathy–Neuberger 2011). Minthogy ezek angol nyelvű szoftverek, ezért a vezérlő felület, valamint a Transcriberben az automatikus címkék is angolul jelennek meg. Az átírt szövegek (alaplehetőségként) a Praatban .txt/TextGrid, a Transcriberben .trs kiterjesztésű adatfájlokban tárolhatók és kezelhetők.

A Praat szoftverben a beszédszakaszokat néma szünettől néma szünetig határozzák meg (a néma szünetet az észlelet és a vizuális információ alapján azonosítják). Jelölik továbbá a fordulókat (átvétel/átadás), a háttércsatorna-jelzéseket és a szünetek típusait. Az átírás alapvetően helyesírásban történik központozás nélkül. A Praat programban többféle annotálás jeleníthető meg. A 8. ábra példájában az annotált szövegrész tartalmazza a rezgésképet (legfelül), a hangszínképet (alatta), az alaphangmagasság változásait (a hangszínképre rajzoltatva), valamint a fonémaszintű (a hangszínkép alatti szövegsorok) és a hangszintű annotálásokat (az ábra legalsó sora). Az itt látható függőleges vonalak a szegmenshatárok. A hangszintű annotálás esetenként nagybetűket is alkalmaz (pl. *s* hang = S), ez az autoszegmentáló (MAUS, vö. Beringer–Schiel 2000) használatából fakad. Az egyszerre beszélések, illetve az érthetetlen vagy nehezen érthető szövegrész jelölése kettős zárójelek használatával történik. A BEA adatbázis felvételeinek mintegy 10%-a van különféle szinten annotálva a Praatban, tíz interjú szakasz-, szó- és hangszinten van felcímkézve.

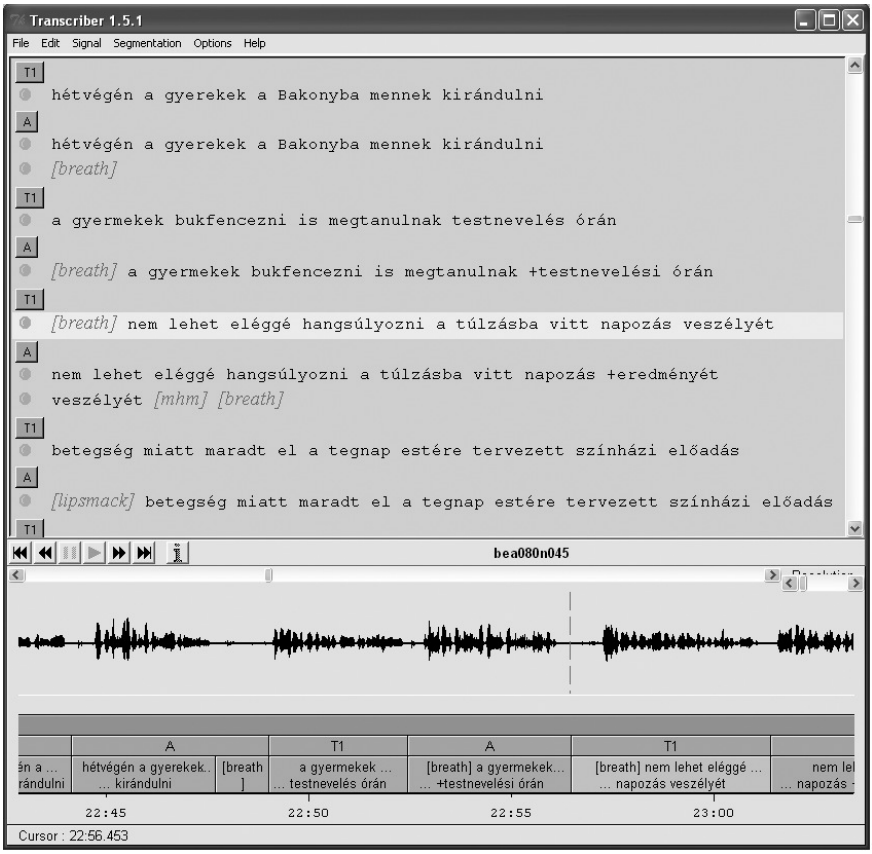
A Transcriber program a beszéd szegmentálására, címkézésére és leírására ad lehetőséget, elsősorban beszédtechnológiai alkalmazásban jelent nagy segítséget. Fonetikai mérésekre nem alkalmas, viszont a beszédfelismeréshez történő felhasználáshoz kiváló. A hanganyag és az írott szöveg itt is egyszerre láthatóvá és hallhatóvá tehető. A szoftver többféle audiofájltípust (.au, .wav, .snd) támogat. Folyamatos fejlesztés alatt áll, de a szabad hozzáférésnek köszönhetően a felhasználók újabb funkciókat adhatnak hozzá attól függően, hogy elsődlegesen mire kívánják használni (pl. Barras et al. 2001). A Transcriber alkalmas a néma szünetek, hezitálások, hümmögések és egyéb, nem beszéd jellegű hangadások (pl. köhögés, nevetés, más zajok) automatikus jelölésére címkékkel (9. ábra). A szegmentálás beszédszakaszok szerint történik, a határokat két szöveges szakasz közötti néma szünet közepére húzzák be (a néma szünetek hossza nincs jelölve, csak az átlagosnál hosszabbnak ítélt jelkimaradásokat jelölik [sil] címkével), vö. Gyarmathy–Neuberger (2011). A hangfájl megnyitását követően a kezelőfelület



8. ábra. Annotálás szemléltetése a Praat programban (a hangszínképen az alábbi látható: *ték a pénzt belőle meg szerintem ez is olyan hogy*)

alsó részén megjelenik a hang rezgésképe, alatta az egyszintű címkézés helye (itt láthatjuk függőleges vonalakkal jelölve a szegmenshatárokat), fölötte pedig – a képernyő nagy részén – a beírt szövegek helyét találjuk (a beszélők, témák megjelölésével). A trs-ben a címkézés helyesírásban történik, vannak azonban egyes esetek, például mozaikszavak, idegen szavak, régi családnevek, ahol jelölni kell a kiejtést is. Jelenleg a BEA felvételeinek 40%-a van Transcriberrel átírva.

A Praat és a Transcriber nyújtotta annotálási lehetőségeknek is vannak előnyeik és hátrányaik. Az egységes lejegyzési útmutató ellenére a lejegyzők sokszor egyéni módon, eltérő részletességgel, illetve pontossággal valósítják meg a lejegyzést. Az egyéni beszédészlelésnek itt is nagyon nagy a szerepe (Grønnum 2009). Kétségtelen előnyük, hogy a hangzó és a leírt szöveg egyidejűleg rendelkezésre áll, ezáltal az elemzések jóval könnyebben és egyszerűbben valósíthatók meg. A Praatban ez egyszersmind az akusztikai-fonetikai méréseket is könnyen lehetővé teszi, és biztosítja az automatikus adatkinyerést. A korábbi, Wordben elkészített lejegyzéshez képest nagy előny itt, hogy a hang hullámformája, valamint a szöveges rész közös felületen, egy ablakban látható és kezelhető, így nem kell váltogatni a hanglejátszó program és a Word között, valamint a hang elindítása, leállítása, újrajátszása egy billentyű segítségével végrehajtható. A Transcriberrel történő átírás bizonyos pontokon megkönnyíti a lejegyzők dolgát, más tekintetben több odafigyelést igényel tőlük. A program lehetővé teszi, hogy a beszéd időszeletekre bontásával a hang és a szöveg szinkronba kerüljön, így kisebb részeket



9. ábra. A Transcriber felhasználói felülete

kell egyszerre feldolgozni; ez megkönnyíti az ellenőrzést, a visszakeresést, valamint segíti a későbbi felhasználást.

A szakaszszintű annotáció alapján a BEA közel 120 órányi anyagában autoszegmentálóval hangszinten annotáltattuk a felvételeket (Transcriberben). Ez a mennyiség, ismereteink szerint, az egyik legnagyobb hangszinten felcímkézett spontánbeszéd-anyag. Az annotálások tovább finomíthatók, de hangsúlyozandó, hogy azok mindig az adott kutatási cél függvényei.

4. A természetesség kérdése a beszédadatbázisokban

A spontán beszédet tartalmazó adatbázisok esetében gyakorta felmerül a természetesség kérdése, amely keveredik a spontaneitásával (Gósy et al. 2009). Előfordul, hogy a természetességet és a spontaneitást rokon értelmű fogalmakként kezelik, így összekeveredik a beszédtervezés és a verbális viselkedés többszintű mechanizmusa. A spontaneitás nem sérül egy mesterségesen létrehozott kommunikációs helyzetben, például egy interjú során, hiszen a beszélő az adott pillanatban válogat a gondolatai között és rendeli hozzájuk a nyelvi formát. A rögzített beszéd tehát egyértelműen spontán akkor, ha a beszélőnek a feltett kérdés(ek)re azonnal kell válaszolnia, megelőzően nem készülhetett fel az adott témára sem a véleménykifejtéshez, sem a társalgáshoz. Ez a spontaneitás kritériuma.

A beszédfelvételek készítésének egy másik fontos kérdése, hogy a beszélő ilyenkor mennyire viselkedik a közléseit tekintetve természetesen („megfigyelői paradoxon": Labov 1979). A megfigyelői paradoxon áthidalásának egyik lehetséges módja az, ha az adatközlőtől többféle típusú beszédet rögzítünk (Wardhaugh 1995), továbbá ha olyan körülményeket tudunk teremteni, amelyben a beszélő szorongása oldódik, és a mikrofon okozta fokozott önellenőrzés megszűnik. Nusbaum és munkatársai (1995) szerint a természetesség a beszéd multidimenzionális, szubjektív minősége, ami úgy is felfogható, hogy a beszéd akkor természetes, ha megfelel a beszélő egyéniségének. A mesterséges helyzet, illetve a beszédrögzítés tényének negatív hatása jelentkezhet abban, hogy a beszélők törekszenek a nyelvi norma megközelítésére, nagyobb mértékben, mint a mindennapi megnyilatkozásokban (Szende 1973; Lindblom 1990). Ez azonban nem általánosítható, és a beszélők nagyobb része nem is képes a „regiszterek" között könnyen váltani. A felvétel előtt az adatközlők egy része bevallja, hogy izgul (ez sokszor a testtartáson, a mozdulatokon is látható), azonban szinte kivétel nélkül hamar feloldódnak, a legtöbbjüket láthatóan nem zavarják a felvételi körülmények. Ennek nyelvi igazolása az (is), hogy nemegyszer olyan szavakat és nyelvi fordulatokat használnak, amelyek egyáltalán nem illenek egy formális verbális kommunikáció kereteibe (az interjúkészítő tapasztaltságának jelentős szerepe van a relatíve természetes helyzet kialakításában). Megállapítható tehát, hogy a BEA felvételeinek rögzítése során a beszéd spontaneitása nem sérül és az adatközlők a körülményekhez képest természetes módon beszélnek.

5. Nyelvészeti kutatások a BEA adatbázison

A BEA főbb jellemzőinek bemutatása jól szemlélteti, hogy az adatbázis a nyelvészet számos területén kínál tanulmányozásra alkalmas anyagot. A beszédhangok képzésének akusztikai-fonetikai következményei, a koartikuláció, a szóejtés sajátosságai, a szupraszegmentális tényezők vizsgálata évtizedeken keresztül abba a módszertani nehézségbe ütközött, hogy nem állt rendelkezésre megfelelő minőségű és mennyiségű spontán beszéd, illetve csak nehezen lehetett ilyen beszédanyagot a szükséges kritériumok betartásával rögzíteni. A szoros értelemben vett fonetikai kutatások mellett lehetőség nyílik a társalgások elemzésére, pragmatikai megközelítésre, a beszédalkalmazkodás tanulmányozására, az idősek spontán beszédének vagy a megakadásjelenségeknek a vizsgálatára (pl. Gráczi 2008; Beke–Horváth 2009; Váradi 2009; Beke–Gyarmathy 2010; Bóna 2010; Dér 2010; Gósy–Beke 2010; Gósy–Horváth 2010; Gráczi–Bata 2010). Az adatbázis révén magyar nyelven először vált lehetővé az összes magánhangzó akusztikai-fonetikai szerkezetének leírása, a koartikulációs mezők jellemzése, a beszédhangok semlegesedésének, a gyakori szavak ejtési sajátosságainak, a zöngeminőség kommunikációs funkcióinak az elemzése, avagy a prozódia szerepének vizsgálata a spontán beszéd tagolásában (pl. Beke–Szaszák 2010; Gráczi–Horváth 2010; Markó 2010; Markó et al. 2010).

A beszédszövegeket szavak, virtuális mondatok, gondolategységek építik fel, amelyekben elemezhetők a morfológiai és a szintaktikai formák, a szókapcsolatok és közlésegységek, a szintaxis és a prozódia összefüggései, pragmatikai jellemzők, és tetten érhetők a jelentésváltozások, a nyelvhasználati módosulások, valamint a kiejtés egyéni és beszédstílusfüggő változatai is. A spontán beszéd grammatikájának leírása mind a mai napig nem történt meg, a BEA ehhez is lehetőséget nyújt. Cél lehet a morfológiai, szintaktikai jelenségek megismerése az élőbeszédben, a lexikális hozzáférés folyamatainak vizsgálata, feltételezett szinkrón nyelvi változások elemzése, stilisztikai vonatkozások, illetve a nyelvhasználat bármely aspektusának vizsgálata az életkor függvényében. A tipikus beszélők rögzített anyagai támpontul szolgálnak az atipikus (pl. afáziás) beszéddel, illetve nyelvhasználattal történő összevetésre. A szegmentumoknak, a szavaknak, a frázisoknak, a morfológiai struktúráknak, a nyelvi szabályszerűségek megvalósulásainak az előfordulási gyakorisága, megterheltségük egy nyelv tulajdonságainak fontos részét jelentik, de csak nagy anyagon kutathatók (pl. Beke et al. 2012). A spontán beszéd tényei hozzájárulhatnak az elméleti nyelvészeti területeken folyó elemzésekhez, mint például a kvantorhatókörök mintázatainak vizsgálata, a szintaxis és a szemantika egyes kérdéseinek elemzése, téma–réma kutatások, szintaxis és prozódia összefüggéseinek vizsgálata. A nagy mennyisé-

gű rögzített anyag jól használható a nyelvi normativitás megismerésében, a főbb irányok meghatározásában, illetve alátámasztásában.

A BEA – protokolljának, felvételi és lejegyzési sajátosságainak, valamint mennyiségi mutatóinak (sok beszélő és beszélőnként relatíve nagy minta) következtében – jól használható beszédtechnológiai kutatásokban, valamint a spontán beszéd mesterséges felismerését és a kriminalisztikai beszélőazonosítást célzó munkálatokban is. Lehetővé teszi a spontán beszéd tipikus prozódiai egységeinek a meghatározását tanuló algoritmusokkal, a társalgások automatikus feldolgozására szolgáló szoftverek fejlesztését, avagy az egyszerre beszélések automatikus osztályozását, a beszédfordulók automatikus előrejelzését, a szavak kezdetének a meghatározását folyamatos szövegben (pl. Beke 2011). Már fejlesztésének jelen szakaszában alkalmas arra, hogy statisztikai eszközökkel elősegítse a beszédfelismerő(k) jó működését (például a megakadásjelenségek automatikus kategorizálásával és statisztikai modellezésével).

A jól megtervezett és kivitelezett, annotált és lekérdezhető adatbázisok kiváltják az időigényes felvételek készítésének munkáját, hatalmas adathalmazt biztosítanak sokféle kutatáshoz, és a nyelv valós használatát tükrözik. A BEA számos tekintetben nemzetközileg is jelentős adatbázis.

Irodalom

Allwood, Jens – Leif Gronqvist – Elisabeth Ahlsen – Magnus Gunnarson 2003. Annotations and tools for an activity based spoken language corpus. In: Jan van Kuppelvelt – Ronnie W. Smith (szerk.): Current and new directions in discourse and dialogue. Dordrecht: Kluwer. 1–18.

Anderson, Anne H. – Miles Bader – Ellen G. Bard – Elisabeth Boyle – Gwyneth Doherty – Simon Garrod – Stephen Isard – Jacqueline Kowtko – Jan McAllister – Jim Miller – Catherine Sotillo – Henry S. Thompson – Regina Weinhert 1991. The HCRC map task corpus. Language and Speech 34: 351–366.

Bael, Christophe van – Lou Boves – Henk van der Heuvel – Helmer Strik 2007. Automatic phonetic transcription of large speech corpora. Computer Speech and Language 21: 652–668.

Barras, Claude – Edouard Geoffrois – Zhibiao Wu – Mark Liberman 2001. Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communication 33: 5–22.

Beckman, Mary E. – Julia Hirschberg – Stefanie Shattuck-Hufnagel 2007. The original ToBI system and the evolution of the ToBI framework. In: Jun Sun-Ah (szerk.): Prosodic models and transcription: Towards prosodic typology. Oxford: Oxford University Press. 9–54.

Beke András 2011. Szókezdetek automatikus osztályozása spontán beszédben. Magyar Nyelvőr 135: 226–241.

Beke András – Gósy Mária – Horváth Viktória 2012. Gyakorisági vizsgálatok spontán beszédben. Beszédkutatás 2012: 260–277.

Beke András – Gyarmathy Dorottya 2010. Zöngétlen résmássalhangzók akusztikai szerkezete. Beszédkutatás 2010: 57–76.

Beke András – Horváth Viktória 2009. A nazális koartikuláció variabilitása a spontán beszédben. Beszédkutatás 2009: 28–45.

Beke András – Szaszák György 2010. Automatic recognition of schwa variants in spontaneous Hungarian speech. Acta Linguistica Hungarica 57: 329–353.

Beringer, Nicole – Florian Schiel 2000. The quality of multilingual automatic segmentation using German MAUS. In: Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, China. 728–731.

Boersma, Paul 2001. Praat, a system for doing phonetics by computer. Glot International 5: 341–345.

Boersma, Paul – David Weenink 2005. Praat: Doing phonetics by computer. (Version 4.4). Letöltve: 2011. március 5. http://www.praat.org/.

Bóna Judit 2010. Bizonytalansági megakadások idősek és fiatalok spontán beszédében. Beszédkutatás 2010: 125–138.

Burnard, Lou – Guy Aston 1998. The BNC handbook: Exploring the British National Corpus. Edinburgh University Press: Edinburgh.

Calhoun, Sasha – Jean Carletta – Jason M. Brenier – Neil Mayo – Dan Jurafsky – Mark Steedman – David Beaver 2010. The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language Resources and Evaluation Journal 44: 387–419.

Chan, Dominic – Adrian Fourcin – Dafydd Gibbon – Björn Grandström – Mark Huckvale – George Kokkinakis – Knut Kvale – Lori Lamel – Börge Lindberg – Asunción Moreno – Jiannis Mouropoulos – Francesco Senia – Isabel Trancoso – Corin 't Veld – Jerome Zeiliger 1995. EUROM – a spoken language resource for the EU. In: Puppel – Demenko (1995, 867–870).

Clark, Herbert H. – Jean E. Fox Tree 2002. Using *uh* and *um* in spontaneous speaking. Cognition 84: 73–111.

Cole, Ronald A. – Mike Noel – Terri Lander – Terry Durham 1995. New telephone speech corpora at CSLU. In: Puppel – Demenko (1995, 821–824).

Dér Csilla 2010. „Töltelékelem" vagy új nyelvi változó? A *hát, úgyhogy, így* és *ilyen* újabb funkciójáról a spontán beszédben. Beszédkutatás 2010: 159–170.

Godfrey, John J. – Edward C. Holliman – Jane McDaniel 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In: Proceedings of ICASSP. 517–520.

Gósy Mária 1999. Az egyéni hangszínezet és a beszélő felismerésének kísérleti-fonetikai megközelítése. Magyar Nyelvőr 123: 424–438.

Gósy Mária 2008. Magyar spontánbeszéd-adatbázis – BEA. Beszédkutatás 2008: 194–207.

Gósy Mária – Beke András 2010. Magánhangzó-időtartamok a spontán beszédben. Magyar Nyelvőr 134: 140–165.

Gósy Mária – Gyarmathy Dorottya – Horváth Viktória 2009. A beszéd természetességéről alkalmazott fonetikai szempontból. Beszédkutatás 2009: 170–181.

Gósy, Mária – Viktória Horváth 2010. Changes in articulation accompanying functional changes in word usage. Journal of the International Phonetic Association 40: 135–161.

Gósy Mária – Horváth Viktória – Nikléczy Péter 2011. A Hegedűs-archívum mint korszerű adatbázis. In: Bárth M. János – Vargha Fruzsina Sára (szerk.): Hangok – helyek. Budapest: ELTE Magyar Nyelvtudományi és Finnugor Intézet. 85–103.

Gráczi Tekla Etelka 2008. Alveoláris spiránsok akusztikai fonetikai vizsgálata. Beszédkutatás 2008: 33–51.

Gráczi, Tekla Etelka – Sarolta Bata 2010. The effect of familiarization on temporal aspects of turn-taking: A pilot study. Acta Linguistica Hungarica 57: 307–328.

Gráczi Tekla Etelka – Horváth Viktória 2010. A magánhangzók realizációja spontán beszédben. Beszédkutatás 2010: 5–16.

Grønnum, Nina 2009. A Danish phonetically annotated spontaneous speech corpus (DanPASS). Speech Communication 51: 594–618.

Gyarmathy Dorottya – Neuberger Tilda 2011. A BEA adatbázis alkalmazásfüggő lejegyzései. Beszédkutatás 2011: 109–121.

Hennebert, Jean – Håkan Melin – Dijana Petrovska – Dominique Genoud 2000. POLYCOST: A telephone-speech database for speaker recognition. Speech Communication 31: 265–270.

Hunston, Susan 2002. Corpora in applied linguistics. Cambridge: Cambridge University Press.

Hutchinson, Ben – Cécile Pereira 2001. Um, one large pizza. A preliminary study of disfluency modelling for improving ASR. In: Robin Lickley – Elisabeth Shriberg (szerk.): Disfluency in spontaneous speech. Edinburgh: University of Edinburgh. 77–81.

Keating, Patricia A. – Dani Byrd – Edward Flemming – Yuichi Todaka 1994. Phonetic analyses of word and segment variation using the TIMIT corpus of American English. Speech Communication 14: 131–142.

Keszler Borbála 1983. Kötetlen beszélgetések mondat- és szövegtani vizsgálata. In: Rácz Ende – Szathmári István (szerk.): Tamulmányok a mai magyar nyelv szövegtana köréből. Budapest: Tankönyvkiadó. 164–187.

KKA 1994. Ismertetés a Keleti Kereskedelmi Akadémia Fonetikai Laboratóriumának munkájáról. In: A KKA 25. évi jelentése az 1915-16-iki iskolaév végén. Bp. 1916. 55–56. Közli: Studia Academiae Nyíregyháziensis. Tomus III., 1994.

Kontra Miklós 1988. Beszélt nyelvi tanulmányok. Budapest: MTA Nyelvtudományi Intézet.

Labov, William 1979. A nyelv vizsgálata társadalmi összefüggésben. In: Csaba Pléh – Tamás Terestyéni (szerk.): Beszédaktus and kommunikáció and interakció. Budapest: Tömegkommunikációs Kutatóközpont. 365–39.

Lindblom, Björn 1990. Explaining phonetic variation: A sketch of the H&H theory. In: William J. Hardcastle – Alain Marchal (szerk.): Speech production and speech modeling. Dordrecht: Kluwer. 403–440.

Maekawa, Kikuo 2003. Corpus of spontaneous Japanese: Its design and evaluation. In: Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003), Tokyo. 7–12.

Markó Alexandra 2010. A prozódia szerepe a spontán beszéd tagolásában. Beszédkutatás 2010: 82–100.

Markó Alexandra – Bóna Judit 2006. A spontán beszéd lejegyzésének néhány módszertani kérdése. Beszédkutatás 2006: 124–133.

Markó, Alexandra – Tekla Etelka Gráczi – Judit Bóna 2010. The realization of voicing assimilation rules in Hungarian spontaneous and read speech: Case studies. Acta Linguistica Hungarica 57: 210–238.

Neuberger Tilda 2009. A spontán beszéd lejegyzése and a BEA adatbázis tapasztalatai alapján. Beszédkutatás 2009: 182–195.

Nusbaum, Howard C. – Alexander L. Francis – Anne S. Henly 1995. Measuring the naturalness of synthetic. International Journal of Speech Technology 1: 7–19.

Olaszy Gábor – Bartalis Mátyás 2008. Jelfeldolgozási és fonetikai algoritmusok kombinációja a gépi hanghatárjelölés javítására. Beszédkutatás 2008: 208–220.

Patterson, Eric K. – Sabri Gurbuz – Zekeriya Tufekci – John N. Gowdy 2002. CUAVE: A new audio-visual database for multimodal human-computer interface research. http://people.uncw.edu/ICASSP2002.pdf (Letöltve: 2012. február 15.).

Pitt, Mark A. – Keith Johnson – Elizabeth Hume – Scott Kiesling – William Raymond 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. Speech Communication 45: 89–95.

Popescu-Belis, Andrei – Jean Carletta – Jonathan Kilgour – Peter Poller 2009. Accessing a large multimodal corpus using an automatic content linking device. In: Michael Kipp – Jean-Claude Martin – Patrizia Paggio – Dirk Heylen (szerk.): Multimodal corpora. Springer lecture notes in artificial intelligence. Berlin: Springer. 50–59.

Puppel, Stanisław – Grazina Demenko (szerk.) 1995. Eurospeech '95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology. Vol. 1. Madrid: ESCA.

Ruhi, Şükriye 2011. Creating a sustainable large corpus of spoken Turkish for multiple research purposes. http://tinyurl.com/cw37zlk (Letölve: 2011. január 27.).

Schmidt, Thomas 2009. Creating and working with spoken language corpora in EXMARaLDA. In: Verena Lyding (szerk.): LULCL II: Lesser Used Languages and Computer Linguistics II. Bolzano: EURAC. 151–164.

Simpson, Adrian – Klaus J. Kohler – Tobias Rettstadt (szerk.) 1997. The Kiel Corpus of read/spontaneous speech and acoustic data base, processing tools and analysis results. (Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 32). Kiel: Universität Kiel.

Svartvik, Jan (szerk.) 1990. The London Corpus of spoken English: Description and research (Lund Studies in English 82). Lund: Lund University Press.

Szende Tamás 1973. Spontán beszédanyag gyakorisági mutató (Nyelvtudományi Értekezések 81). Budapest: Akadémiai Kiadó.

Tóth László – Kocsor András 2003. A Magyar Telefonbeszéd-adatbázis (MTBA) kézi feldolgozásának tapasztalatai. Beszédkutatás 2003: 134–146.

Váradi Tamás 2003. A Budapesti Szociolingvisztikai Interjú. In: Kiefer Ferenc – Siptár Péter (szerk.): A magyar nyelv kézikönyve. Budapest: Akadémiai Kiadó. 339–359.

Váradi Viola 2009. Hallásalapú és vizuális alapú közlések. Beszédkutatás 2009: 228–239.

Vicsi Klára 2001. Beszédadatbázisok a gépi beszédfelismerés segítésére. Híradástechnika 2001/1: 5–13.

Vicsi Klára 2010. Adatbázisok a beszédtechnológia szolgálatában. In: Német Géza – Olaszy Gábor (szerk.): A magyar beszéd – beszédkutatás, beszédtechnológia, beszédinformációs rendszerek. Budapest: Akadémiai Kiadó. 262–332.

Vicsi Klára – Tóth László – Kocsor András – Gordos Géza – Csirik János 2002. MTBA – magyar nyelvű telefonbeszéd-adatbázis. Híradástechnika 57: 35–43.

Vicsi Klára – Vig Attila 1998. Az első magyar nyelvű beszédadatbázis. Beszédkutatás 1998: 163–178.

Wardhaugh, Ronald 1995. Szociolingvisztika. Budapest: Osiris–Századvég.

Weisser, Martin 2003. SPAACy – a semi-automated tool for annotating dialogue acts. International Journal of Corpus Linguistics 8: 63–74.

A multifunctional spontaneous speech data base: BEA

Abstract: There is an increasing demand for studies of the phonetic properties of spoken language using large data bases. The aim of developing a phonetically-based multi-purpose data base of Hungarian spontaneous speech, dubbed BEA, is to accumulate a large amount of spontaneous speech material of various types (including conversations) together with sentence repetition and reading. The recorded material of BEA amounts to a total of 230 hours and is produced by 265 adult Budapest speakers (aged between 20 and 85, 157 females and 108 males), providing annotated audio materials and transcripts for various types of research and practical applications.

Keywords: large data base, spontaneous speech, various types of annotations

Főszerkesztői utószó

Az *Általános Nyelvészeti Tanulmányok* 24. kötete, bármily meglepőnek látszik is a kötetcím alapján, hagyományt folytat, pontosabban hosszú idő után éleszt fel a sorozatban korábban nagy reményeket keltő tudományközi tárgykört. A jelen kötet tematikája azért lehet meglepő, mert a számítógépes nyelvészet, illetve a nyelvtechnológia nemigen jut eszébe annak, aki az általános, vagy – újabb és elterjedtebb nevén – az elméleti nyelvészet részterületeit sorolná fel. Az ilyen irányú kutatások ugyanis a nyelvészet alkalmazásainak a körébe tartoznak, és aligha találunk elméleti nyelvészeti programot, amely nem csak érintőlegesen foglalkozna velük.

Csakhogy már az *ÁNyT* születésekor, amikor a matematikai, számítógépes, statisztikai és/vagy algebrai nyelvészet sokkal idegenebb volt a hagyományos bölcsészkeretekben gondolkodó általános nyelvészettől, és épp ezért jóval távolabb is volt tőle, mint manapság, tehát már az 1960-as évek elején, a Telegdi Zsigmond gondozásában elindult sorozatnak a II. kötete is e témakörökből közölt tanulmányokat, ráadásul olyan jelentős, és ma is nagyhírű tudósok tollából, mint Kalmár László, a magyar kibernetika „atyja", a világhírű matematikus Rényi Alfréd, a viharos életű „általános társadalomtudós" és polihisztor Szalai Sándor, a szoftverfejlesző szaktekintéllyé vált Dömölki Bálint, és persze a nyelvészek: mások mellett például Dezső László, Fónagy Iván, Hell György, Kiefer Ferenc, Papp Ferenc, Petőfi S. János, Szépe György, akik mind az akkor matematikai nyelvészetnek nevezett szakterület valamelyik ágát (is) művelték.

Lehet persze, hogy az akkori főszerkesztőnek kapóra jött egy olyan konferencia, amelynek anyagából, valamint annak kiegészítéséből sorozattá alakíthatta a sorszám nélkül kiadott első *ÁNyT*-kötet folytatását. De ismerve Telegdi Zsigmond elvekhez való ragaszkodását, koncepcionális szigorát, aligha feltehető, hogy pillanatnyi előnyök kedvéért tett volna erőszakot következetességén.

Az a II. kötet sok érdekes írást tartalmazott, sőt csaknem a felét a ma is aktuális gépi fordítás problémáinak szentelték a szerkesztők: Kalmár László és Telegdi Zsigmond. Most csak remélni tudjuk, hogy ötven év múlva ezt a cikkgyűjteményt is érdemes lesz majd levenni a polcról – ha lesz még könyvespolc, és lesznek még papírra nyomott könyvek.

A nem (csak) papíron hozzáférhető kiadványok sorát most mi is gyarapítjuk. Jelen kötetünkkel ugyanis tovább újítjuk az *ÁNyT* formátumát: a sorozat ezentúl e-változatban is megjelenik, azaz akik nem nyomtatott könyvként, hanem számítógép segítségével kívánják elolvasni, azok a Kiadó honlapjáról vásárolhatják meg és tölthetik le. Ezzel egyébként nemcsak a postaköltséget és a kézbesítési időt csökkentjük

nullára, hanem remélhetőleg mindörökre eltűnik az „elfogyott" címke a kiadói lista tételei mellől. Talán még szimbolikus jelentőségűnek is mondhatjuk, hogy éppen a számítógépes nyelvészeti és nyelvtechnológiai kötettel kezdődik az új korszak.

De nem ez az egyetlen újdonság az *ÁNyT* körül. Mint az előző kötet utószavában is utaltam rá, szerkesztőbizottsággal, egyfajta tanácsadó testülettel kívánom megtámogatni a mindenkori főszerkesztő munkáját, hiszen legyen bármilyen széles látókörű vagy ambiciózus is, egyetlen ember aligha képes a nyelvtudomány, vagy csupán az elméleti nyelvészet minden területét áttekinteni. Eleve nemzetközi tagságú testületre törekedtem és hálás vagyok mindazoknak, akik elfogadták a felkérést: nevüket és intézményüket a kötet elején soroltuk fel. Itt csak arra hívom fel a figyelmet, hogy nyelvfilozófustól fonológusig, tipológustól kognitív pszichológusig terjed a bizottság hozzáértése. Külön szeretném megköszönni az *ÁNyT* két előző főszerkesztőjének, Szépe Györgynek és Kiefer Ferencnek beleegyezésüket, hogy a szerkesztőbizottság tiszteletbeli tagjaiként hajlandók segíteni munkámat. Jelenlétük is biztosíték a sorozat folyamatosságára.

Végül a további lépéseket is hadd vázoljam itt. Az évenkénti megjelenéssel az *ÁNyT* egyre inkább folyóiratszerű kiadvánnyá válik, bár egyelőre továbbra is őrizzük a Telegdi Zsigmond által bevezetett tematikus számok gyakorlatát, ráadásul úgy, hogy körülbelül két évre előre rögzítjük a kötetek tárgykörét és a szerkesztő(k) személyét, ezzel is a minél nagyobb biztonságra törekedve.

Végül visszatérve a jelen kötetre: a két társszerkesztő, Prószéky Gábor és Váradi Tamás, a hazai számítógépes nyelvészet minden fontosabb műhelyéből talált reprezentatív szerzőt, s az általuk bemutatott kutatások és fejlesztések remélhetőleg pontos képet adnak e szakterület közelmúltjáról, jelenéről és biztató jövőjéről is.

Kenesei István

A nyomdai munkálatok közben érkezett a szomorú hír, hogy Szépe György 2012. szeptember 12-én elhunyt. Ő nem csupán a magyar nyelvészek generációinak felnevelője, a magyarországi nyelvtudomány számos területének megújítója, az alkalmazott nyelvészet nemzetközileg elismert nagy kezdeményező szelleme volt, hanem a kezdetektől fogva, előbb Telegdi Zsigmond alapító főszerkesztő segítőjeként, majd társszerkesztőjeként és gyakorlati sorozatszerkesztőként az Általános Nyelvészeti Tanulmányokat *is gondozta. Amint Szépe György 70 éves születésnapi köszöntőjében Péter Mihály írta: „Az ötlet eredetileg Telegdi Zsigmondtól származott, de Szépe szinte valamennyi kötet megszületésénél bábáskodott mint szerkesztő, lektor s nem utolsósorban szerző. Ahogyan ő maga tömören megfogalmazta: én mozgattam, ennyi az egész." A jelen kötetet ezért is szenteljük Szépe György emlékének.*