DOI:10.15774/PPKE.ITK.2011.004

Igei szerkezetek gyakorisági szótára

Egy automatikus lexikai kinyerő eljárás és alkalmazása

doktori (Ph.D.) disszertáció

Sass Bálint

témavezető:
Prószéky Gábor,
az MTA doktora

Pázmány Péter Katolikus Egyetem,
Információs Technológiai Kar,
Multidiszciplináris
Műszaki Tudományok
Doktori Iskola

Budapest, 2011.

DOI:10.15774/PPKE.ITK.2011.004

Dórinak, Micinek, Lencsinek, Csöpinek, Jáninak

Lencsi: A papa mit fog csinálni éjjel?
Dóri: Gondolom, a dolgozatát írja.
Lencsi: Szegény papa, éjjel is nappal is a dolgozatát írja.

DOI:10.15774/PPKE.ITK.2011.004

„Elégedjünk meg azzal, hogy a szavak sem fix pontok, és minden végleges megfogalmazás, és így a definíció is komikus."

Hamvas Béla: Karnevál

„Ami kérem a mi adattárunkban nincs benne, az nem létezik."

Star Wars II. – A klónok támadása

DOI:10.15774/PPKE.ITK.2011.004

DOI:10.15774/PPKE.ITK.2011.004

Kedvcsináló

'*Részt vesz vmiben.*' '*Górcső alá vesz vmit.*' Bár a természetes nyelvfeldolgozás kutatási hagyományában és a szótárírási hagyományban is két elkülönült területet jelentenek az igei vonzatkeretek és a többszavas kifejezések, számos nyelvben előfordulnak az effajta bonyolultabb szerkezetek, melyek *egyszerre* igei vonzatkeretek és kollokációk. Ezeket a szerkezeteket (legalább) két tartalmas elem – alapesetben egy ige és egy (ragos/névutós/elöljárós) névszó – alkotja, és ezen kívül még egy (vagy több) vonzat is szerves része a szerkezetnek. A fenti magyar nyelvűek mellett valóban számos nyelven látunk ilyenre példát: '*get rid of*' (angol; megszabadul vmitől), '*få lov til*' (dán; engedélyt kap vmire), '*imati pravo na*' (szerb; joga van vmihez), '*houden rekening met*' (holland; számításba vesz vmit), '*zijn van toepassing op*' (holland; vonatkozik vmire), '*avoir effet sur*' (francia; hatása van vmire).

Az idézett példákban az ige mellett mindig két bővítményt találunk: az egyiket egy konkrét, kötött szó tölti ki, ez alkot jelentéshordozó kollokációt az igével, a másik bővítménynek viszont csak a helyét jelöli ki a szerkezet egy esetrag vagy elöljáró segítségével. Látható, hogy általában ugyanazokkal a nyelvi eszközökkel – ragokkal, névutókkal, elöljárókkal vagy szórendi megkötéssel – kapcsoljuk a bővítményeket az igéhez; függetlenül attól, hogy a kollokátumról, vagy pedig a vonzati helyen éppen megjelenő tartalmas szóról (ilyen lenne például a '*játék*' a '*részt vesz a játékban*' esetén) van szó. A '*részt vesz vmiben*' szerkezetben például a kollokátum jelenik meg tárgyként, a '*górcső alá vesz vmit*' szerkezetben pedig a vonzat. Ez a váltakozás persze egyazon ige szerkezetei között is előfordulhat, a '*pillantást vet vkire*' és a '*szemére vet vmit*' szerkezet egyaránt tárgyat és egy '*-ra/-re*' ragos bővítményt tartalmaz, de az elsőben a tárgy a kollokátum és a '*-ra/-re*' ragos bővítmény a vonzat, a másikban pedig éppen fordítva.

Az ilyen szerkezetek – sokszor anyanyelvi intuíciónkkal ellentétes módon – kifejezetten gyakoriak, nagyon fontos szeletét képviselik egy nyelv szerkezeteinek, nem kezelhetők marginális esetként. Sokszor bírnak a részeikből nem levezethető, azaz nem kompozicionális, idiomatikus jelentéssel, kiemelt fontosságú részét kell hogy képezzék az emberi felhasználásra szánt szótáraknak és az automatikus nyelvfeldolgozó eszközök nyelvi erőforrásainak egyaránt. Fordításaikat a legtöbb esetben érdemes külön egységként tárolni, mert gyakran nem megjósolható elemeket tartalmaznak.

Szükség van tehát egy olyan adatközpontú számítógépes eljárásra, mely rendet vág a bővítménykijelölő nyelvi eszközök egymást átfedő erdejében, szétválasztja a konkrét, kötött szót tartalmazó és a szabadon kitölthető bővítményeket. Megállapítja, „felfedezi", hogy az egyes igei szerkezeteknek melyik bővítményi szó szorosan vett része kollokációként, és mely szükséges vonzati helyek kapcsolódnak még hozzá, azaz képes

DOI:10.15774/PPKE.ITK.2011.004

arra, hogy korpuszból kinyerje a jellegzetes igei szerkezeteket. A dolgozat fő eredménye ez az algoritmus (3.3. rész az 54. oldaltól), illetve az ennek közvetlen felhasználásával készülő magyar, egynyelvű igeiszerkezet-szótár (4.2. rész a 73. oldaltól).

A szótár – mely az igei szerkezetek legegyszerűbb modelljére építve készült – kézzelfoghatóvá teszi az igei szerkezeteket kinyerő algoritmus hasznosságát. A módszer igazi jelentőségét azonban az adja, hogy több irányban is kiterjeszthető. Egyrészt a modell nyelvfüggetlensége miatt megfelelő nyelvspecifikus előfeldolgozást követően számos nyelvre módosítás nélkül alkalmazható a kinyerő eljárás, így különféle nyelvű hasonló szótárak állíthatók elő. Másrészt nagyobb bonyolultságú szerkezetek – ld. például a fent említettekhez képest még egy jelzői kollokátumot is tartalmazó '*gyenge lábakon áll*' vagy '*száraz lábbal kel át vmin*' szerkezeteket –, valamint főnévi, melléknévi stb. központú szerkezetek feltérképezésére is alkalmas. Harmadrészt pedig – a modell speciális alkalmazása révén – ugyanez az említett kinyerő algoritmus alkalmassá tehető párhuzamos igei szerkezetek, azaz igei szerkezetek és fordításaik azonosítására is. Ezen a módon az eljárás képes felfedni az egymásnak megfelelő, de formailag egymásra nem hasonlító aszimmetrikus szerkezetpárokat is, valamint a szerkezetek különféle (szinonim) idegen nyelvű megfelelőit és azok gyakorisági viszonyait is. Annak, aki a dolgozat legizgalmasabb részeire kíváncsi, ajánlom figyelmébe a fenti kiterjesztéseket tárgyaló 5. fejezetet (89. oldal).

DOI:10.15774/PPKE.ITK.2011.004

Tartalomjegyzék

DOI:10.15774/PPKE.ITK.2011.004

DOI:10.15774/PPKE.ITK.2011.004

DOI:10.15774/PPKE.ITK.2011.004

DOI:10.15774/PPKE.ITK.2011.004

1. fejezet

Bevezetés

1.1. Szótárírás ma: automatizálás és frazémák

Már tíz évvel ezelőtt felmerült (Grefenstette, 1998), hogy meddig lesz szükség lexikográfusokra, manapság az is kérdés, hogy meddig lesznek egyáltalán szótárak – és itt általában a hagyományos papíralapú szótárakra gondolnak – az online világban. Az biztosnak tűnik, hogy az idegen nyelvek megismerésének vágya örök, azaz mindig lesz igény olyan eszközökre, amelyek segítik egy nyelv megértését és használatát; következésképpen olyan szakemberekre is, akik ezeket az eszközöket készítik és fejlesztik. A jövő szótárai azonban minden bizonnyal a mostani szótárakhoz képest teljesen más formában és módosult tartalommal fognak megjelenni. Egyes vélemények szerint a következő 5-10 évben a szótárírás folyamata teljesen automatizálódhat, nagyrészt ki fogjuk váltani automatikus eszközökkel a manuális lexikográfiai munkát. Már a mai szótárak is valójában lexikai adatbázisok, amiknek csak az egyik „kimenete" a klasszikus szótár, másik fontos felhasználásuk, hogy különböző nyelvtechnológiai alkalmazásokba építhetők be, ezen alkalmazások nyelvi tudását adják.

A hagyományos szótárírás nagyon munkaigényes, időigényes és költséges tevékenység. A XXI. század elején az egyik fő kérdés az, hogy a mai, nagy kapacitású számítógépek segítségével mennyire tudjuk *automatizálni* a szótárírás egyes lépéseit (Atkins és Rundell, 2008). Számos területen nagy előrelépés tapasztalható. Nagy méretű korpuszokból gyűjthetünk adatokat, az adatelemzést konkordanciák és kollokációs viszonyokat bemutató eszközök (Kilgarriff et al., 2004) segítik. A szócikkírás technikai aspektusait automatizálják a szótáríró rendszerek (dictionary writing system, DWS), formailag és szerkezetileg ellenőrizve a készülő szótárt. A valóban intelligenciát igénylő feladatok – mint a szavak, kifejezések egyes jelentéseinek meghatározása, illetve a definícióírás – természetesen ma is emberi munkával készülnek (Rundell, 2009).

A számítógépes korpuszok – mint nyelvi adatok hiteles forrása – használata a modern lexikográfiában elengedhetetlen követelménnyé vált. A COBUILD szótár óta ismert, hogy a korpuszok fontos segédeszközt jelentenek a lexikográfiai munkában. A korpuszból származó adatok, konkordanciák elemzése a hagyományosnál objektívebb

DOI:10.15774/PPKE.ITK.2011.004

munkát tesz lehetővé, eredményeképpen a szótár anyaga teljesebb lehet. A szótáríró elszakadhat idiolektusától, szembenézhet a valós nyelvhasználattal, és egyes szavaknak olyan jelentésére, használatára bukkanhat, melyek a korábbi szótárakban nem szerepelnek. A korpuszhasználat következő két alapvető módját szokás elkülöníteni (Tognini-Bonelli, 2001). A *korpuszalapú* szótárak esetében a szótárat a lexikográfusok írják, ők határozzák meg a felépítését, a korpusz pusztán segédeszköz, a korábban hagyományosan, cédulán gyűjtött idézeteket pótolja vagy egészíti ki. A *korpuszvezérelt* szótárak esetében ezzel szemben a korpusz nem csupán az alkalmas idézeteknek, hanem a szótár teljes anyagának forrása, a korpuszból nyert adatok határozzák meg a szótár struktúráját és tartalmát, így a nyelv korpuszban megjelenő szerkezete közvetlenebbül tükröződik a szótár szerkezetében. Az első korpuszvezérelt szótár a Cobuild (Sinclair, 1987). Szerkesztői a szócikkek belső elrendezésében elsődlegesnek tekintették a gyakorisági szempontot, a korpuszbeli gyakoriság csökkenő sorrendje szerint közölték a szavak jelentéseit. Ezt a megoldást az a megfigyelés indokolta, hogy az átlagos szótárhasználók rendszerint csak az elsőként megadott jelentést olvassák el, a legritkább esetben olvasnak végig egy sok jelentésből és aljelentésből álló szócikket. A pusztán gyakoriságra alapozott megoldásokat ugyan számos kritika érte, a korpuszvezérelt lexikográfia eredményei mégis sok tekintetben forradalmasították a szótárkészítést (Rundell, 1998).

Az egyik ilyen eredmény a *toz@több szóból álló lexikai egység* – kollokációk, idiomatikus kifejezések, állandósult szókapcsolatok, frazémák – jelentőségének felismerése és a korábbinál sokkal hangsúlyozottabb megjelenítése az új szótárakban. Sinclair (1998) úgy látja, hogy a nyelv valójában részben előre megkonstruált szókapcsolatokból épül fel, nem pedig egyes szavakból. A korpuszvezérelt szótárírás tapasztalatait így foglalta össze (saját fordítás):

> „A lexikográfia számos régóta elfogadott hagyománya megkérdőjeleződött: például az, hogy egy szónak inherensen van egy vagy több jelentése. A munkahipotézis az volt, hogy ha ezeket a jelentéseket értelmezzük, vagy többnyelvű szótár esetén megadjuk az ekvivalensét, és jobb szótáraknál még példákkal is ellátjuk, a lexikográfus munkája készen van. Bebizonyosodott azonban, hogy ez a gyakorlat képtelen a markáns, ismétlődő minták kezelésére, amelyek – mint azt a korpuszelemzés megmutatta – jelen voltak a szövegek nyelvhasználatában: a jellegzetes szövegkörnyezet messze fontosabbnak bizonyult, mint az a kérdés, hogy hány jelentése is van a szónak és ezek a jelentések milyen viszonyban vannak egymással... a legtöbb jelentés realizációjához szükséges, hogy egynél több szó jelenjen meg a szövegben."

Sinclair (1998, 2. oldal) végeredményben tehát arra a következtetésre jut, hogy a szó nem a legjobb kiindulópont a jelentés megragadáshoz, mivel az aktuális jelentés rendszerint szavak bizonyos kombinációjával realizálódik.

A komplex, több szóból álló lexikai egységek szótárban való megfelelő súlyú reprezentálását a szótári médium átalakulása is elősegíti. A nyomtatott szótáraknál mind a terjedelmi korlátok, mind a több szóból álló lexikai egység következetes elhelyezé-

DOI:10.15774/PPKE.ITK.2011.004

sének problematikája önkorlátozásra késztette a szótárírókat. Az elsődlegesen számítógépen publikálandó szótárak esetében ezek a korlátok már sokkal rugalmasabbak, annak sincs akadálya, hogy egy nyomtatott szótár CD melléklete lényegesen bővebb anyagot tartalmazzon. A több elemű lexikai tételek a számítógépen minden nehézség nélkül megtalálhatók, függetlenül attól, melyik elemük szócikkének részletei. Ennek köszönhetően mind a kétnyelvű, mind az egynyelvű szótárakban egyre gazdagabban szerepelnek nem csak az idiomatikus kifejezések, hanem a legkülönfélébb gyakran együttesen előforduló szabad szókapcsolatok is.

Az elmúlt években, több évtizednyi szünet után (O. Nagy, 1966), a magyar lexikográfiában is egyre nagyobb teret kap a különféle szókapcsolatok összegyűjtése, szótárba szerkesztése és elemző kutatása. A kollokációk kezelésének igénye az Akadémiai Nagyszótár munkálatai során is felmerült korábban (Pajzs, 2000, 2002), az egynyelvű lexikográfia kollokációkkal kapcsolatos legfrissebb eredményei közül pedig az alábbiakat kell megemlítenünk: Bárdosi (2003), Forgács (2003), T. Litovkina (2005), Forgács (2007), Bárdosi (2009). Bár a korpuszok használata már e szótárszerkesztőknek sem idegen, ők még általában a sajátos értelműnek tekinthető állandósult szókapcsolatok gyűjtésére és értelmezésére, illetve példákkal való illusztrálására helyezik a hangsúlyt, azaz a hagyományosabb korpuszalapú megközelítéssel dolgoznak.

Ha a gyarkorlati, kézzel fogható végtermék felől tekintünk rá, akkor jelen dolgozat témája egy új, korpuszvezérelt szótárkészítési módszer bemutatása, illetve annak alkalmazása egy konkrét szótár esetében. A módszer kulcslépése a már említett jellegzetes igei szerkezeteket kinyerő eljárás. Amint látni fogjuk, módszerünk illeszkedik a fent leírt két fő fejlődési irányhoz. Egyrészt nyelvtechnológiai eszközök kiterjedt használatával a szorosan vett nyelvi elemzésen túl egy konkrét lexikográfiai részfeladatot, nevezetesen az anyaggyűjtés feladatát *automatikusan* végezzük el: automatikusan dől el, hogy mi kerül be a szótárba és mi nem. Másrészt a többszavas és egyszavas nyelvi elemeket egységes keretben kezeljük, ezzel a többszavas kifejezéseket teljes jogú lexémákként a szótárkészítési folyamat középpontjába állítjuk. A szótárkészítő eljárás váza a következő: az első szakaszban nyelvtechnológiai eszközök segítségével, valamint egy speciális lexikális kinyerő eljárással korpuszból előállítjuk a nyers szótárat; a második szakaszban pedig ezt manuális munkával javítjuk és véglegesítjük. Azt vizsgáljuk, hogy meddig tudunk eljutni automatikus eszközökkel, azaz mennyire tudjuk csökkenteni a szükséges manuális lexikográfiai munka mennyiségét. Munkánk tehát egy kis lépés az automatizált lexikográfia felé.

1.2. Célkitűzés

Kutatásom célját egy mondatban foglalom össze, majd a kiemelt aspektusok kifejtése következik. **Kutatásom célja tehát egy olyan *nyelvfüggetlen* modell és módszer kifejlesztése, aminek segítségével *automatikus* úton lehet egy nyelv *igei szerkezeteinek* adatbázisát, szótárát létrehozni *korpuszból* kiindulva.**

A korábbi meggondolások alapján nyilván *korpuszból* indulunk ki, ha lexikai erőforrást akarunk építeni, egy automatikus nyelvfeldolgozó eljárás nyelvi adatainak forrá-

DOI:10.15774/PPKE.ITK.2011.004

sa legtöbbször a korpusz. Nem egyszerűen korpuszból indulunk ki, hanem szigorúan korpuszvezérelt módszertant követünk, amint ezt az 1.4.1. részben látni fogjuk.

Automatikusan fog előállni a nyers szótár egy speciális lexikai kinyerő eljárás segítségével, azaz a szótári anyaggyűjtés teljesen automatikusan történik. A szótár tényleges kiadásához lexikográfusok által végzett kézi ellenőrzés szükséges, ez a munka azonban nagyságrendileg kevesebb, mint ami egy teljes szótár hagyományos előállításához kellene. Az automatikus anyaggyűjtés tehát hozzájárul a gyorsabb és olcsóbb szótárkészítéshez.

A szótár alapelemei, „lexémái" nem szavak, hanem *igei szerkezetek* lesznek. A megnyilatkozások, mondatok általában egy központi igéből és annak bővítményeiből állnak, mondataink lényegében rendre egy-egy igei szerkezet megvalósulásai. Ez indokolja, hogy egy általános célú szótárban az igei szerkezetekkel foglalkozzunk. A szótári mikrostruktura nemcsak, hogy tartalmazza a többszavas kifejezéseket (frazeológiát), hanem kifejezetten frazeológia-központú, tekintve, hogy az alapelemek szókapcsolatok, frazémák.

Az igei szerkezetek formai leírására egy olyan függőségi nyelvtan alapú általános modellt alakítunk ki, mely *nyelvfüggetlen* módon képes megragadni az igei szerkezeteket. A kulcselem az, hogy nyelvfüggő korpuszból nyelvfüggő feldolgozó lépésekkel nyelvfüggetlen korpuszreprezentációt fogunk előállítani. Bármely nyelvű, ilyen reprezentációvan bíró korpuszon már közvetlenül futtatható a kinyerő eljárás, mely így tehát az egységes reprezentáció révén tud nyelvfüggetlen lenni. Ez a modell kiterjeszthető egyéb struktúrákra is. E kiterjesztés által eljárásunk nemcsak az igei szerkezetek kinyerésére lesz alkalmas, hanem valójában egy általános szótárépítő algoritmusnak tekinthető. A nyelvfüggetlenség kis nyelvek szótárainak hatékony és olcsó elkészítéséhez járulhat hozzá.

A nyelv- és korpuszfüggetlenség valamint az adatközpontúság révén a módszer *rugalmas*, azaz lényegében tetszőleges modell szerinti reprezentációjú korpuszból – például szaknyelvi korpuszokból – „gombnyomásra" előállítható a nyers lexikai adatbázis.

Megjegyzendő, hogy a szótári definíciók automatikus megalkotása nem volt célkitűzés, így a létrejött szótárban nem szerepelnek definíciók, a szótár a jellegzetes igei szerkezeteket mutatja be, a szerkezeteket és jelentésüket gondosan válogatott autentikus korpuszpéldák világítják meg. Látni fogjuk, hogy az effajta szótárnak is számos hasznos felhasználási lehetősége van.

1.3. A kapcsolódó szakirodalom áttekintése

A dolgozat folyamatosan építkezik, egymásra épülő, de viszonylag független és több kutatási területet felölelő fejezetekből áll. Nem tartottam hasznosnak, hogy az egymással nem szorosan kapcsolódó szakirodalmi utalásokat egy közös fejezetbe gyűjtsem. A dolgozat természetéhez jobban illő módon a korábbi megközelítések és eredmények, megfontolások a dolgozat különböző fejezeteiben, mindig a megfelelő résznél találhatók. Ez a könnyebb érthetőségét is biztosítja, mivel mindig csak a szükséges fogalmak

DOI:10.15774/PPKE.ITK.2011.004

bevezetése, és a szükséges előzmények tárgyalása után kerül sor az aktuális részhez kapcsolódó irodalom tárgyalására.

E helyen csak utalunk a dolgozat azon pontjaira, ahol lényeges szakirodalmi kapcsolatok bemutatása található. A különféle többszavas kifejezésekkel foglalkozó magyar lexikográfiai munkákat a 13. oldalon említettük röviden. A módszertani fejezetben érintjük a többszavas kifejezés bigram-központúságát és ennek kritikáját (18. oldal); a bevezető elején példaként említett bonyolultabb igei szerkezetekkel foglalkozó kutatásokra a magyar vonatkozásokkal együtt a 24. oldalon térünk ki. A korpusz egy igei szerkezetet tartalmazó egységekre bontása kapcsán a korábbi tagmondatra bontó eljárásokat a 35. oldalon mutatjuk be. A bővítmények lényegességének mérésére használt mérték a (Kilgarriff és Tugwell, 2001) cikkből való, részletesen ld. a 45. oldalon. A jellegzetes igei szerkezeteket kinyerő algoritmushoz az alapötlet a (Zeman és Sarkar, 2000) cikkből származik, ennek ismertetése az 55. oldalon található. A többszavas kifejezéseket kinyerő eljárások kiértékelésének módszereire és korábbi eredményekre a 63. oldalon térünk ki.

1.4. Módszertan

A bevezető rész második felében a kutatás módszertani megközelítéséről szólok, illetve ismertetem azokat az elvi megfontolásokat, melyek a kutatás során vezérfonalul szolgáltak.

1.4.1. Korpuszvezéreltség

A XX. század második felében a nyelvészet fő irányvonalát a generatív nyelvészet jelentette, de mindig jelen volt ezzel párhuzamosan az adatközpontú, korpuszokat használó megközelítés is. (Firth, 1957) szerint *„You can know a word by the company it keeps"*, azaz egy szót legjobban a környezete alapján ismerhetünk meg. A korpusznyelvészet hagyománya a generatív nyelvtan igen/nem grammatikalitási döntéseivel szemben a gyakorisági szempontok fontosságát emeli ki, illetve a valós, élő nyelvhasználat vizsgálatát tartja elsődlegesnek az introspekcióval és a konstruált példamondatok alkalmazásával szemben. Mára a korpuszok alapvető, széles körben használt eszközzé váltak a nyelvtudomány valamennyi területén, azok az állítások az igazán hitelt érdemlőek, melyeket korpuszból származó nyelvi adatokkal támasztanak alá, a korpuszkutatás a nyelvészet egyik kulcsterületévé vált (Teubert, 2005).

Jól elkülöníthető a korpuszok felhasználásának alábbi két módja. A *korpuszalapú* (corpus-based) felfogásban a korpusz segédeszköz, mely empirikus adataival támogatja az intuíciót, mérhetővé teszi a nyelvi jelenségeket, meglévő elméleteket bizonyít/cáfol. A radikálisabb *korpuszvezérelt* (corpus-driven) megközelítésben viszont a korpusz maga szolgáltatja az „elméletet", a nyelvész előzetes feltevések és elvárások nélkül fordul az adatokhoz. Minden következtetést kizárólag korpuszmegfigyelésekből von le, minden állítás empirikus alapokon: a korpusz-megfigyeléseken nyugszik (Tognini-Bonelli, 2001).

DOI:10.15774/PPKE.ITK.2011.004

1. Bevezetés

A korpuszok használata a különböző elméleti és alkalmazott nyelvészeti vizsgálódások során nem új ötlet. Már Simonyi Zsigmondnál tisztán megjelenik az adatközpontú felfogás a nyelvtanítás kapcsán. „Simonyi új grammatikai módszert akar behozni, könyve inductive halad, azaz a példákból kiindulva tanítja a szabályt, nem pedig dogmatice. A grammaticát tehát valami olvasmány alapján akarja előadni, úgy hogy a szabályokat a tanár tanítványai közreműködésével vonhatja le ésszerű következtetések útján." (Riedl, 1882) A lényegi változás az, hogy a mai informatikai környezet lehetővé teszi, hogy nagy méretű korpuszokat építsünk és hatékonyan kezeljünk nagy mennyiségű nyelvi adatot. Ma viszonylag kis ráfordítással vizsgálhatók nagy méretű szövegek, ami korábban elképzelhetetlen volt.

A korpuszhasználat tehát az utóbbi időben a nyelvészet szinte minden területén hódít, mégis tapasztalható ellenérzés, amit általában úgy fogalmaznak meg, hogy a korpusz „csal" vagy „félrevezet". Jellemző a két alábbi mondat, mindkettő magyar nyelvész szájából hangzott el: „A korpusznál jobban semmi nem vezetheti félre az embert." illetve „Manapság már a tisztán introspektív nyelvészre nézünk furán." (azaz az olyanra, aki sosem ellenőrzi az elméleti hipotéziseket korpuszból származó valós nyelvi adatokon). Fontos megjegyezni, hogy nem állja meg a helyét az a széles körben elterjedt vélekedés, miszerint egy jelenség korpuszbeli hiánya nem értékelhető negatív evidenciaként. Bizonyos esetekben statisztikailag biztosra vehető egy adott jelenség nem-létezése (Stefanowitsch, 2006). Természetesen egy korpusz mindig véges, és nem képes az elvben interpretálható megnyilatkozások sokaságát tükrözni, csak a valóban megjelenő, jellemző nyelvi formákról tud képet adni. Ez a kép azonban a korpuszméret növelésével egyre pontosabb a viszonylag ritkább jelenségek tekintetében is.

A manapság elérhető korszerű, nagyméretű korpuszok elég pontos képet adnak a nyelvről, de méretükből adódóan a legtöbb esetben képtelenség a belőlük nyert nagy mennyiségű releváns adat manuális feldolgozása, „átolvasása". Olyan kutatóeszközre van szükség, amely egy bizonyos nyelvi jelenségről valamiképpen összegzi a korpuszokból leszűrhető tényeket, és ezt olyan formában adja a kutatók elé, hogy hatékonyan felhasználhassák adott nyelvészeti kérdések vizsgálatához, elméleti meggondolások alátámasztására, illetve cáfolatára. Az egyik első ilyen eszköz az ún. *Sketch Engine* (Kilgarriff et al., 2004). Ez a rendszer egy tömör táblázatban foglalja össze a lekérdezett szó statisztikailag lényeges kollokációit, grammatikai viszony szerint csoportosítva. Jelen kutatásnak is fontos eredménye lesz egy hasonló speciális korpuszlekérdező eszköz, mely az igei szerkezetek vizsgálatát teszi lehetővé (3.2. rész).

Hogyan fedhetjük fel a megnyilatkozások jelentését, hogyan érthetjük meg a megnyilatkozásokban kódolt üzenetet? Az általánosan elfogadott hagyományos generatív keret szerint: a megértéshez először az egyes szavak jelentését kell egyértelműen meghatároznunk, majd ez alapján a nagyobb szerkezeti egységek, mondatok szerkezeti felépítését figyelembe véve juthatunk el azok jelentéséhez. Szükséges a szöveg szintaktikai elemzése, az igei vonzatkeretek felderítése valamint az egyes szemantikai összetevők azonosítása, az argumentumszerkezet feltérképezése. A korpusznyelvészet elveit és küldetését összegző cikkében Wolfgang Teubert a jelentésnek a fentitől gyökeresen eltérő megközelítését fogalmazza meg (Teubert, 2005): „A jelentés körülírás." („Meaning is paraphrase.") E felfogás szerint adott jelentéssel bíró egység („unit of meaning") jelentését az egység körülírásai, átfogalmazásai adják, másképp

DOI:10.15774/PPKE.ITK.2011.004

pen azon megnyilatkozásoknak az összessége, melyekben az adott egységről szó esik. („The meaning of the unit lemon is everything that has been said about lemons.") Teubert két dolgot mond tehát: adott jelentéssel bíró egység jelentését (1) az egység átfogalmazásai adják; (2) azon megnyilatkozásoknak az összessége adja, melyekben az adott egységről szó esik. Itt a hagyományossal ellentétes irány rajzolódik ki: mintegy a mondatokból, a többszavas egységekből származtatjuk a szavak jelentését.

1.4.2. Többszavas kifejezések

Minden nyelvben vannak olyan több elemből álló nyelvi alakulatok, amelyek az elemzés valamely szintjén egy egységként viselkednek. A szemantikai szinten ilyen az, mikor több egymás melletti (vagy egymás közelében lévő) elem együttesen hordoz olyan speciális jelentést, mely az elemek jelentéséből és kapcsolódási módjukból nem vezethető le közvetlenül. Az ilyen egységeknek – a *többszavas kifejezés*eknek (továbbiakban TSZK-k) – a szó szerinti jelentése mellett (mely sok esetben szinte el is tűnik) van nemkompozicionális vagy idiomatikus jelentése is. Ezt a nem kikövetkeztethető jelentést ismernünk kell, ha intelligens módon akarjuk feldolgozni a szövegeket, legyen szó az NLP bármely területéről, az információvisszakereséstől egészen a gépi fordításig. A TSZK-k a nyelvtan és a lexikon határterületén helyezkednek el, ez lehet az oka annak, hogy a számítógépes nyelvfeldolgozásban a legutóbbi időkig marginális jelenségnek, kivételnek tartották a TSZK-kat, jelentőségüket alábecsülték (Sag et al., 2002). Valójában a TSZK-k száma igen nagy, egy mérés szerint folyó szövegben az igék legalább egyötöde TSZK alkotórésze (Kaalep és Muischnek, 2008).

A TSZK-k definíciója a következőképpen fogalmazható meg (Sag et al., 2002; Oravecz et al., 2004, 2005):

1. definíció. *Többszavas kifejezés (TSZK).* Idioszinkratikus értelemmel rendelkező szósor, ami a nyelvi elemzés valamely szintjén egy egységként jelenik meg.

Eszerint a TSZK-k szósorok, azaz mindenképpen tartalmaznak szóhatárt (szóközt). Az alábbi jellemző tulajdonságokkal rendelkezhetnek:

- jelentésük nem teljesen kompozicionális;
- formájuk többé-kevésbé rögzített, rigid, variabilitásuk csökkent;
- a nyelv bizonyos (pl.: szintaktikai) szabályait megsértik.

Az idiomatikusság, idioszinkratikus jelentés nem bináris tulajdonság, megfigyelték, hogy e tekintetben inkább fokozatosságról beszélhetünk, a TSZK-k elhelyezhetők egy idiomatikusság szerint folytonos skála mentén (McCarthy et al., 2003). A számítógépes nyelvészetben elfogadottá vált, hogy TSZK-knak alábbi osztályait különítjük el (Sag et al., 2002; Oravecz et al., 2004, 2005; Kaalep és Muischnek, 2008) nagyjából a csökkenő idiomatikusság szerint:

1. teljesen rögzült kifejezések – pl.: '*ad hoc*', angol összetett szavak;
2. idiómák – pl.: '*felveszi a kesztyűt*';
3. ige + partikula szerkezetek, igekötős igék – pl.: '*elárul vmit*';
4. kiüresedett, „funkcióigés" kifejezések – pl.: '*döntést hoz*';

DOI:10.15774/PPKE.ITK.2011.004

1. Bevezetés

5. intézményesült kifejezések, azaz olyan szókapcsolatok, melyek kompozicionálisak viszont tagjaik nem cserélhetők fel rokonértelmű szóval – pl.: '*fáj a feje*'.

A TSZK-k szokásos kezelési módja, hogy lexikonban tároljuk őket a megfelelő idiomatikus jelentéssel együtt, és szükség esetén kiolvassuk őket (Kis et al., 2004). A klasszikus feladat tehát egy ilyen lexikon felépítése, az adott nyelv lehetőleg összes TSZK-jának összegyűjtése. A TSZK-kat általában nehéz egzakt kritériumok alapján kategóriákba sorolni, sőt egyáltalán azonosítani, a lexikonban való tárolásukhoz pedig alkalmas reprezentáció szükséges.

Az utóbbi évtizedben jelentős mértékű kutatómunka folyt különféle nyelvek vonatkozásában ebben a témában. Az alkalmazott módszerek legnagyobb része egy sémát követ: arra építve, hogy a TSZK-k elemei a vártnál, a véletlenszerűnél gyakrabban fordulnak elő együtt, különféle *asszociációs mérték*eket alkalmaznak, melyek az együtt előfordulás erősségét mérik. Az asszociációs mértékek 2×2-es kontingenciatáblán alapulnak, *két* elem közötti viszonyt tudnak megragadni, azaz a kétszavas kifejezések (bigramok) vizsgálatára alkalmazhatók közvetlenül. Természetesen számos fontos típusa van a két elemből álló TSZK-knak: ilyen például az univerzálisnak mondható ige+tárgy, melléknév+főnév szerkezet, vagy például angolban a főnév+főnév formában megjelenő összetett szavak.

A többszavas kifejezésekkel foglalkozó szakirodalom legnagyobb része valóban a két elemű, két tagból álló kifejezésekkel foglalkozik, ahogy ez az egyik jelentős áttekintő munka címében is megjelenik: *The Statistics of Word Cooccurrences: Word Pairs and Collocations* (Evert, 2005). Siepmann (2005, 412. oldal) szerint általánosan elfogadott a kutatók között, hogy a kollokációk bináris egységek. Számtalan asszociációs mértéket dolgoztak ki melyekkel két tag közötti kapcsolat szorossága mérhető, Pecina (2008) 55 különböző ilyen mérték valamint a gépi tanulással kialakított kombinációik teljesítményét veti össze. A kettőnél több tagú kifejezések kezelésével ritkábban foglalkoznak, az ide tartozó módszerek Seretan (2008, 5.1 fejezet) szerint három csoportra oszthatók: egyrészt megpróbálhatjuk az asszociációs mértékeket kettőnél több elemre kiterjeszteni; alkalmazhatunk iteratív kollokációkinyerő módszereket, ahol a már kinyert kéttagú kollokációk a következő iterációban összevont elemként egy nagyobb kiterjedésű kollokáció részét képezhetik; valamint a kinyert bigramokat utólag feldolgozva is következtethetünk bizonyos többtagú kollokációk meglétére.

A két elemre koncentráló felfogás egyértelműen leszűkítő, mert bár a TSZK-k alapesete valóban a kételemű szerkezet, nyilvánvalóan számos többelemű TSZK is létezik, álljon itt most illusztrációként egy nagyon egyszerű háromelemű angol példa:

(1) '*get rid of*' (megszabadul vmitől)

A TSZK-kinyerő módszerek felé megfogalmazzuk az alábbi elvárást, mely a fent vázolt felfogást szeretné kitágítani: a kinyerés során ne kössük meg előre a TSZK elemeinek számát, legyen az algoritmus feladata, hogy „kitalálja", hogy hány (természetesen akár kettőnél több) elemű egy adott TSZK.

Megfigyelhetjük, hogy a többszavas kifejezések bizonyos elemei „tartalmi" elemek, mások viszont csak valamiféle (szintaktikai) „viszonyt" fejeznek ki vagy jelölnek két

DOI:10.15774/PPKE.ITK.2011.004

tartalmi elem között. Arra gondolunk, amit fenti (1) példa esetében láttunk, ti. hogy itt a '*get*' és a '*rid*' tartalmi (teljes jogú, önmagában megálló) elem, az '*of*' viszont egy olyan elem, mely két másik elemet kapcsol össze, jelen esetben a '*rid*'-et egy kifejezésen kívüli elemmel (ti. *amitől* megszabadul vki). Így az '*of*' elöljáró felfogható e két tartalmi elem viszonyát kifejező nyelvi eszköznek. E fogalmakra még visszatérünk, most nézzük az alábbi példákat:

(2) '*beleüti az orrát vmibe*'

(3) '*szó van vmiről*'

(4) '*zur Verfügung stellen*' (rendelkezésre bocsát)

A (2) példában tartalmi elem a '*beleüt*' és az '*orr*', a '*-ba/-be*' rag pedig – hasonlóan az említett angol '*of*' elöljárószóhoz – nyilvánvalóan viszonyt jelöl, függetlenül attól, hogy a magyar ezt kötött morfémával fejezi ki. A (3) példában hasonlóan tartalmi elem a '*szó*' és a '*van*', a '*-ról/-ről*' rag pedig viszonyjelölőként része a TSZK-nak.

A (4) számú német példa egy olyan cikkből (Evert és Krenn, 2001) származik, melyben elöljárószó+főnév+ige hármasokat vizsgáltak. Tartalmi elem a '*Verfügung*' és a '*stellen*', a '*zur*' pedig e két elemet összekötő, azaz a TSZK-n belüli viszonyt jeleníti meg (ez tehát fontos eltérés az előző két szerkezetben említett viszonyjelölőtől!). Első pillantásra talán fel sem tűnik, de ez a TSZK nem teljes, hiányos. Két fontos elem is hiányzik belőle: a tárgy illetve a részeshatározó viszonyjelölője, hogy ti. *mit* és *kinek* bocsátanak rendelkezésére. Ez olyan típusú hiba, mintha az (1) példából az '*of*' a (2) példából a '*-ba/-be*' vagy a (3) példából a '*-ról/-ről*' maradna el. A hiba oka pontosan az, hogy a cikkben a vizsgált TSZK-k körét eleve korlátozták az említett elöljárószó+főnév+ige hármasokra, így esély sem volt az ettől eltérő struktúrájú TSZK-k megjelenésére.

Ezzel kapcsolatos a másik elvárás, amit a TSZK-kinyerő eljárások felé megfogalmazunk, hogy az algoritmus „fedezze fel", hogy egy TSZK-ban csak bizonyos viszony inherens rész, vagy az adott viszonyhoz kötődő tartalmi elem is.

A dolgozatban egy olyan igei szerkezeteket kinyerő eljárást fogok bemutatni, mely a fenti szakaszban megfogalmazott két elvárásnak megfelel.

1.4.3. Függőségi elemzés

A magyar nyelv szórendje szabad, legalábbis abban az értelemben, hogy a mondatban az ige és bővítményei szinte tetszőleges sorrendben elhelyezkedhetnek, közéjük egyéb szereplők ékelődhetnek. Más szóval: az említett TSZK-k – (2) és (3) példa – lehetnek folytonosak és megszakítottak, bármilyen sorrendi variánsban előfordulhatnak. A szórendi variabilitás kezelése nem oldható meg úgy, hogy az TSZK-k összes sorrendi variációját nyilvántartjuk, sokkal hatékonyabb, ha a nyelv természetéhez jobban illeszkedő függőségi viszonyokra alapozhatunk, a magyar nyelv leírására a *fuzggőségi nyelvtan@függőségi nyelvtan* (Prószéky et al., 1989; Koutny és Wacha, 1991; Oravecz et al., 2004, 2005) nyelvelméleti keretet választjuk.

DOI:10.15774/PPKE.ITK.2011.004

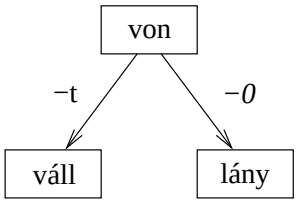
1. Bevezetés

Korábban már folytak kutatások egy magyar függőségi nyelvtan irányában (Koutny és Wacha, 1991; Prószéky et al., 1989). Központi elemnek már ez a javaslat is az igét tekinti „nagy bővítményfelvevő képessége" miatt. Az igéhez vonzatok és szabad határozók kapcsolódnak, a formai viszonyokat esetvégződések (és névutók) jelenítik meg – szemben az indoeurópai nyelvekkel, ahol prepozíciók vannak és a sorrendnek van funkciókijelölő szerepe. Fontos megjegyezni, hogy jelen dolgozatban kizárólag formai oldalról közelítjük meg a dependenciaviszonyokat, azaz alanyi, tárgyi valamint különféle „esetragi" ('*-ban/-ben*'-i, '*-ról/-ről*'-i stb.) és névutói viszonyokról lesz szó. Nem foglalkozunk a szemantikai dependenciával, a thematikus szerepekkel, melyeknek formai megjelenése változatos lehet. Ennek következménye, hogy megközelítésünkben a vonzatok és a szabad határozók nem különülnek el közvetlenül.

A függőségi leírásban általában szavak szoktak lenni az alapelemek, ugyanakkor elengedhetetlen – az előző szakaszban már érintett – viszonyt kifejező elemek és tartalmi elemek szétválasztása. Mivel a magyarban a viszonyjelölők általában a tartalmi elemek végén lévő toldalékok, szokatlan, de kifejezetten alkalmas választás, ha a *morfémák* lesznek az alapelemeink. A morféma alapelemnek választása több szempontból hasznos döntés: a viszonyokat, viszonyjelölőket elválaszthatjuk a tartalmi elemektől (ti. az esetragokat a szótól, amin megjelennek); lehetővé válik a nem-folyamatos elemek, sorrendi variációk kezelése; a szóalakok egyébként sem lehetnének alapelemek kezelhetetlenül nagy számuk miatt.

Mondatok és különféle TSZK-k ábrázolására egyaránt a *fuzggőségi fa@függőségi fa* tűnik jó reprezentációs eszköznek. A tartalmi elemek a csomópontokba, a viszonyjelölők pedig az élekre kerülnek. Az (5) példamondat függőségi fája az 1. ábrán látható, a 19. oldalon látható (2) és (3) szerkezet függőségi fája pedig a 2. ábrán.

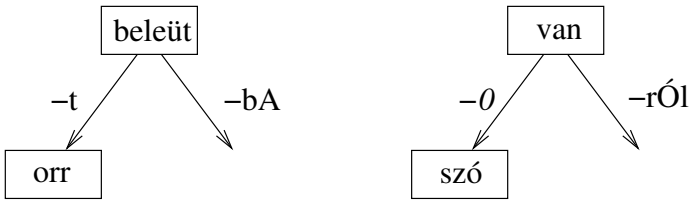
(5) '*A lány vállat vont.*'



1. ábra. '*A lány vállat vont.*' mondat függőségi fája. Az igén kívül két tartalmi elemet ('*lány*', illetve '*váll*'), valamint két (alanyi ét tárgyi) viszonyjelölőt látunk. A viszonyjelölők itt esetragok, közöttük zérómorféma – a magyar alanyeset jelölésében (jele: *-0*) – is előfordulhat.

Lényegében tehát *egyfajta* dependenciaviszonyt kezelünk: az ige és a névszói csoport bővítmény közötti relációt. Ez a relációtípus nagyon gazdag, számos alesete van az egyes eseteknek és névutóknak megfelelően. Annyira gyakori relációtípus ez, hogy az igeközpontú TSZK-k jelentős részénél megtaláljuk ezt a relációt, azaz ha csak az ilyen relációval bíró igeközpontú TSZK-k kinyerésével foglalkozunk, akkor is megkapjuk lényegében az összes ilyen szerkezetet. A TSZK-k kigyűjtésekor tehát nem a szoká-

DOI:10.15774/PPKE.ITK.2011.004



2. ábra. A (2) és (3) szerkezet függőségi fája. Jól látszik, hogy mikor része a szerkezetnek a tartalmi elem, és mikor csak a viszonyjelölő. A szerkezetben kollokátumként megjelenő szót (‘*orr*’ illetve ‘*szó*’) is viszonyjelölő kapcsolja az igéhez.

sos megközelítést követjük, mely csak a szavak egymás-mellettiségét tekinti, hanem számunkra egy TSZK elemei mindig konkrét függőségi viszonyban vannak egymással (Debusmann, 2004), illetve ezek a függőségi viszonyok maguk is teljes jogú elemei lesznek a TSZK-knak.

1.4.4. Többmorfémás kifejezések

A TSZK-k kapcsán eddig mindig nyelvi *elem*ekről volt szó, pedig a többszavas kifejezés terminus egyértelműen részt vevő *szavak*ra utal, és valóban így is szokás értelmezni. Az előző fejezetben láttuk, hogy az agglutináló nyelvekre morféma-alapú megközelítést érdemes alkalmazni. Már az (2) és (3) példából (ld. a 2. ábrát is) látszódott, hogy a viszonyjelölő morfémák saját jogukon képesek több elemből álló speciális jelentésű egységekben részt venni. Az ötlet tehát az, hogy az ige és bővítményei közötti viszonyokat függetlenül attól, hogy (az adott nyelv szabályainak megfelelően) hogyan jelennek meg a felszínen, bevesszük a vizsgálandó elemek közé. Nem releváns, hogy egy nyelvben adott viszonyt elöljáró (önálló szó) vagy esetrag (kötött morféma) fejez ki. A funkció azonos, és hasznos az azonos jellegű jelenségeket egységes keretben vizsgálni.

A fentiek alapján a TSZK (vagy most már TMK) új definíciója az 1. definíció egyetlen szavának megváltoztatásával a következő:

2. definíció. *Többmorfémás kifejezés (TMK).* Idioszinkratikus értelemmel rendelkező *morfémasor*, ami a nyelvi elemzés valamely szintjén egy egységként jelenik meg.

A jellegzetes tulajdonságok megmaradnak, kivéve, hogy nem releváns tulajdonság többé, hogy a szerkezet tartalmaz-e szóhatárt. A klasszikus több önálló szóból álló TSZK-k mellett most már ide tartoznak az egy szóból és egy (vagy több) esetragból álló TMK-k is, mint például (6).

(6) ‘*hisz vmiben*’

Ez a definíció kizárja a egyszerű ragozott alakokat (‘*asztalt*’), a kompozicionális jelentésű igekötős igéket (‘*bemegy*’), de megtartja a nem kompozicionális jelentésű igei

DOI:10.15774/PPKE.ITK.2011.004

szerkezeteket (2. ábra), és a (magyarban egybeírt) összetett szavakat (‘*számítógép*’). Arra is lehetőséget ad, hogy a magyar igekötős igék egybe és különírt (elváló) változatait egységesen kezelhessük, korábban kénytelenek voltunk csak az elváló változatot TSZK-nak tekinteni (Oravecz et al., 2004, 2005). Az indoeurópai nyelvekben egységesen kezelhetjük a főnévi (NP) és prepozíciós (PP) frázisokat, a főnévi csoportokból „hiányzó” elöljárót sorrendi megkötés helyettesíti. A (4) példában bemutatott hiányzó tárgy probléma is megoldódik, az ott szereplő kifejezés teljes egészében (tárggyal és részeshatározóval együtt) egy TMK-ként ábrázolható (‘*jm. etw. zur Verfügung stellen*’). A sejtés az, hogy a fenti definícióval a „valamilyen nyelven szóhatárt tartalmazó” kifejezéseket ragadjuk meg.

Nyelvtanulói szemszögből mindegy, hogy egy adott nyelvi elem szó vagy frazéma, ha fontos és gyakori, akkor szükséges az ismerete. Ez a definíció lehetőséget ad arra, hogy egységes keretben foglalkozzunk a több morfémából összetevődő speciális kifejezésekkel, függetlenül attól, hogy egy adott nyelven hány szóból állnak. Így a látóterünkbe kerülnek olyan kifejezések is, melyek – esetleg csak éppen a vizsgált nyelven – nem frazémák. Ilyenek például a fenti definíció szerint a TMK-k közé tartozó, az igén kívül csak viszonyjelölő(k)ből álló igei vonzatkeretek, mint amilyen az imént említett ‘*hisz vmiben*’ is volt. Az 1.4.2. oldalon közölt osztályozás tehát egy újabb, 6., osztállyal egészül ki. Az igei vonzatkeretekkel a TSZK-któl elkülönítve szokás foglalkozni, ez a definíció közös, általános sémába foglalja bele mindkét jelenséget.

1.4.5. Igei szerkezetek

A dolgozatban elsősorban igei szerkezetekkel – olyan többmorfémás kifejezésekkel, melyeknek a központi eleme egy ige – foglalkozunk, ezt az alapvető fogalmat tárgyaljuk ebben a szakaszban. Ezek a szerkezetek egy igéből és annak bővítményeiből állnak. (A bővítmény lehet vonzat és szabad határozó is.) Ilyen volt a fenti (1), (2), (3), (4) és (6) példa is. A lehetséges bővítmények körét leszűkítjük a névszói csoportokra – példáink csak névszói csoport bővítményt tartalmaztak –, így a következőképpen fogalmazhatjuk meg a definíciót:

3. definíció. *Igei szekezet.* Központi igéből és annak névszói csoport bővítményeiből álló többmorfémás kifejezés. Az igei szerkezeteknél nem mindig követeljük meg a jelentés idiomatikusságát, bizonyos esetekben kompozicionális szerkezeteket is ideértünk.

Fontos megjegyezni, hogy valóban az összes ilyen formájú kifejezést ideértjük a vonzatkeretektől (pl.: ‘*néz vmit*’, ‘*foglalkozik vmivel*’), az összetettebb kifejezéseken (pl.: ‘*vállat von*’, ‘*hasznot húz vmiből*’) át egészen a szólásokig (pl.: ‘*pontot tesz a végére*’, ‘*más malmára hajtja a vizet*’).

A korábbi megfontolások alapján nem váratlan, hogy a bővítményeknek formailag két típusát különböztetjük meg. Az egyik esetben csupán a viszonyjelölő képezi részét a szerkezetnek (‘*vmit*’, ‘*vmivel*’, ‘*vkinek*’). Ezek a bővítmények sok esetben az ige vonzatainak felelnek meg. Itt a bővítményként megjelenő tartalmi elem – gyakran szinte korlátozás nélkül – számos lehetőség közül választható (pl.: ‘*néz vmit*’ – ‘*képet*’, ‘*adást*’,

DOI:10.15774/PPKE.ITK.2011.004

'*lányt*', '*mennyezetet*', '*vizet*', '*eget*' stb.). A másik esetben viszont a viszonyjelölő és az általa az igéhez kapcsolt tartalmi elem – egy konkrét, kötött szó – is lényeges részét alkotja a szerkezetnek, ('*váll*' + '*-t*', '*haszon*' + '*-t*', '*malom*' + '*-ra*'). Az első esetben tehát a bővítményt az esetragja (vagy névutója) képviseli, a második esetben ezen felül még az adott esetraggal (névutóval) szereplő konkrét szó is.

A TSZK-k 17. oldalon idézett csoportosítását szemügyre véve látjuk, hogy valamennyi TSZK-csoportban találunk igei szerkezeteket: a teljes mondatként megjelenő szólások, közmondások tartoznak az 1. csoportba (pl.: '*Veri az ördög a feleségét.*'); a 2. és az 5. csoportba főként különféle igei szerkezetek tartoznak (pl.: '*bakot lő*', '*hasznot húz vmiből*', '*szerződést köt vkivel*'), az igekötős igék (3. csoport) és a kiüresedett funkcióigés kifejezések (4. csoport) pedig természetszerűleg tartalmaznak igét. Az 1.4.3. részben leírt függőségi fák egységes keretet adnak az igei szerkezetek kezeléséhez, az itt említett összes típus ábrázolható ezen a módon.

Az igei szerkezetek tehát a TSZK-k széles rétegét képviselik, az ige a tagmondat „pillére", a különféle ige-központú kifejezések a megnyilatkozások túlnyomó részét lefedik, általuk az egész nyelv struktúrájáról kaphatunk információt. Ebben a dolgozatban ezzel a széles osztállyal foglalkozom egységes keretben.

1.4.6. Komplex igék

A '*beleüti az orrát vmibe*', '*szó van vmiről,*' '*vállat von*', '*hasznot húz vmiből*' '*kétségbe von vmit*', '*kockán forog vmi*', '*górcső alá vesz vmit*' típusú szerkezeteket, melyeknek a bennük szereplő ige önálló jelentésétől eltérő együttes jelentése van, *komplex igének* nevezem. Ide tartoznak az igemódosítós igék, azok a szerkezetek, ahol az ige jelentése kiüresedett (pl.: '*moziba megy*', '*egyetemre jár*'); ahol a vonzat jelentése kiüresedett (pl.: '*útnak indul*', '*ott marad*'); valamint azok is, ahol az inkorporáció jelenségével találkozunk (pl.: '*fogat mos*') (Kálmán, 2006). Általában véve mindazok a szerkezetek ide tartoznak, amikor egy (vagy több) névszói csoport szervesen hozzátartozik az igei szerkezethez, a szerkezetnek csak a névszói csoporttal együtt van meg a speciális jelentése.

4. definíció. *Komplex ige.* Olyan igei szerkezet, melynek az ige mellett egy vagy több névszói csoport is szemantikailag szerves része. Azaz az ige és a névszói csoport együttes jelentése valamilyen mértékben nem-kompozicionális, idiomatikus. Másképp fogalmazva: ha megváltoztatjuk a névszót, akkor elvész vagy megváltozik a komplex ige együttes jelentése. (A fenti példákkal ellentétben a '*sört iszik*' tehát nem komplex ige.)

Külön kiemelendők azok a szerkezetek, melyekben mindkét említett formai bővítménytípus jelen van: ezek a *vonzatos komplex igék*. E rész elején felsorolt példák közül a '*vállat von*' kivételével mind ilyen. Az ilyen típusú szerkezetek *egyszerre* vonzatkeretek és többszavas kifejezések: a kollokációk közül (és a kollokációs szótárakból) vonzatuk miatt, a vonzatkeretek közül (és a vonzatszótárakból) pedig a jelen lévő kollokátum miatt „lógnak ki".

Nem véletlen az elnevezés. A komplex igék, annak ellenére, hogy több szóból állnak, valóban tekinthetők önálló igéknek, az igék egy csoportjának. Négy érvet sorakozta-

DOI:10.15774/PPKE.ITK.2011.004

tunk fel ennek alátámasztására: (1) egyrészt látjuk, hogy igék helyén jelenhetnek meg a mondatban (vö: '*megvizsgál vmit*' ↔ '*górcső alá vesz vmit*'); (2) másrészt az alapigétől eltérő új jelentéssel bírnak; (3) harmadrészt az alapigétől független új vonzatkerettel rendelkezhetnek: a '*részt vesz*' mellett megjelenő '*-ban/-ben*' vonzat vagy az '*hírt ad*' melletti '*-ról/-ről*' az alapige ('*vesz*' illetve '*ad*') mellett nem szerepelt; valamit (4) sokszor egy hangsúllyal ejtjük (akár egybe is írjuk) őket, ilyenkor a kötött névszó igekötőként viselkedik ('*egyetért* ↔ *egyet ért vmivel*').

A komplex igék sokkal gyakoribbak, mint azt az általános nyelvi intuíciónk sugallja. Gyakoriságuk és a fenti elméleti érvek szólnak amellett, hogy érdemes ezzel a jellegzetes, határterületre eső csoporttal külön is foglalkozni.

A többszavas kifejezések kinyerésével foglalkozó kutatásokon belül nem jelentéktelen részt képviselnek a kifejezetten a többszavas igékre, komplex igékre irányuló vizsgálatok. A figyelem a legtöbb esetben csak egy jól meghatározott szerkezettípusra irányul, erre szűkítik le a vizsgálódást (Manning, 1993). Baldwin és Villavicencio (2002) ige–partikula szerkezetekkel, Fazly és Stevenson (2006) pedig ige–főnév idiomatikus konstrukciókkal foglalkozik. Kifejezetten komplex igékkel kevés cikk foglalkozik, a 2008. évi TSZK workshop keretében készült észt nyelvre egy komplex igékkel annotált korpusz, illetve komplex igék gyűjteménye (Kaalep és Muischnek, 2008). Ebben a cikkben találkozunk a többszavas ige – az eredetiben *multiword verb* – fogalmával. Ez teljesen azonos a mi komplex ige fogalmunkkal, ami nem véletlen a magyar és az észt nyelv nagy szerkezeti hasonlósága folytán. Egy cikk tanulmányozza kifejezetten a komplex igék vonzatait, azonban mindössze a tranzitivitására vonatkozó vizsgálatokat végez (Baldwin, 2005).

Két fontos magyar nyelvre vonatkozó korábbi kutatást említek. A (Kis et al., 2004) publikációban ige+(főnév+esetvégződés) hármasokat vizsgáltak. Az általuk vizsgált hármasok az igei szerkezetek egy csoportját alkotják: a vonzat nélküli komplex igéket. Egy másik kutatásban pedig részletesen elemzik a TSZK-kinyerés különböző aspektusait, valamint egy kinyerő módszert tesztelnek amely a TSZK-k rigiditására alapul, pontosabban arra, hogy a feltételezés szerint a bennük szereplő szavak nem cserélhetők szinonimájukra (Oravecz et al., 2004, 2005).

1.4.7. Igei szerkezetek mint konstrukciók

A módszertani rész lezárásaként megemlítjük, hogy az igei szerkezetek nagy része valódi konstrukció. Konstrukció, azaz „forma és jelentés pár" (Goldberg, 2006), jelentésük a teljes formához rendelődik, nem lehet őket kisebb elemekre bontani, ha meg akarjuk tartani az együttes jelentést. Az igei szerkezetek lehetséges használati mintázatokat jelenítenek meg, és általában hozzárendelhetők az (egyszerű vagy komplex) alapige egyik jelentéséhez. Érdekes gondolat, hogy nem érdemes az alapigékhez ('*vesz*', '*ért*' stb.) tucatnyi jelentést absztrahálni, célravezetőbb, ha egyszerűen megjelenítjük az alapigéhez tartozó igei szerkezeteket, amelyek jó eséllyel egy- vagy legalábbis kevesebb jelentésűek (Kilgarriff, 1997), és jól bemutatják az alapige jelentéseit és használati módjait.

DOI:10.15774/PPKE.ITK.2011.004

Szemben az általában többjelentésű szavakkal, „a kollokációk több mint 90%-a pontosan egyjelentésű” (Yarowsky, 1993). Az igei szerkezetek, azon belül főként a komplex igék, az esetek nagy részében egyjelentésűek, a benne szereplő elemek egy kollokáció tagjaiként meghatározzák, leszűkítik az egyes elemek jelentését. Egy ige különböző vonzatkeretei, szerkezetei gyakran megfelelnek a különböző szótárbeli jelentéseknek (Briscoe és Carroll, 1997), azaz ha az összes jellegzetes igei szerkezet a birtokunkban van, akkor közülük mindig kiválaszthatjuk az épp kívánt jelentésnek megfelelőt. Ha tehát az igei szerkezeteket tesszük meg egy szótár alapegységének, a poliszémia jelentős részétől automatikusan megszabadulhatunk.

DOI:10.15774/PPKE.ITK.2011.004

DOI:10.15774/PPKE.ITK.2011.004

2. fejezet

Igei szerkezetek modellje

Az alapvető új tudományos eredményeket a 2. és a 3. fejezetben ismertetem. Jelen fejezetben először felvázolom az igei szerkezetek ábrázolására szolgáló modellt (2.1. rész), aztán arról lesz szó, hogy hogyan lehet egy morfoszintaktikailag annotált korpuszból kialakítani a modell szerinti reprezentációt (2.2. rész). A modell szerint reprezentált igei szerkezetek korpuszból való kinyerésével a 3. fejezetben foglalkozom majd.

2.1. Modell és reprezentáció

Ebben a részben a módszertani (1.4. rész) megfontolásokra építve, azok alapján kialakítom, pontosan definiálom az igei szerkezetek modelljét.

2.1.1. A modell alapfogalmai

Az ige legszorosabb környezetét a bővítményei alkotják. Absztrakt szinten egy egyszerű mondat (illetve ezzel teljesen egyenértékűen egy tagmondat) tekinthető egy központi ige és a hozzá tartozó bővítmények összességének. Egy tagmondat alapesetben pontosan egy igei szerkezetet (3. definíció a 22. oldalon) tartalmaz, ezért választottuk a tagmondatot a modell alapegységének. A továbbiakban az alábbi definíciók alapján gondolkozunk ezekről a fogalmakról:

5. definíció. *Tagmondat.* Egy igét és a hozzá tartozó bővítményeket tartalmazó nyelvi egység.

6. definíció. *Bővítmény.* A bővítmények körét a dolgozat törzsrészében leszűkítjük a névszói csoportként megjelenő bővítményekre. Fontos kiemelni, hogy a bővítmények közé számítjuk az alanyt is, mely természetesen legtöbbször névszói csoportként jelenik meg. (A modell általánosításával tetszőleges bővítmény kezelhető lesz, amint ezt majd a 5.2 fejezetben látni fogjuk.)

DOI:10.15774/PPKE.ITK.2011.004

'*A lány vállat vont.*'

tartalmi elem	viszonyjelölő
lány	∅ (alany)
váll	-t (tárgy)

'*A huszonkilenc éves Bobbi McCaughey hét és fél hónapos terhesség után császármetszéssel hozta világra a négy fiú- és három leánygyermeket.*'

tartalmi elem	viszonyjelölő
Bobbi McCaughey	∅ (alany)
terhesség	után
császármetszés	-vAl
világ	-rA
gyermek	-t (tárgy)

3. ábra. Az alapfogalmak illusztrálása két példamondaton. A kis táblázatok a megfelelő tartalmi elemeket és viszonyjelölőket tartalmazzák. Látjuk, hogy függetlenül attól, hogy az adott névszói csoport vonzat vagy szabad határozó, ugyanúgy esetragok, illetve névutók a viszonyjelölők bennük.

A bővítményeket, azaz a névszói csoportokat – számos tulajdonságukat figyelmen kívül hagyva – két legfontosabb jellemzőjükkel reprezentáljuk. A névszói csoport fő tartalmi elemével: a névszói csoport fejével, az ott megjelenő névszóval illetve a morfoszintaktikai viszonnyal, mely a csoportot az igéhez kapcsolja. A bővítmény reprezentációja tehát egy *tartalmi elemből* és egy *viszonyjelölőből* áll.

7. definíció. *Viszonyjelölő.* Nyelvi elem, mely az ige és a bővítmény közötti felszíni viszonyt megtestesíti, jelöli. A viszonyjelölőt a magyarban esetrag vagy névutó képviseli.

8. definíció. *Tartalmi elem.* A konkrét névszó, mely a névszói csoport fejét alkotja, és amit a viszonyjelölő kapcsol az igéhez.

A modell tehát kizárólag a névszói szerkezetként megjelenő bővítményeket tekinti, ezeket reprezentálhatjuk a szerkezet fejét adó szótővel és a fej esetragjával, illetve névutójával. Itt jegyezzük meg, hogy Kis et al. (2004) javaslatának megfelelően a magyar esetragokat és névutókat teljesen egyenrangúaknak tekintjük, egységesen, egy kategóriaként kezeljük. Eltekintve attól, hogy az esetragok kötött morfémák, a névutók pedig önálló szavak, szerepük azonos. Például névutók ugyanúgy képviselhetik egy ige vonzatát, mint az esetragok (pl.: '*tartozik vmi közé*', '*vki elé tár vmit*'). Koutny és Wacha (1991) szerint az esetragok és a névutók ugyanazt a funkcionális szerkezetet hozzák létre, ezért azonos módon kezelendők. A 3. ábrán bemutatunk két példamondatot, a bennük szereplő viszonyjelölőket és tartalmi elemeket.

DOI:10.15774/PPKE.ITK.2011.004

2.1.2. A tagmondat reprezentációja

A magyar nyelv diskurzus-konfigurációs nyelv, a magyar tagmondatban az ige és az egyes bővítményeket képviselő szerkezetek sorrendjét a topik-fókusz viszonyok befolyásolják (É. Kiss et al., 2003). Lényegében bármilyen sorrend előfordulhat, azaz a magyar tagmondat szórendje ebből a szempontból szabadnak tekinthető.

Reprezentációnkban nem jegyezzük fel, hogy adott bővítmény adott tagmondatban éppen hol szerepelt: a tagmondatokat tehát *halmazként* kezeljük, amiben egy ige és valamennyi bővítmény van. E felfogás miatt a reprezentáció képes kezelni a nem folytonos igei szerkezeteket, és a változó szórendű igei szerkezeteket is, melyek számos különböző felszíni formában jelenhetnek meg.

Az eddig mondottak alapján tehát a magyar tagmondat reprezentációja a következő:

tagmondat = **ige + bővítmények halmaza**
bővítmény = **viszonyjelölő + tartalmi elem**

Az, hogy a tagmondatot halmaznak fogjuk fel, megfelel a függőségi elemzéses (ld. 1.4.3. rész) megközelítésnek, mely a szabad szórendű nyelvekhez, így a magyarhoz is jól illeszkedő nyelvleírási elmélet (Prószéky et al., 1989; Koutny és Wacha, 1991). A reprezentáció által ábrázolt egységek tekinthetők 1-mélységű függőségi struktúrának is, melyben az ige a gyökér-csomópont, a tartalmi elemek a dependensek, a viszonyjelölők pedig a függőségi relációk.

A reprezentáció nyelvészeti szempontból egyfajta *kevert* szintaktikai felfogást valósít meg, mivel első szinten *függőségi* viszonyokat ábrázolunk, azonban a dependensek belső függőségi szerkezetét már nem ábrázoljuk, hanem a dependenseket *frázisokként* kezeljük. Ez jól illeszkedik a magyar nyelv szórendi tulajdonságaihoz, ugyanis a szabad szórend csak a frázisok között, a mondat szintjén érvényesül, itt megfelelő a függőségi elemzés; a névszói frázisokon belül már kötött a szórend, ott már érdemes szintaxist frázisstruktúrával megragadni.

2.1.3. A reprezentáció megjelenítése

A reprezentációkat a továbbiakban a folyó szövegben következő két mód egyikével jelenítem meg.

Ha a reprezentáció szerkezetét akarom kiemelni, akkor a '`ige=`von `-t=`váll `-0=`lány' formát használom: elöl az ige szerepel, ezt követik a viszonyjelölők (az alany jele itt a `-0`) és egyenlőségjel után a hozzájuk tartozó tartalmi elem.

Ha csak mint nyelvi egységre hivatkozom, a jobban olvasható 'von VÁLL-t LÁNY' formát használom: szintén az igét tüntetem fel először, utána a bővítmények következnek tartalmi elem szótöve + kötőjel + viszonyjelölő formában a tartalmi elemeket itt kiskapitális szedés emeli ki.

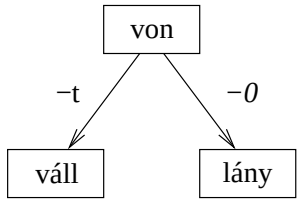
DOI:10.15774/PPKE.ITK.2011.004

A fentieken kívül a reprezentációkat természetesen függőségi fa formájában is megjeleníthetjük, ábraként. A három egymással egyenrangú, izomorf megjelenítési módot a 4. ábrán foglaljuk össze.

a) ‘ige=von −t=váll −0=lány’

b) ‘von VÁLL-t LÁNY’

c)



4. ábra. A reprezentáció három használatos megjelenítését a 3. ábra 1. mondatán mutatjuk be. Az *a)* forma bemutatja a reprezentáció szerkezetét, a *b)* forma az eredeti nyelvi formához közelebb álló, könnyebben olvasható megjelenítés, a *c)* forma pedig függőségi faként jeleníti meg a tagmondatot: itt a viszonyjelölők élekként, a tartalmi elemek csomópontokként szerepelnek.

A magyar esetragokat úgy jelenítjük meg, hogy az hangrendileg illeszkedő magánhangzó helyén mindig a hátulképzett változat nagybetűs alakját használjuk (pl.: ‘-tÓl’). A magyarban sok szerkezetnél érdemes hangsúlyozni a birtokos személyrag meglétét, a birtokos személyrag jele: ‘-A’. A ‘*csóválja a fejét*’ szerkezet megjelenítése tehát ‘csóvál FEJ-A-t’.) Névutó – és más szabad morfémaként megjelenő viszonyelölő – esetén az összetartozás jelzésére kötőjel helyett pont szerepel, pl.: ‘VÉKA·alá’. Az üres magyar alanyi esetrag el is maradhat. Ha ki akarom emelni az igekötőt, akkor ’|’ jellel választom el az alapigétől, pl.: ‘el|távolít -t’.

A megjelenítés – a halmazos felfogásnak megfelelően – nem ad információt az elemek eredeti vagy szokásos sorrendjéről. Az elemek mindig a következő rend szerint követik egymást: először az igét tüntetjük fel, utána a bővítmények következnek (az alany kivételével) a viszonyjelölő szerinti ábécésorrendben, és legvégül az alany. Ez a megjelenítés nem mellesleg közvetlenül alkalmas arra, hogy egy sor – egy reprezentáció formában számítógépen hatékonyan tároljuk, kezeljük.

A továbbiakban az igei szerkezeteket – illetve részeiket – a most ismertetett egyik forma szerint, az egyéb nyelvi példákat továbbra is ‘*aposztrófok között kurzívan szedve*’ közöljük.

2.1.4. Mit reprezentál: LSzB és LKB

Fontos tulajdonsága a modellnek, hogy segítségével a tagmondatokon kívül olyan szerkezeteket is ábrázolhatunk, melyben csak adott viszonyjelölő meglétét akarjuk kifejezni, a hozzá tartozó tartalmi elemet nem akarjuk rögzíteni. Ez például a formailag kötött vonzatok ábrázolásakor fordul elő. Ilyenkor a tartalmi elemet egyszerűen nem tüntetjük fel. A ‘*bocsánatot kér vkitől*’ szerkezet megjelenése tehát: ‘ige=kér

DOI:10.15774/PPKE.ITK.2011.004

-t=bocsánat -tÓl' vagy 'kér BOCSÁNAT-t -tÓl'. Látjuk: a '-tÓl' viszonyjelölő esetében a konkrét tartalmi elem, szó elmarad.

Itt érkeztünk el az igei szerkezetek szempontjából alapvető fogalompárhoz, melyek éppen ezt a jelenséget ragadják meg. Ti. bizonyos igei szerkezeteknek egyértelműen meghatározott inherens része egy-egy tartalmi elem (a tartalmi elem megváltoztatásával sok esetben megváltozik a szerkezet jelentése is, egy új igei szerkezetet kapunk), másoknak pedig csak a viszonyjelölő (a hozzá tartozó tartalmi elem pedig szabadon választható a szerkezet jelentésének változása nélkül). Ezen kívül hasznos, ha van arra eszközünk, hogy aktuális szándékunk szerint bizonyos esetekben a tartalmi elemet is fel akarjuk tüntetni, más esetekben pedig csak a viszonyjelölőt, függetlenül az igéhez fűződő viszony szorosságától.

9. definíció. *Lexikálisan kötött bővítmény (LKB).* Olyan bővítmény, melyben a viszonyjelölő és a tartalmi elem is szerepel. A komplex igék kötött (névszói) eleme tipikusan LKB-ként jelenik meg: 'kér BOCSÁNAT-t -tÓl' szerkezetben a 'BOCSÁNAT-t' elem LKB. Itt azt akarjuk kifejezni, hogy csak akkor teljes ez a szerkezet és csak akkor hordozza speciális jelentését, ha ez a kötött szó jelen van.

10. definíció. *Lexikálisan szabad bővítmény (LSzB).* Olyan bővítmény, melyben csak a viszonyjelölő szerepel. A vonzatok tipikusan LSzB-ként jelennek meg: 'kér BOCSÁNAT-t -tÓl' szerkezetben a '-tÓl' elem LSzB. Itt azt akarjuk kifejezni, hogy a szerkezetnek csak a viszonyjelölő része, csak az releváns, a '-tÓl' viszonyjelölőhöz kapcsolódó tartalmi elem viszont – az adott szerkezetre jellemző szematikai korlátok mellett – szinte bármi lehet.

LKB-t használunk tehát, ha egy szerkezetnek elengedhetetlen eleme az adott tartalmi elem (pl.: 'von VÁLL-t', 'jut ÉSZ-A-bA'), de akkor is ha csak valamiért hangsúlyozni akarjuk az aktuális kötött szót (pl.: 'vesz SZEKRÉNY-t', 'iszik SÖR-t'). A komplex igék (4. definíció a 23. oldalon) az igével szoros kapcsolatban álló LKB-t tartalmaznak, a vonzatos komplex igék pedig – mostani példánkhoz hasonlóan – LKB-t (kollokátumot) és LSzB-t (vonzatot) is. Utóbbi esetben a két fogalom nagyjából megfelel a *belső valencia* (LKB) és a *kuzlső valencia@külső valencia* (LSzB) fogalmának (Burger, 2003, 41. oldal).

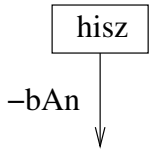
A teljesség kedvéért említjük az alábbi fogalmakat.

11. definíció. *Egyszerű ige.* Nem tartalmaz LKB-t. Például '*fut*' vagy '*néz vmit*'. Egy egy LSzB-t tartalmazó, vonzatos egyszerű ige látható az 5. ábrán.

12. definíció. *Puszta ige.* Sem LKB-t, sem LSzB-t nem tartalmaz, azaz nem komplex és vonzata sincsen. Vonzat nélküli egyszerű ige. Egyetlen (elvben) lehetséges bővítménye az LSzB alany. Ide tartoznak tehát a csak alannyal bíró igék (pl.: '*történik*') és az alanytalan igék (pl.: '*villámlik*') is.

Itt jegyezzük meg, hogy az igei szerkezetekben az alanyt csak akkor tüntetjük fel, ha LKB-ként szerepel (pl: 'kerül SOR -rA'), a nagyon sok szerkezetben megjelenő LSzB alanyt elhagyjuk, úgy is fogalmazhatunk, hogy implicite feltesszük, hogy alanya minden igének, igei szerkezetnek van.

DOI:10.15774/PPKE.ITK.2011.004



5. ábra. Az *'hisz vmiben'* vonzatos egyszerű ige függőségi fája.

13. definíció. *Igei rész.* Igei szerkezet igei része alatt az ige és az esetleges LKB-k együttesét értjük.

2.1.5. Mit reprezentál: mondatváz és bővítménykeret

14. definíció. *Mondatváz.* Egy tagmondatnak a reprezentáció által megjelenített jellemzőinek összességét (viszonyjelölők + tartalmi elemek) mondatváznak nevezzük. Tartalmazza az igét és a bővítmények halmazát, a bővítményeket a viszonyjelölők és a tartalmi elemek képviselik.

15. definíció. *Bővítménykeret* vagy *keret.* A bővítménykeret a mondatvázhoz hasonlóan egy igét és bővítmények halmazát tartalmazza, melyek az igéhez tartoz(hat)nak. A bővítménykeretben azonban LSzB-kként is megjelenhetnek a bővítmények. Minden tagmondat több bővítménykeretnek egy megvalósulása. A '*Mártonnak gólpasszt adott*' tagmondat például megvalósulása az alábbi kereteknek: 'ad -t', 'ad -nAk', 'ad -nAk -t', 'ad GÓLPASSZ-t', 'ad -nAk GÓLPASSZ-t'.

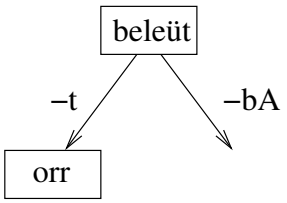
A mondatváz és a bővítménykeret fogalmát is bizonyos esetekben az ige nélkül fogjuk érteni, így fogjuk használni.

A mondatvázak természetükből adódóan csak LKB-ket tartalmaznak: '*az EU csak abba üsse bele az orrát*' tagmondat mondatváza '`ige=`beleüt `-bA=`az `-t=`orr `-0=`EU'. E mondat a '*beleüti az orrát vmibe*' vonzatos komplex ige egy megvalósulása, ez utóbbi igei szerkezet reprezentációja: '`ige=`beleüt `-bA -t=`orr', mely egy LSzB-t és egy LKB-t tartalmaz (6. ábra). Amint látjuk, az LSzB alanyt az utóbbi reprezentációban nem tüntettük fel.

16. definíció. *Típus.* A különféle bővítmények alapján az igei szerkezeteket formai alapon csoportokra lehet osztani, ezek a típusok. Az azonos típusba tartozó igei szerkezetek azonos számú LKB-t és LSzB-t tartalmaznak. A típus – jelölésére bevezetjük a `[01]` formát – két számból áll: először az LKB-k majd az LSzB-k száma következik.

A 6. ábrán látható szerkezet típusa `[11]`, a 19. oldalon látható (4) szerkezeté `[10]` – illetve az ott említett elmaradó tárgyat és részeshatározót is hozzávéve `[12]` –, az 5. ábrán látható szerkezet típusa `[01]`, a 14. ábrán (56. oldal) láthatóé pedig `[02]`. Amint látjuk, a modell alkalmas a mondatok (mondatvázak), és a korábban említett igei szerkezetek minden fajtájának ábrázolására.

DOI:10.15774/PPKE.ITK.2011.004



6. ábra. A ‘beleüt -bA ORR-t’ vonzatos komplex ige függőségi fája. A szerkezet egy LKB-t (‘ORR-t’) és egy LSzB-t (‘-bA’) tartalmaz.

2.1.6. Ige bővítményszerkezete

17. definíció. *Bővítményszerkezet.* Ige bővítményszerkezetén legfontosabb/legjellegzetesebb/legtipikusabb/leggyakoribb bővítménykereteinek összességét értjük.

A bővítményszerkezetből látszik, hogy az ige mely névszói csoport bővítményekkel szokott általában együtt előfordulni. A bővítményszerkezet fogalma implicit módon tartalmazza a korpusznyelvészet gyakorisági szempontját is: a gyakoribb bővítménykeret a fontosabb A *‘von’* ige öt leggyakoribb bővítménykerete az 1. táblázatban látható, angol megfelelőikkel együtt.

1. táblázat. A *‘von’* ige bővítményszerkezete: az öt legfontosabb bővítménykeret. A táblázat jól illusztrálja, hogy a különböző szerkezetek gyakran az ige különböző jelentéseit képviselik (vö: 1.4.7. rész), ez abból is látszik, hogy egy másik nyelvre való fordításkor magát az igét rendre különbözőképpen kell fordítanunk. (Az elöljárókat – szabad morfémák lévén – a névutókhoz hasonló jelöléssel kapcsoljuk a megfelelő tartalmi elemhez: ‘to·ACCOUNT’.)

magyar szerkezet	angol megfelelő
‘von KÉTSÉG-bA -t’	‘to question sg’
‘von VÁLL-t’	‘to shrug (one’s) SHOULDER’
‘von -t MAGA·után’	‘to entail sg’
‘von FELELŐSSÉG-rA -t’	‘to call sy to·ACCOUNT’
‘von -t’	‘to pull sg’

2.1.7. Összefoglalás

A modell összefoglalása **az 1. tézisben** olvasható a 109. oldalon. A kialakított modell az összes szóba jöhető igei szerkezetet képes ábrázolni, egységes, általános keretet biztosít az igei szerkezetek kezeléséhez. A modell elméleti szempontból függőségi leírásként fogható fel, mely jól illeszkedik a magyarban a mondat szintjén meglévő szabad szórendhez.

DOI:10.15774/PPKE.ITK.2011.004

A fentiekben egy konkrét esetre „vezettük le” (mutattuk be) a modellt: a magyar nyelv igei szerkezeteire. Ez az a szcenárió, amit részletesen kidolgozunk a következőkben, de már most előrevetítjük, hogy a modell nagy mértékben, több irányban kiterjeszthető. Látni fogjuk, hogy a magyartól szerkezetében eltérő nyelvek kezelésére is alkalmas lesz (5.1. rész) valamint, hogy jóval bonyolultabb függőségi szerkezeteket is képes leírni, amennyiben azok beleillenek egy sokkal általánosabb felépítésű keretbe (5.3. rész).

2.2. A reprezentáció megvalósítása

Ebben a fejezetben arról lesz szó, hogy klasszikus nyelvfeldolgozó eszközök segítségével hogyan tudjuk egy korpusznak a modell szerinti reprezentációját kialakítani.

A modellből következik, hogy a számítógépes feldolgozás során milyen lépéseket szükséges megtenni, hogy a nyers szövegből a modell szerinti reprezentációval bíró korpuszt kapjuk. A végső termékként előálló szótár (4.2. rész) is a Magyar Nemzeti Szövegtár teljes anyagára épül majd, ezért most is ennek a korpusznak a példáján mutatjuk be a feldolgozási lépéseket. (A 5. részben egyéb korpuszokkal is foglalkozunk majd.)

Kiinduló korpuszunk tehát a Magyar Nemzeti Szövegtár (`http://mnsz.nytud.hu`) (Váradi, 2002). Az MNSZ az ezredforduló magyar írott köznyelvének általános célú reprezentatív korpusza. 187,6 millió szónyi magyar szöveget tartamaz öt különböző stílusrétegből és öt különböző határontúli regionális nyelvváltozatból. Az MNSZ automatikus, egyértelműsített morfológiai annotációt tartalmaz. A morfológiai elemzés a Morphologic *Humor* elemzőjével (Prószéky és Tihanyi, 1996) készült. A morfoszintaktikai elemző és egyértelműsítő rendszer összesített pontossága 97,5%-os, azaz az összes szóalak 97,5%-a van helyesen elemezve (Oravecz és Dienes, 2002). Ennél jobb eredményt csak a kézi elemzés biztosíthatna, ami ekkora méretű anyag esetén megvalósíthatatlan. Az automatikus morfológiai elemzés és egyértelműsítés eredményeképpen az MNSZ-ben minden egyes szóhoz hozzá van rendelve a szótő, a szófaj és a morfológiai elemzés információ.

Mivel morfológiailag elemzett és egyértelműsített korpuszból indulunk ki, a következő két lépés szükséges: tagmondatra bontás (Sass, 2006b); és részleges szintaktikai elemzés (Sass, 2005).

2.2.1. Tagmondatra bontás

Az első feldolgozó lépés a tagmondatra bontás. E lépés célja az, hogy olyan egységeket kapjunk, melyek egy igét és annak bővítményeit tartalmazzák, azaz előállítsuk a modell által megkövetelt alapegységet. A szöveg tagmondatai általában egy bővítménykeretet tartalmaznak, megfelelnek a nagy valószínűséggel egy bővítménykeretet tartalmazó alapegységnek. A tagmondat kifejezést ebben az értelemben használom: a mondat egy bővítménykeretet tartalmazó része, így lényeges követelmény lesz annak garantálása, hogy a tagmondat egy igét tartalmazzon. Sok helyen találkozhatunk

DOI:10.15774/PPKE.ITK.2011.004

a mondatok bizonyos szempontból könnyebben elemezhető, kisebb részekre darabolásával (Kim és Hong, 2006), itt is erről van szó.

Azáltal, hogy az alárendelést tartalmazó mondatokat tagmondatra bontjuk, az alárendelt tagmondatban szereplő igei szerkezetekhez is hozzáférünk. Azaz attól függetlenül számba vehetjük a szerkezeteket, hogy szintaktikai szerkezetnek éppen mennyire elrejtett szintjén fordulnak elő. Ha egy szerkezet gyakoriságát akarjuk megállapítani, akkor nyilván minden előfordulása számít. A tagmondatra bontás tehát azt is biztosítja, hogy a gyakoriságok számításakor minden igei szerkezet ugyanannyit ér, ugyanolyan jogon számít.

Korábbi megoldások

A tagmondatra bontó rendszer kialakításakor az alábbi kutatásokból indultam ki. A (Váradi és Gábor, 2004) cikk ismertet egy az INTEX/NooJ nyelvfeldolgozó rendszerben implementált eljárást. Ezenkívül két kézirat állt rendelkezésemre: az imént említett eljárás részleit tartalmazó kézirat (Gábor, 2005), illetve egy másik megközelítés (Varasdi, 2005).

A (Gábor, 2005) kéziratban ismertetett tagmondathatár-azonosító rendszer tizenegy szabályból áll. Az egyik szabály például tagmondathatárt helyez el vessző után, amennyiben a vesszőt (esetleges kötőszó vagy határozószó közbeszúrásával) vonatkozó névmás követi. Adott szabály illeszkedése esetén a szabály által meghatározott helyre kerül a szövegbe a tagmondathatár. Az eljáráshoz tartozik még egy a szabályalkalmazások után futó program, mely lehetséges tagmondathatárként megjelöli az összes kötőszót, mely két olyan finit ige között helyezkedik el, melyek között még nincs tagmondathatár. A szabályrendszerben részletesen benne foglaltatik, hogy az egyes kötőszók hányadik pozícióban szoktak állni a tagmondathatárhoz képest, és milyen típusú elemek előzhetik meg őket.

A (Varasdi, 2005) kéziratban leírt, de nem implementált eljárás igazi célja, hogy megállapítsa a szöveg kötőszavairól, hogy szerkezeteket koordinálnak vagy esetleg tagmondatokat kötnek össze, így mintegy melléktermékként kapjuk meg a tagmondatokat.

Több helyen (Gábor et al., 2003; Varasdi, 2005) megfogalmazott fontos elv, hogy a finit ige vonzatai az igét tartalmazó tagmondaton belül vannak. A magyar névszói állítmány kezelése minkét módszernek nehézséget jelent, előfordulhat, hogy hibásan bekerülnek az ige bővítményei közé a szomszédos névszói prédikátum bővítményei is. Probléma lehet még, ha a magyar mondatból elmarad a kötőszó, ilyenkor a kötőszóra építő szabályok természetesen nem működnek.

A tagmondatra bontó eljárás

Az általam kialakított módszer főként a fent ismertetett első eljárásra épít, ez alapján egy szabályalapú rendszert alakítottam ki a morfológiailag elemzett szöveg tagmondatokra bontására. A szabályok a szövegszavak és írásjelek sorozata fölött megfogalmazott reguláris kifejezések. Azon alapulnak, hogy milyen a szövegben a központozás

DOI:10.15774/PPKE.ITK.2011.004

2. Igei szerkezetek modellje

2. táblázat. A kialakított tagmondatra bontó eljárás szabályai. A szabályokat reguláris kifejezésre emlékeztető szintaxissal írom le, adott szabály illeszkedése esetén a '@' jel helyére kerül be egy tagmondathatár.

[: | ;]	@	
[, | -]	@	[kötőszó | határozószó]? [vonatkozó névmás]
[-]	@	[kötőszó | határozószó]? [vonatkozó névmás] [bármi]+ [-] [,]? @
[, | -]	@	[bármi]{0,3} ['*pedig*' | '*akár*' | '*azonban*' | '*viszont*' | '*ellenben*' | '*mihelyt*' | '*tehát*' | '*ugyanis*']
[, | -]	@	[határozószó]? ['*nehogy*' | '*mintha*']
[,]	@	[kötőszó, kivéve: '*de*' | '*illetve*' | '*illetőleg*' | '*mintegy*']
[, | -]	@	[múlt idejű, egyes szám harmadik személyű ige]
[,]	@	['*az*' szótőként]? [határozói igenév] [,] ['*hogy*']

és a kötőszavak elhelyezkedése. A szabályok a (Gábor, 2005) kéziratból származnak, korpuszvizsgálatok alapján a morfológiai elemzés hibái (pl.: a '*meg*' és a '*ki*' elváló igekötő gyakori hibás elemzése) vagy más okok miatt rosszul teljesítő néhány szabályt elhagytam, illetve néhány újat vettem hozzá (Sass, 2006b) (2. táblázat).

A tagmondatra bontó algoritmus végighalad a korpusz mondatain. Az adott mondat minden egyes szavára sorra illeszti a szabályokat úgy, hogy az adott szó utáni ponton próbál tagmondathatárt keresni. Ha az egyik szabály tagmondathatárt jelez, akkor létrehozza a tagmondathatárt, majd továbblép a következő szóra.

Ez az algoritmus a következő eljárással egészül ki. Tudjuk, hogy az ige vonzatai vele egy tagmondatban vannak (Gábor et al., 2003; Varasdi, 2005). Ezt kiegészíthetjük azzal, hogy csak a tagmondat igéjének a vonzatai vannak a tagmondatban. Ebből következik, hogy az ige-koordinációt nem engedjük meg, két finit ige közé akkor is megpróbálunk tagmondathatárt elhelyezni, ha szabállyal ez nem sikerült. Megfigyelhető, hogy nemcsak kötőszó, hanem legalább ugyanolyan gyakran közbeeső központozás (vessző, pontosvessző, kötőjel) is lehet tagmondathatár. Tehát két finit ige között megjelöljük ezen írásjelek *utáni* és a kötőszavak *előtti* összes pozíciót, mint lehetséges tagmondathatárt (Váradi és Gábor, 2004). Ha egyetlen ilyen közbeeső pozíció van, akkor az lesz a tagmondathatár (Varasdi, 2005). Ha pedig több ilyen megjelölt hely van, akkor ezek közül – heurisztikus döntéssel – a leginkább jobbra esőt választjuk, csökkentve az esélyét annak, hogy hibásan, felsorolás közepére helyezzünk el tagmondathatárt.

Kiértékelés

A kiértékeléshez az MNSZ részét képező *Magyar Nemzet* napilap anyagából választottam ki véletlenszerűen 200 mondatot (Sass, 2006b). Ezen a kisméretű tesztkorpuszon a következő nagyon egyszerű útmutató szerint végeztem a tagmondatok manuális

DOI:10.15774/PPKE.ITK.2011.004

bejelölését: (1) jelöljük be a szövegben a tagmondatokat; (2) minden finit ige külön tagmondatba kerüljön; (3) a tagmondatvégi központozás minden esetben a megelőző tagmondathoz tartozzon. A kiértékelés eredménye a 3. táblázatban látható.

3. táblázat. A szabályalapú tagmondatra bontó eljárás a 171 bejelölt tagmondathatárból a program 148-at talált meg (23-at hagyott ki), emellett 29 helytelen tagmondathatárt jelölt meg.

pontosság	= **83,6**%
fedés	= **86,5**%
F-mérték	= **85,0**%

Ezen mérőszámokat befolyásoló tényező lehet az, hogy a szöveg egy viszonylag bonyolult jogi nyelvezeten írt részletet, egy rendeletszöveget tartalmazott, valamint, hogy az eredeti korpuszban sokszor helytelen volt a mondatok határainak megállapítása. Egyszerűbb szerkezetű szöveg esetén, valamint jobb mondatrabontás alkalmazásával minden bizonnyal még növelhetők ezek az értékek. Amint várható volt, a hibák főleg olyan pontokon jelentkeznek, ahol szinte semmi konkrét jel nem utal arra, hogy ott egy tagmondat kezdődik, nincs kötőszó (sőt esetleg központozás sem), illetve névszói állítmány van valamelyik tagmondatban (például: '*A kérdés második felére azt felelném, minden lehetséges s minden az erőviszonyoktól függ.*') Ez a teljesítmény a további feldolgozáshoz elegendő, sok esetben csak olyan hibáról van szó, melyek a bővítményekre nincsenek kihatással.

2.2.2. Szintaktikai elemzés

A tagmondatra bontást követő részleges szintaktikai elemzés során nem törekszünk a tagmondatok teljes szintaktikai fájának felépítésére. Ehelyett az elemzés célja: a központi, „kerethordozó" ige és a mellette álló főnévi csoport bővítmények azonosítása. A modellnek megfelelően csak az igét és a névszói csoportokat dolgozzuk fel, a jelen lévő határozószókat például figyelmen kívül hagyjuk. Ezek alapján a reprezentáció már kialakítható.

Az elemző algoritmus és a felhasznált nyelvtan

A tagmondatra bontáshoz hasonlóan itt is szabályalapú megközelítéssel dolgozunk. A szabályok szintén a szövegszavak és írásjelek sorozata fölött megfogalmazott reguláris kifejezések, a kidolgozott morfológiai reprezentáció részletekbe menő lekérdezéseket tesz lehetővé, az elemzési lépésekben részletekbe menően hivatkozhatunk a magyar morfológia különféle jellemzőire. Ezek a szabályok – a tagmondatra bontó szabályoktól eltérően – többszintű reguláris nyelvtant (*cascaded regular grammar*) (Abney, 1996) alkotnak: egymásra épülnek, azaz a felismert csoportokból további szabályokkal, rekurzívan újabb, nagyobb kiterjedésű csoportok képezhetők (Sass, 2005).

DOI:10.15774/PPKE.ITK.2011.004

2. Igei szerkezetek modellje

Az elemző algoritmus végighalad a korpusz tagmondatain, és egységek sorozatára sorrendben illeszti a szabályokat. Illeszkedés esetén a szabály által lefedett egységekből a szabály bal oldalának megfelelő címkével ellátott új egységet képez. Az egységek kezdetben a szavak, később a szabályok alkalmazása révén létrejött több szóból álló csoportok. A felhasznált szabályrendszert, mely képes a különféle névszói csoportok és az ige felismerésére, a 7. ábrán mutatom be. A névszói csoportokat érintő szabályok megalkotása során építettem a (Váradi, 2003) cikkben ismertetett szabályokra. Nemrégen készült magyar nyelvre egy jó minőségű főnévi csoport felismerő rendszer (Recski, 2010), ezt természetesen a jövőben a reprezentáció előállítása során alkalmazni lehet.

A modell alapján a névszói csoportok két számunkra fontos tulajdonsága az esetrag és a csoport feje mint tartalmi elem: ezeket azonosítjuk és ezek fognak a reprezentációba kerülni. Amint a 28. oldalon említettük, a névutókat az esetragokkal azonos módon kezeljük, a bővítmények tehát esetragos vagy névutós névszói csoportok. Egy szabály alkalmazása során alapesetben a létrejövő új egység a benne szereplő utolsó szó tulajdonságait örökli, ennek köszönhető, hogy a névszó esetragja a névszói csoport eset jegyébe kerül a többszörösen összetett névszói csoportok esetén is. A névutók (főként a személyragos névutók) természetesen ettől eltérő speciális kezelést igényelnek. A rendszer tartalmazza azt az egyszerűsítést, hogy a bővítménykeretek minden bővítményi helyén csak egy darab névszói szerkezetet enged meg, ha egy mondatban több azonos esetragú névszói csoport szerepel, akkor azok közül csak a legutolsót vesszük tekintetbe. Ha egy tárgyas ragozású igével bíró mondatban nincs explicit tárgyesetű csoport, akkor a speciális `NULL` tartalmi elemmel veszünk fel egyet, elfogadjuk, hogy az igei személyrag egy tárgyi bővítményi hely meglétére utal az adott bővítménykeretben.

Az ige meghatározása

Amint azt a 7. ábrán látjuk, az elemzés megjelöli a tagmondat központi igéjének meghatározásához szükséges információkat is: a finit igét, az esetleges elváló igekötőt és az esetleges jelen lévő főnévi igenevet. Az igető azonosítása során az ige morfológiai elemzéséből kinyerhető igető elé kapcsoljuk az igekötőt. Elhagyjuk a '*-hat/-het*' képzőt, mivel az nem befolyásolja az ige vonzatkeretét. Ha a tagmondatban főnévi igenevet találunk, akkor a főnévi igenév tövét tekintjük főigének. Persze sok esetben nem igaz, hogy a tagmondat főnévi igenevéhez tartozik a tagmondatban lévő összes bővítmény. Az ilyen hibák javítására számos szabály tesztelése után egy megbízhatóan működő szabályt tartottunk meg: ragos főnévi igenév esetén ha nincs a tagmondatban alanyesetű névszói csoport, akkor a '*-nAk*'-ragos névszói csoportot tekintjük alanynak. Ez alapján a '*Péternek meg kellett csinálnia a feladatot.*' mondat elemzése után '*megcsinál*' lesz az ige, '*Péter*' lesz az alany és a '*feladat*' a tárgy.

DOI:10.15774/PPKE.ITK.2011.004


```
1.
   X          <- {position='0';'Det|Adv|Con|Pro|V|Num'}
   NE         <- {capit='yes';unknown='yes'}{2,}
   NE         <- {capit='yes';unknown='no'}{2,}
   NE         <- {capit='yes';unknown='yes'} {capit='yes';unknown='no'}
   NE         <- {form='dr.'} NE{}
   X          @delete
2.
   NP:d       <- [{'Det'} {form='egy'}]?  [{'A'} {'Num'}]* NE{}
   NP:d       <- {'Det'} [{'A'} {'Num'}]* {'N'}
   NP:n       <- [{'A'} {'Num'}]* {'N'}
3.
   NUragos    <- {lemma~'_';'NU'}
4.
   NP:pro     <- {'Pro'}
   NP:d:adj   <- {'Det'} {'A'}
   NP:d:num   <- {'Det'} {'Num'}
   NP:n:adj   <- [{form='egy'}]? {'A'}
   NP:n:num   <- [{form='egy'}]? {'Num'}
5.
   NP:ps:1    <- NP{case='NOM';pspers!='3'} NP:n{ps='sing3'}
   NP:ps:2    <- NP{case='DAT';pspers!='3'} NP:d{ps='sing3'}
   NP:ps:3    <- NP{case='DAT'} NP:d{ps='sing3'}
   NP:ps:4    <- NP{case='DAT'} NP:n{ps='sing3'}
6.
   NP:nu      <- NP{} {'NU'}
7.
   MNI        <- [{'MIF'} {'MIB'}]
   A          <- NP{} MNI{}
   NP:mif     <- [{'Det'} {form='egy'}]? MNI{} NP{}
   NP:mif     <- A{} NP{}
   NP:mif     <- A{} NP{}
8.
   I_         <- {'V'}
   PRE        <- {'Pre'}
   INF        <- {'INF'}
9.
   A_         <- NP{case='NOM'}
   T_         <- NP{case='ACC'}
```

7. ábra. A részleges szintaktikai elemzés nyelvtana belső formátumban. Egy kapcsoszárójelen belüli feltételek egy egységre vonatkoznak. Kezdetben egy szó egy egység, de a szabályok alkalmazásával több szóból álló egységek is képződnek. A szögletes zárójel vagylagosságot jelöl. A szabályokat sorban alkalmazzuk, a szabály jobb oldalának megfelelő egység(ek) a szabály bal oldalán látható címkét kapják meg. Az 1. szabálycsoport nagyon egyszerű tulajdonnévfelismerőt valósít meg: lényegében nagybetűs szavak sorozatait keresi meg, kiegészítve azzal, hogy bizonyos szófajú mondatkezdő (`position='0'`) nagybetűs szavakat nem enged meg tulajdonnév részeként. A 2. szabálycsoport a legegyszerűbb határozott (`NP:d`) és határozatlan (`NP:n`) főnévi csoportokat azonosítja. A 3. szabály elkülöníti a személyragos névutókat (a lemmában található aláhúzás karakter alapján). A 4. szabálycsoport felépíti a névmási, melléknévi és számnévi csoportokat. Az 5. csoportban következnek a birtokos szerkezeteket kezelő szabályok. A 6. szabály a névutós csoportokat ismeri fel. A 7. szabálycsoport a melléknévi igeneves szerkezeteket kezeli. A 8. szabálycsoport számbaveszi a tagmondat igéjével kapcsolatos elemeket: a finit igét, az esetleges elváló igekötőt, illetve főnévi igenevet; végül a 9. szabálycsoport a legfelsőbb szintű névszói csoportok közül esetrag alapján külön megjelöli az alanyt és a tárgyat.

DOI:10.15774/PPKE.ITK.2011.004

A kapott reprezentáció

Az elemzés végén az esetrag/névutó mint viszonyjelölő révén a mondat igéjéhez rendeljük a fej által reprezentált névszói csoportokat, kialakítva a modell által megkívánt függőségi struktúrát.

Az ismertett részleges szintaktikai elemzés tehát alkalmas arra, hogy előállítsa egy tagmondatnak a modell által megkívánt reprezentációját, mely szerint a tagmondat igéből és névszói csoport bővítményekből áll, a bővítmények reprezentációja pedig az esetrag, illetve az esetraggal megjelenő tartalmi elem, azaz konkrét szó. Az említett '*Péternek meg kellett csinálnia a feladatot.*' tagmondat elemzése végén tehát előáll a következő kívánt reprezentáció: '`ige=`megcsinál `-0=`Péter `-t=`feladat'

2.2.3. Összefoglalás

Kutatásom további részéhez szükséges előfeltétel volt egy nagy méretű korpusznak a kidolgozott modell szerinti reprezentációja. Ennek előállításához a fent ismertetett közelítő módszereket használtam. A szabályalapú tagmondatra bontás és részleges szintaktikai elemzés (igeazonosítás és névszói csoportok felismerése) révén viszonylag kis erőfeszítéssel egy egyszerű felépítésű függőségileg elemzett korpuszhoz jutottunk, mely a modellnek megfelelően csak a mondat legfelső szintjén megjelenő dependenciákat ábrázolja.

Nem állítom, hogy e nyelvi elemző lépések megvalósítása kiemelkedő minőségű, kiértékelésük is korlátozott mértékű. Ezek részletes kidolgozása és tökéletesítése önmagukban önálló dolgozatok témáját adhatják. Elegendő leszögezni, hogy a Magyar Nemzeti Szövegtár kialakított reprezentációja megfelelő alapot biztosít kutatásom további lépéseihez, illetve eredményeim bemutatásához. Megjegyzendő, hogy a kapott korpusz *kifejezetten* nagy méretű (147 millió szavas), ami lehetővé teszi a ritka jelenségek jellemzését is. A valódi, teljes függőségileg elemzett korpuszok általában ennél két (vagy akár három) nagyságrenddel kisebbek.

A reprezentáció kialakításáról szóló **2. tézis** a 110. oldalon olvasható.

DOI:10.15774/PPKE.ITK.2011.004

3. fejezet

Igei szerkezetek kinyerése

A dolgozat leghosszabb fejezetében folytatom az új tudományos eredmények ismertetését: a modell (2. fejezet) ismeretében a modell szerint reprezentált igei szerkezetek korpuszból való kinyeréséről szólok. A 3.1. részben azt indoklom, hogy miért megfelelő az idiomatikus bővítmények helyett a lényeges bővítményekkel foglalkozni. Bemutatok egy a reprezentációhoz illeszkedő korpuszlekérdező rendszert (3.2. rész), végül pedig a jellegzetes igei szerkezetek kinyerésére szolgáló algoritmust ismertetem (3.3. rész). Ezen algoritmus alkalmazásáról (4.2. rész) és kiterjesztéseiről (5. fejezet) lesz aztán szó a dolgozat további részeiben.

3.1. Idiomatikusság helyett lényegesség

Ebben a részben bemutatok egy korábbi kísérletet, melynek célja az idiomatikus igei keretek kinyerése volt. A kísérlet tapasztalatai és egyéb megfontolások alapján indoklom, hogy a továbbiakban nem a szorosan vett idiomatikus, hanem az ennél nagyobb halmazt jelentő lényeges igei szerkezetekkel foglalkozom. Bemutatok egy hasznos kollokációs mértéket, és ismertetem azt a módot, ahogyan ezt a két szó kollokacionalitásának vizsgálatára kifejlesztett mértéket az igei szerkezetekre alkalmaztam. Ez a mérték alkalmas lesz a lényeges bővítmények, és ezáltal a lényeges igei szerkezetek megragadására.

3.1.1. Kísérlet idiomatikus igei szerkezetek kinyerésére

Nyelvtechnológiai alkalmazások – például a gépi fordítás – szemszögéből elsősorban azokat az igei szerkezeteket érdemes összegyűjteni, és a lexikonban külön nyilvántartani, melyeknek a jelentése nem kompozicionális, idiomatikus, és ezáltal a fordításuk speciális (nem triviális), azaz a fordítás nem vezethető le a szavak fordításaiból (Bojar és Hajič, 2005). A most ismertetendő kísérlet (Sass, 2006a) célja az volt, hogy a létrehozott korpuszreprezentáció alapján kinyerjem a kötött névszót – azaz LKB-t – is tartalmazó idiomatikus szerkezeteket.

DOI:10.15774/PPKE.ITK.2011.004

3. Igei szerkezetek kinyerése

Ebben a korai vizsgálatban az MNSZ 3–10 szavas, írásjelet nem tartamazó mondatainak 10 millió szavas korpuszát használtuk. Itt tagmondatra bontást nem kellett végezni, ezek a mondatok jó eséllyel egy igei keretet tartalmaznak.

Az idiomatikus jelentéssel bíró, LKB-t tartalmazó igei keretek kinyerésére szolgáló módszerünk két lépésből állt. Az első lépésben összegyűjtöttük a modellnek megfelelő korpusz-reprezentációból az összes mondatvázat. Ezt a listát LSzB-ket is tartalmazó keretekket egészítettük ki (hasonlóan ahhoz, amit majd az 57. oldalon a valódi algoritmusban alkalmazunk): a mondatvázak minden egyes LKB-jéből három változatot készítettünk: egyrészt megtartottuk az LKB-t, másrészt töröltük a tartalmi elemet, azaz LSzB-vé alakítottuk, harmadrészt teljesen elhagytuk a mondatvázból. Ezt minden lehetséges variációban megcsináltuk, így egy n bővítményt tartalmazó mondatvázból 3^n származtatott keret lett. (A keretek kezelhetetlenül nagy száma miatt, az alanyt – mely a leggyakrabban tartalmaz gyakori, de nem idiomatikus jelentésű szót – elhagytuk a mondatvázakból, elvesztve ezáltal a 'derül -rA FÉNY'-típusú szerkezeteket.) Figyelmen kívül hagyva, hogy a fenti módon minden mondatvázból számos származtatott mondatváz keletkezik, és emiatt az eredeti gyakorisági viszonyok sérülnek, az így kapott összes igei keretből gyakorisági listát készítettünk, ez lett a kiinduló lista a következő lépéshez.

A második lépésben az idiomatikus keretek kinyerése céljából egy konkrét idiomaticitási mértéket alkalmaztunk (Tapanainen et al., 1998) javaslatának megfelelően. Eszerint a mérték szerint az a keret az idiomatikusabb, melynek bővítményei az adott formában kevés (szélső esetben egyetlen) igével fordulnak elő (a '*fittyet vmire*' bővítménykeret például kizárólag a '*hány*' igével fordul elő). Tapanainen et al. (1998) az ige-tárgy relációval foglalkoznak, erre fogalmazzák meg az *elosztott gyakoriság (distributed frequency, DF)* mértéket, mely a következők szerint működik: ha egy tárgy csak kevés igével fordul elő együtt, akkor a DF értéke magasabb lesz. Pontosabban: ha egy adott tárgy (**o**) n különböző igével ($V_{1..n}$) jelenik meg egy gyakorisági küszöbnél ($C = 5$) többször (F_k jelöli a $(V_k, \mathbf{o})$ kollokációk gyakoriságát), akkor a DF kiszámítására szolgáló formula a következő:

$$\mathrm{DF}(\mathbf{o}) = \sum_{k=1}^{n} \frac{F_k}{n}$$

Esetünkben ezt a mértéket nem két szóra (az igére és a tárgyra), hanem az igére és a bővítménykeretre (most nem beleértve az igét!) kell alkalmaznunk. Egyszerűen vettük a bővítménykeretet egy sztringként, és így alkalmaztuk a mértéket.

A DF mértéket megszoroztam az igének az adott bővítménykereten belül mért relatív gyakoriságával, így kaptam a végső idiomaticitási mértéket: a *DF-pontszám*ot, mely nem csak a keretet, hanem az igét is számításba veszi, így különböző értéket ad a kereteknek attól függően, hogy mely igével kollokálnak.

$$\text{DF-pontszám}(V_k, \mathbf{o}) = \mathrm{DF}(\mathbf{o}) \cdot \frac{F_k}{\sum_{i=1}^{n} F_i}$$

Ha ez a pontszám egy küszöbérték felett van, a keret bekerül az idiomatikus keretek

DOI:10.15774/PPKE.ITK.2011.004

listájába. Az idiomatikusságban megfigyelhető gradualitás (McCarthy et al., 2003) miatt nem mondhatjuk, hogy bizonyos keretek idiomatikusak, bizonyosak pedig nem, csak annyit mondhatunk, hogy a lista elején lévő keretek idiomatikusabbak, mint a kevesebb pontszámmal lejjebb következők.

A módszert először kézi annotálás segítségével értékeltük ki. Azon kiértékelési feltétel mellett, hogy „idiomatikus az a szerkezet, melynek az angol fordítása speciális" a pontossági értékek 12 és 75, a fedés értékek pedig 46 és 81 százalék között mozogtak. (Briscoe és Carroll, 1997) munkájához hasonlóan összevetettük a kinyert kereteket egy tekintélyes igei keret adatbázissal. Mivel magyar nyelvre nincs elektronikus igei keret adatbázis, a Magyar Értelmező Kéziszótárhoz (Pusztai, 2003) fordultunk: 17 kiválasztott keretet vetettünk össze a szótár megfelelő igei címszavainak anyagával. A szótárban 15 keret van, ebből a módszerünk mindössze 5-öt talált meg, azaz a szótárhoz viszonyított fedés csak 33%. Viszont az is kiderült, hogy a kiválasztott 17 keretből 14 helyes idiomatikus keret, azaz 9 olyan gyakori keretet találtunk, mely a szótárban nem szerepelt. A szótárakhoz viszonyított kiértékelés ismert problémájával találkoztunk: bizonyos ritka elemeket nem találunk meg, mert nem szerepelnek a korpuszunkban, viszont találunk további helyes elemeket, mert a szótár nem teljes (Manning, 1993; McCarthy et al., 2003).

Elmondható, hogy bár voltak biztató részeredmények – a 'mond PÉLDA-t' keretet helyesen nem-idiomatikusnak, a 'mutat -nAk PÉLDA-t' pedig helyesen idiomatikusnak ítélte a módszer – a kiértékelés azt mutatja, hogy a módszer az idiomatikus szerkezetek kinyerésére nem elég megbízható.

3.1.2. A lényegesség és a gyakoriság szerepe

Amint láttuk, nem egyszerű feladat az idiomatikus igei szerkezetek kinyerése, az idiomatikusság és a kompozicionalitás elkülönítése. Ez a szembenállás nemcsak a szerkezetek szintjén, hanem az egyes bővítmények szintjén is megjelenik. A vonzatok tekinthetők idiomatikus bővítménynek: ekkor a viszonyjelölő jelentése nem megjósolható (pl.: *'hisz vmiben'*). A szabad határozók pedig a kompozicionális bővítmények: ekkor a viszonyjelölő jelentése megjósolható (pl.: *'ül a fotelben'*). A bővítmények két alapvető osztályának, a vonzatoknak és szabad határozóknak az elkülönítése a magyarban nyelvészetileg sem megoldott kérdés (Komlósy, 1992). A valódi vonzatkeretek megragadására sincs megbízható automatikus eszközünk.

Létezik azonban az igei kereteknek egy, a valódi vonzatkereteknél bővebb halmaza mely több szempontból – lexikográfiailag, vagy a gépi fordítás szemszögéből is – lényegesnek mondható. Ezek között a szerkezetek között már nemcsak idiomatikus, hanem kompozicionális szerkezetek is vannak; a szereplő bővítmények között pedig nemcsak vonzatok, hanem szabad határozók is. Ide tartozik például: *hajat vág*, *fésüli a haját*, *választ ad valamire*, *véleményének ad hangot*, *nem tud semmit vmiről*, *csökken a száma*, *problémát okoz*, *örömmel fogad vmit*.

E dolgozatban a vizsgálódást tehát nem korlátozzuk az idiomatikus szerkezetekre, illetve a valódi vonzatkeretekre, helyettük az igék bővítményszerkezetével, a lényeges

DOI:10.15774/PPKE.ITK.2011.004

bővítményekkel és a lényeges igei keretekkel foglalkozunk. Azaz a továbbiakban nem mérlegelem, hogy mi vonzat és mi szabad határozó, csak azzal foglalkozom, hogy melyik bővítmény lényeges. Érdemes ezzel a tágabb körrel foglalkozni, mert ezek azok a szerkezetek, melyeket jellegzetességük, intézményesültségük és gyakoriságuk miatt érdemes belefoglalni egy szótárba, vagy egy nyelvtechnológiai rendszer nyelvi, lexikai adatbázisába.

Lexikográfiai szempontból a lényegességnek valóban fontos eleme a gyakoriság: egy szótárnak a gyakori nyelvi egységeket kell tartalmaznia. A Magyar Értelmező Kéziszótárban (Pusztai, 2003) például nem szerepel a '*nemet mond vmire*', szerepel viszont a '*rosszat mond vkire*'. Mindkét szerkezet kompozicionálisnak vagy csak kis mértékben idiomatikusnak mondható, emelett mindkettő nagyon jellegzetes szerkezet. A első szerkezet azonban jóval gyakoribb (a Magyar Nemzeti Szövegtárban $7\times$ gyakrabban fordul elő) mint a másik, ezért a fent idézett gyakorisági elv azt kívánná, hogy a gyakoribb szerkezetet tüntessük fel. A gyakorisági elv alapján változtatni lehet a szótárban feltüntetett jelentések sorrendjén is. A szokásos gyakorlattal szemben, mely az „alapjelentést" dolgozza ki először, érdemes lehet a gyakoribb jelentéseket előrevenni. Így nem fordulna elő az, hogy a '*kezébe/nyakába vesz vmit*' előrébb szerepel, mint a nagyságrendekkel gyakoribb '*részt vesz vmiben*', ahogy ezt most az ÉKSz-ben látjuk.

Az, hogy gyakori kompozicionális szerkezeteket kellően fontosnak tartunk ahhoz, hogy egy szótárba belekerüljenek nem új gondolat. Sinclair (1998) a szótárban szereplő szókapcsolatok esetében nem tartja feltétlenül szükséges kritériumnak, hogy a szókapcsolatnak önálló, nem-kompozicionális jelentése legyen. A több szóból álló, rendszeresen együttesen előforduló szókapcsolatokat pusztán e rendszeres, gyakori előfordulás miatt címszóként rangjára emeli. Célszerűnek tartja, hogy a szótárak minél inkább maguknak a kollokációknak az értelmezésére törekedjenek, mivel a szavak sohasem önmagukban, hanem mindig valamilyen szövegkörnyezetben jelennek meg. Goldberg (2006, 5. oldal) pedig a konstrukciók (vö: 24. oldal) között is nyilvántart kompozionális kifejezéseket. Ha egy szerkezet kellően gyakori ahhoz, hogy egy egységként rögzüljön, akkor konstrukciónak számít, legyen akár teljesen kompozicionális.

Az sem okoz gondot, ha bizonyos lényeges kompozicionális szerkezeteket egy számítógépes alkalmazásban a lexikonban tárolunk. Nyilván lehetetlen az összes kompozicionális szerkezetet az adatbázisban felsorolni, de az olyan szerkezetek esetében például, mikor bizonyos bővítményi helyeken csak egyetlen szó fordulhat elő, ez a megoldás, hogy a lexikonban kezeljük, nem igényel több erőforrást (McCarthy et al., 2003).

3.1.3. Igei szerkezetek mint kollokációk

A számítógépes nyelvészetben bevett fogalom az *n-gram*, amely egyszerűen n darab egymást követő szót jelent. Ezt a fogalmat terjeszthetjük ki úgy – ezt nevezik *concgram*-nek –, hogy egyrészt a szavak között egyéb közbeékelődő szót is megengedünk, másrészt a szavak sorrendjét sem kötjük meg (Cheng et al., 2006). Egy magyar bővítménykeret elemei a mondatban tetszőleges sorrendben fordulhatnak elő, és mellettük további bővítmények is megjelenhetnek, így – még egy kiterjesztést téve: a szavak

DOI:10.15774/PPKE.ITK.2011.004

helyett frázisokat tekintve alapegységnek – a magyar egyszerű mondatot egy olyan concgram-nek foghatjuk fel, melyben az egyes egységeket frázisok képviselik.

A kollokáció szokásos két egymás melletti szóra (egy *2-gramra*) (vö: 18. oldal) vonatkozó definícióját kiterjeszthetjük a most bevezetett concgram struktúrára. Másképp fogalmazva arról van szó, hogy a kollokáció kifejezést használhatjuk abban a tág értelemben, hogy „együttes előfordulás". Az egy tagmondaton belüli tetszőleges sorrendű, akár megszakított együttes előfordulásról van itt szó, a bővítmények sorrendje illetve egymás mellettisége nem számít, csakis az, hogy az igével egy tagmondatban vannak. Ezáltal a bővítménykeretek felfoghatók kollokációknak, és a lényeges kereteket mint lényeges kollokációkat vizsgálhatjuk. Kollokáción tehát tág értelemben az ige, és különféle bővítményeinek összessége együttes előfordulását értjük, bármilyen formában illetve szórendben és közbeszúrt elemekkel jelenjenek is meg.

3.1.4. A salience kollokációs mérték

A fentiek alapján egy alkalmas kollokációs mérték megfelelő alkalmazásával kinyerhetők a lényeges bővítménykeretek. A kollokációk keresésére használt klasszikus mérték, a *kozlcsönös információ@kölcsönös információ* (*mutual information, MI*) a következő képlettel adható meg:

$$\mathrm{MI}(x,y) = log_2\left(N\frac{f(x,y)}{f(x)\cdot f(y)}\right)$$

ahol N a korpusz mérete, f az előfordulási szám, x és y pedig a két elem, melyeknek a kollokacionalitását vizsgáljuk. E mérték akkor ad magas értéket, ha a két elem a véletlenszerű együttes előfordulásnál gyakrabban fordul elő együtt. Hátrányos tulajdonsága, hogy túlzottan kiemeli a ritka elemeket (Sass, 2006b). Gondoljuk meg:

1. Ha y hapax és éppen x-szel együtt fordul elő, akkor $f(y) = 1$, $f(x,y) = 1$, azaz
$$\mathrm{MI}(x,y) = log_2\left(N\frac{1}{f(x)\cdot 1}\right) = log_2\left(N\frac{1}{f(x)}\right)$$
2. Ha y előfordulási száma 500, és ebből 250-szer x-szel együtt fordul elő, akkor $f(y) = 500$, $f(x,y) = 250$, azaz
$$\mathrm{MI}(x,y) = log_2\left(N\frac{250}{f(x)\cdot 500}\right) = log_2\left(N\frac{1}{2f(x)}\right)$$

Az első esetben nagyobb értéket kapunk, mert ez a mérték annak tulajdonít nagy jelentőséget, hogy az *összes* y-re igaz, hogy x-szel együtt fordult elő, hiába igaz az is, hogy y-nak ez az összes előfordulási száma mindössze 1.

E tulajdonság ellensúlyozására elfogadott megoldás az, hogy az MI értéket korrigáljuk a vizsgált elem (y) előfordulási számának a logaritmusával (hasonlóan a DF módosításához, a 42. oldalon), így kapjuk meg a szakasz címében említett *salience* mértéket (Kilgarriff és Tugwell, 2001).

DOI:10.15774/PPKE.ITK.2011.004

$$\mathrm{S}(x, y) = (log_2 f(y)) \cdot \mathrm{MI}(x, y)$$

A salience szerint rendezett listában valóban a tipikus, lényeges kollokációk kerülnek a lista elejére, az egyszerű előfordulási számhoz képest a salience szerinti ranglistán hátrébb sorolódnak a nagyon gyakori (mindennel előforduló) szavak, és kiküszöböli az MI mérték említett hibáját is. Megállapíthatjuk, hogy a lényeges kollokációk korpuszból való kinyerésére a salience mérték alkalmas.

A lényeges kollokációkat tehát ezzel a mértékkel hatékonyan meg tudjuk ragadni, az a kérdés marad, hogy hogyan tudjuk alkalmazni a 3.1.3. részben bemutatott struktúrára.

3.1.5. A salience alkalmazása az igei szerkezetekre

A két elem együttes előfordulásának vizsgálatára kidolgozott salience mértéket a következő módon alkalmazzuk az igei szerkezetekre. A kollokáció egyik eleme egy szó lesz: a vizsgálandó bővítménykeret egyik (kiválasztott) bővítményi helyén megjelenő tartalmi elem; a kollokáció másik eleme viszont egy összetett struktúra: az ige és az esetlegesen mellette megjelenő vagy megkövetelt egyéb bővítmények együttese, azaz egy igei bővítménykeret. Ezt megtehetjük, szabadon lehet dönteni arról, hogy mit veszünk egy kollokáció egy elemének (Kilgarriff és Tugwell, 2001). Így valójában az adott bővítménynek a bővítménykeret többi részéhez viszonyított lényegességét tudjuk mérni.

A tipikus kérdés tehát, amit vizsgálni tudunk: adott ige illetve igei keret melletti adott bővítményi helyen mely szavak fordulnak elő legjellemzőbben. A megjelenő egyéb bővítmény bármi lehet: igemódosító, vonzat vagy szabad határozó is, a bővítménykeret fogalmába mindegyik beletartozik. A kérdésben megadhatunk egy igetövet és valamennyi bővítményt, függetlenül attól ezeknek a bővítményeknek adott esetben mi a szerepe, és megnézhetjük, hogy egy további bővítményi helyen milyen jellegzetes szavak jelennek meg. Példa: x = 'ad HANG-t -nAk'; y = 'MEGGYŐZŐDÉS', 'VÉLEMÉNY' stb. A fix elem az x, a vizsgált elem az y, a kérdés pedig az, hogy az egyes y-ok közül melyek a jellemzőek. A salience érték akkor lesz magas, ha az y szó gyakrabban fordul elő az x keretben a vártnál, és az y szó maga is gyakori.

Nézzük meg egy konkrét példán az MI és a salience mérték különbségét. Az 'ad -t' keret esetében az MI mérték szerinti csökkenő sorrendben a *tanújel, életjel, ízelítő, személyleírás, áldás* szavakat kapjuk. A salience viszont a *hang, lehetőség, válasz, otthon, tájékoztatás* listát szolgáltatja. Az előbbiek ritka, különleges szavak, az utóbbiak a triviálisabbnak tűnők, mégis ezek a lényegesebbek. Mondhatjuk: az MI nem a lényegeset, hanem a különlegeset mutatja. Az MI által mutatott listára az anyanyelvi beszélő is rácsodálkozhat, hogy tényleg ezeket is 'ad -t' általános keret használatával fejezzük ki, de amiket leginkább érdemes tudni, ha meg akarunk érteni egy magyar szöveget, azok a salience által adott listában találhatók. Egyszerűen fogalmazva hasznosabb ha egy gépi fordító rendszer helyesen le tudja fordítani a 'ad -rA VÁLASZ-t' keretet, mintha helyette az 'ad -bÓl ÍZELÍTŐ-t' keretet kezelné jól.

Említettük, hogy a mai nagyméretű korpuszok méretéből adódóan képtelenség az

DOI:10.15774/PPKE.ITK.2011.004

összes releváns adat manuális feldolgozása. Szükség van olyan eszközökre, mely egy bizonyos nyelvi jelenségről összegzi a korpuszban található információt. Az ismertetett lényegesség-mérési módszer felhasználásával az igei szerkezetek vizsgálatára elkészült egy ilyen eszköz, erről lesz szó a következő fejezetben.

3.2. A „Mazsola” korpuszlekérdező

A *Mazsola* egy internetes felületen hozzáférhető nyelvészeti kutatóeszköz, melynek segítségével a magyar igei vonzatkereteket, az igék bővítményszerkezetét tudjuk kvantitatívan tanumányozni korpuszalapú vagy korpuszvezérelt módszertani keretben. Az elnevezés onnan ered, hogy reményeim szerint izgalmas nyelvi tényeket mazsolázhatunk ki vele a korpuszokból. Pontosan olyan korpusz kezelésére alkalmas, amit a korábbi fejezetek előrevetítettek: a reprezentációnak a 2.1. részben ismertetett modellnek kell megfelelnie, és ezt például a 2.2. részben leírtak szerint tudjuk megvalósítani, előllítani. A modellnek megfelelően a vizsgálható nyelvi alapegység a tagmondat, pontosabban az egy igét és a hozzá tartozó bővítményeket tartalmazó egység. A Mazsola a fentiek szerint előkészített korpuszhoz való speciális korpuszlekérdező eszköz (Sass, 2008, 2009b). Amint látjuk a Mazsola nem pusztán egy klasszikus, konkordanciákat készítő korpuszlekérdező – mint például (Dura, 2006) –, mivel egy olyan speciális korpuszreprezentcióra épül, mely a szerkezetek különféle szórendi megjelenéseit egységesen kezeli.

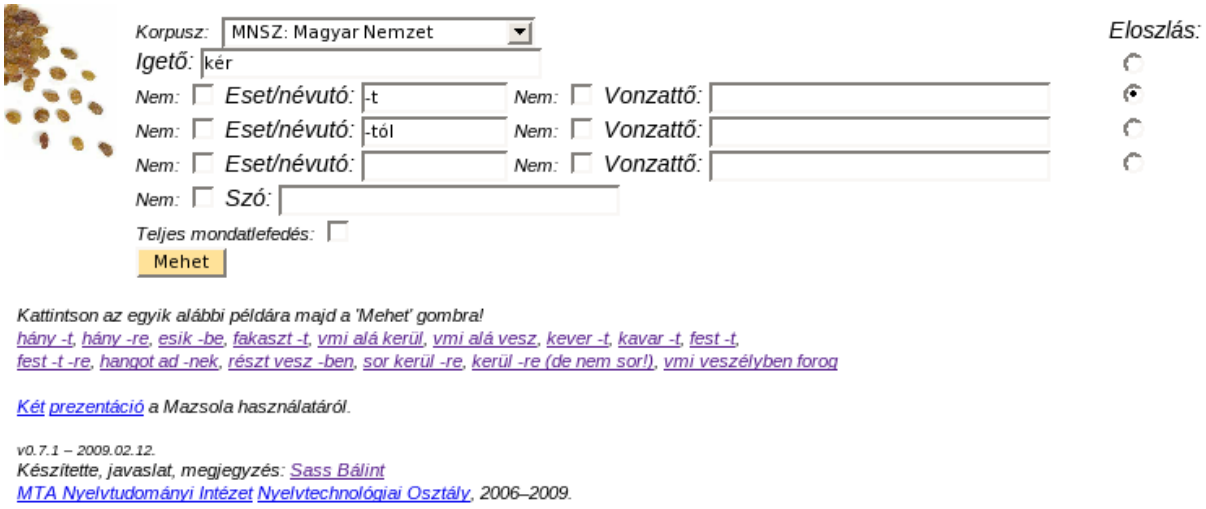
Ingyenes regisztráció után szabadon elérhető a `http://corpus.nytud.hu/mazsola` címen, de regisztráció nélkül is kipróbálható ideiglenesen a `vendeg` felhasználói névvel és a `mazsola` jelszóval. (A kapott közös jelszóval a Magyar Nemzeti Szövegtár közvetlen lekérdezőfelülete is használható.)

Alapvető funkciója, hogy bemutassa a keresett ige leggyakoribb bővítményeit, bővítménykereteit, az ige mellett adott toldalékkal előforduló legjellegzetesebb kollokátumokat. A kollokátumokat – a *salience* (Kilgarriff és Tugwell, 2001) mértékkel mért – jellegzetességük (ld. a 3.1.4. részt) szerint sorbarendezve prezentálja. A Mazsola tehát egy önálló nyelvészeti kutatóeszköz igék és bővítmények, illetve igei szerkezetek korpuszvezérelt tanulmányozására. A 8. ábrán látható a Mazsola felülete, az alábbiakban konkrét példákon keresztül mutatjuk be az eszköz használatát.

3.2.1. Lekérdezhető korpuszok

A felületen (8. ábra) az első mezőben a korpuszt választhatjuk ki. Vizsgálatainkat a 4. táblázatban látható korpuszokon végezhetjük el. Rendelkezésre áll a teljes Magyar Nemzeti Szövegtár anyaga, valamint ennek néhány kiemelt részkorpusza. A 3–10 szavas, írásjelet nem tartalmazó mondatok esetében az előfeldolgozás során nem futtattuk a tagmondatra bontó (2.2.1. rész) modult, ezek a mondatok jó eséllyel pontosan egy bővítménykeretet tartalmaznak. A másik három részkorpusz (amelyek az elsőt értelemszerűen átfedik), az MNSZ három egymástól kellően elütő stílusrétegét képviseli.

DOI:10.15774/PPKE.ITK.2011.004



8. ábra. A Mazsola felülete.

4. táblázat. A Mazsola közzétett kereshető korpuszai.

a teljes Magyar Nemzeti Szövegtár	147,8 millió szó
és ennek alábbi részkorpuszai:	
– 3-10 szavas mondatok	8,0 millió szó
– Magyar Nemzet napilap anyaga	10,6 millió szó
– Index fórum anyaga	12,2 millió szó
– egy kisebb szépirodalomi részkorpusz	10,6 millió szó

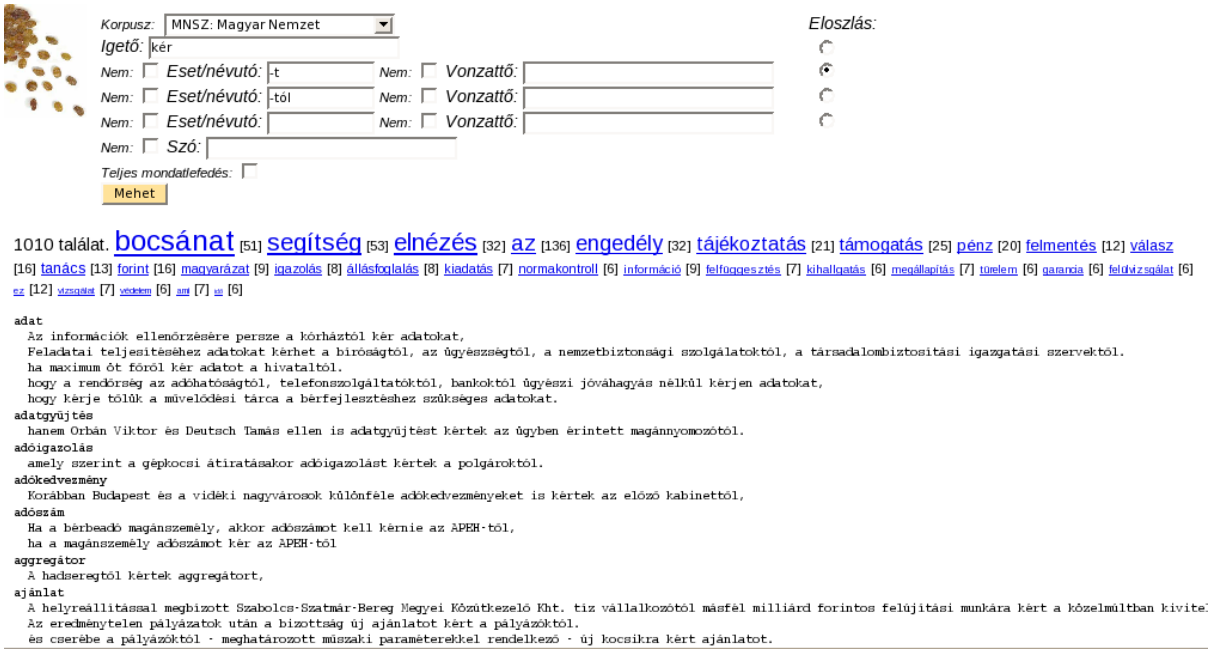
3.2.2. A Mazsola felülete és használata

A felületen megadhatjuk a vizsgálni kívánt igetövet (*Igető* mező), alatta pedig (legfeljebb három) bővítményt specifikálhatunk. A bővítményeket a modellnek (2.1. rész) megfelelően viszonyjelölő (*Eset/névutó* mező) illetve tartalmi elem (*Vonzattó* mező) révén adhatjuk meg. LSzB esetén csak az *Eset/névutó* mezőt töltjük ki, LKB esetén pedig mindkettőt. Az igéhez hasonlóan a tartalmi elemet is szótő formájában kell megadni, erre utal a *Vonzattó* elnevezés, mely abból a szempontból kicsit félrevezető lehet, hogy itt valójában bármilyen bővítmény tartalmi eleme szerepelhet. A *Vonzattó*nél használhatunk szóközzel elválasztott szótőlistát is. Az esetet többféle kiírt formában is elfogadja a felület (helyes megadás például: 't', 'tárgy', '-bA', 'babe', '-ba / -be' stb.), emellett a szokásos latin elnevezés hárombetűs kódja is megfelelő (ACC, DAT, ILL stb.).

A *Nem* jelölőnégyzetek megjelölésével a találati halmazból kizárni kívánt elemeket határozhatunk meg. Kétféleképpen használható: vagy kizárjuk adott viszonyjelölős bővítmény jelenlétét (sor elején álló *Nem*), vagy pedig amellett, hogy megköveteljük adott viszonyjelölős bővítmény jelenlétét, kizárunk bizonyos tartalmi elemeket (sor közepén lévő *Nem*).

Lejjebb, a szintén tagadható *Szó* mezőben szabadszavas kereséssel szűkíthető a vizsgá-

DOI:10.15774/PPKE.ITK.2011.004



9. ábra. A Mazsola válaszképernyője. Fent a lekérdezőfelület, alatta a ‘kér -t -tÓl’ bővítménykeret tárgyként megjelenő jellegzetes szavak, legalul pedig a korpuszpéldák láthatók.

lat. Itt szóközzel elválasztva több szót is megadhatunk, illetve tetszőleges kiterjesztett reguláris kifejezést használhatunk. Ha a *Teljes mondatlefedés*t jelöljük meg, csak azokat a mondatokat, tagmondatokat kapjuk meg eredményül, amelyekben kizárólag a megadott bővítmények fordulnak elő. Ilyenkor a találati halmaz természetesen általában lényegesen kisebb, esetenként üres is lehet.

A képernyő jobboldalán az *Eloszlás* alatt, a megfelelő sor mellett lehet megjelölni, hogy melyik az a bővítmény, amelyet vizsgálni akarunk, azaz hogy melyik bővítmény fejeként megjelenő jellegzetes szavak listáját kérjük.

A 8. ábrán látható példában arra kérdezünk rá, hogy a ‘kér -t -tÓl’ bővítménykeretben melyek a tárgyként megjelenő tipikus, jellegzetes szavak. A honlapon található példákra kattintva azonnal világossá válik az egyes beviteli mezők szerepe.

Itt jegyezzük meg, hogy bár a Mazsola program elsődlegesen az igék tipikus bővítményeinek vizsgálatára készült, használhatjuk az ellenkező irányban is: kereshetjük vele adott bővítményekhez tartozó jellegzetes igéket is. Ha a felületen (ld. a 8. ábrát, illetve a 12. ábrát az 54. oldalon) szótőként a ‘*szerződés*’ szót adjuk meg, és az *Eloszlás* gombot az (üresen hagyott) igető mező mellé állítjuk, az eredményben legelöl a ‘*köt*’, ‘*megköt*’, ‘*aláír*’ igék szerepelnek, de mindjárt ezután következik a ‘*felmond*’, ‘*felbont*’, ‘*lejár*’, majd kicsit hátrébb a ‘*bont*’ és a ‘*felrúg*’ is.

DOI:10.15774/PPKE.ITK.2011.004

3.2.3. A Mazsola válaszképernyője

A Mazsola 9. ábrán látható válaszképernyője három részből áll. A már ismert lekérdezőfelület alatt látjuk az eredményt: a kívánt bővítményként tipikusan megjelenő szavak listáját, jellegzetesség szerinti csökkenő sorrendben. A lényeges szavak e listájában csak az 5-nél gyakoribb szavak szerepelnek, nagyobb betűméret jelzi a nagyobb jellegzetességet (nagyobb salience-értéket), szögletes zárójelben tájékoztatásképpen az előfordulási szám szerepel. A képernyő alján az összes releváns korpuszpéldát is megkapjuk. A találatok az *Eloszlás* alatt megjelölt szempont – a vizsgált bővítményként megjelenő szó – szerint csoportosítva, betűrendben jelennek meg. A fent kék színnel megjelenített jellegzetes szavakra kattintva a hozzájuk tartozó releváns korpuszpéldákhoz jutunk. Az MNSZ-ben meglévő morfológiai elemzésnek és a korpuszelemzésnek (vö: 2.2. rész) köszönhetően természetesen a lekérdezések során megadott igének az összes alakját megkapjuk, az elváló igekötő is a helyére kerül. A rendszer válaszideje – százmillió szavas korpuszméret mellett – mindössze néhány másodperc.

A 9. ábrán példaként látható lekérdezésben a következő kérdésre keressük a választ: „A '*kér vmit vkitől*' keretben mik a jellemző tárgyként megjelenő szavak?", köznyelvi megfogalmazásban: „Mi mindent szoktunk általában kérni?" A válaszban (9. ábra) a *bocsánat, segítség, elnézés, engedély, tájékoztatás, támogatás, pénz* stb. szavakat, azaz a 'kér BOCSÁNAT-t -tÓl', 'kér SEGÍTSÉG-t -tÓl', 'kér ELNÉZÉS-t -tÓl' stb. szerkezeteket kapjuk, ami plauzibilis, nyelvi intuíciónknak is megfelelő eredmény.

3.2.4. Mire szolgál?

A tipikus kutatási kérdés tehát, amit a Mazsolával vizsgálni tudunk: „Melyek a lényeges szavak, melyek egy adott keret egy (adott viszonyjelölő által meghatározott) bővítményi helyét tartalmi elemként betölthetik?"; másképp: „Mik a jellegzetes szavak, amik egy bővítménykeret LSzB-jében tipikusan megjelennek?"; még másképp: „Mik egy bővítménykeret legfontosabb kollokátumai egy adott LSzB-ben?" A lekérdező fontos tulajdonsága tehát, hogy a kérdésben nem csak egy igét, hanem egy teljes vagy részleges *bővítménykeretet* adhatunk meg és azt vizsgálhatjuk, hogy egy további bővítményi helyen milyen tipikus szavak jelennek meg mellette. Kiderül például, hogy a '*hány vmire*' keret legtöbbször (644-ből 288 esetben) '*szemére hány vkinek vmit*' szerkezetként jelenik meg, vagy hogy a '*megköszörül*' igének szinte kizárólag (147-ből 134 esetben) a '*torok*' lehet a tárgya, az ige szó szerinti jelentése (a várt '*kés*', '*olló*' stb. tárggyal) kivételesen ritka. Vizsgálhatóvá válik a komplex igék önálló bővítményszerkezete (pl.: 'rejt VÉKA·alá -t'), valamint összevethető egy alapige és egy komplex ige bővítményszerkezete is (pl.: 'rejt VÉKA·alá -t' vs. 'titkol -t').

Azáltal, hogy a lekérdezésben teljes bővítménykeretet adhatunk meg nemcsak arra van lehetőség, hogy egy ige jellegzetes tárgyait számba vegyük, hanem hogy ige–tárgy párok jellegzetes alanyait, vagy ige–alany–tárgy hármasok például '-tÓl' ragos jellegzetes bővítményeit vizsgáljuk és így tovább. Ezen a „rekurzív" módon feltérképezhetjük egy ige jellegzetes mintázatait.

A Mazsolával kinyert jellegzetes kollokátumok két részre oszthatók (Sass, 2009d).

DOI:10.15774/PPKE.ITK.2011.004

1. gyakori szavak „szó szerinti” jelentésben; ezek gyakran egy szemantikailag koherens csoportot alkotnak – mint például az ‘*eszik*’ tárgyaként megjelenő különféle ételek, vagy a ‘fest -rA -t’ keretben ‘-rA’ ragos bővítménnyel megjelenő színek;

2. olyan szavak, melyek az igével komplex igét (vagy szólást) alkotnak – mint például a ‘*próba*’ a ‘tesz -rA -t’ ‘-rA’ ragos bővítményei között, a ‘*konyha*’ a ‘hoz -rA -t’ ‘-rA’ ragos bővítményei között vagy az ‘*ördög*’ a ‘fest FAL-rA -t’ tárgyai között.

Ha egy bővítménykeret egy bővítményi helyén igemódosítók, vagy igemódosítók *is* vannak, lényegességi mértékünk – a 2. pontnak megfelelően – ezeket hozza elő. Ez lehetőséget ad komplex igék felfedezésére illetve a komplex igék saját, önálló, az alapigétől legtöbb esetben független bővítményszerkezetének vizsgálatára. Láttuk, hogy az ‘ad -t’ keretben megjelenő lényeges tárgyi bővítmények is sok esetben állandósult szókapcsolatot, komplex igét alkotnak az alapigével. További példák láthatók az 5. táblázatban.

5. táblázat. A Mazsola általá szolgáltatott komplex igék néhány bővítménykeret esetében. Érdekes jelenség, hogy egy anyanyelvi beszélő a bal oldalon található keretből nehezen találja ki a jellemző bővítményi fejet (‘hány -t’ → fitty), ugyanakkor a fordított irányú asszociáció (fittyet → hány) azonnali. Az összes alábbi esetben ilyen aszimmetrikus asszociációs viszonnyal van dolgunk. (E példák kapcsán is látjuk, hogy a rendszer az esetragokat és a névutókat valóban teljesen egyenrangúan kezeli (vö: 28. oldal).)

‘hány -t’	→ fitty
‘hány -rA’	→ szem
‘kerül alá’	→ víz, kalapács, fennhatóság
‘rejt alá’	→ véka
‘hoz alá’	→ tető
‘helyez alá’	→ vád
‘vesz alá’	→ górcső, kalap, tűz

A (vonzat nélküli vagy vonzatos) összetett igék azok a tipikus szerkezetek, melyeknek a vizsgálatára a Mazsola szolgál.

3.2.5. A ritka hibák jelentősége

Ha alaposabban megnézzük az egyes korpuszpéldákat, látszik hogy számos esetben valamilyen hiba folytán helytelen eredményre jut a rendszer, azaz helytelenül állapítja meg az igét és/vagy a bővítményeket.

Amint láttuk, az automatikus feldolgozó lépések egyike sem tökéletes, a nyelvtechnológiában 100%-os pontosságot elérni lényegében lehetetlen.

Egyetlen automatikus számítógépes nyelvészeti program sem tud tökéletes eredményt szolgáltatni, így tartalmaz hibákat a Magyar Nemzeti Szövegtár morfoszintaktikai

DOI:10.15774/PPKE.ITK.2011.004

elemzése, illetve a különböző előfeldolgozó (ld. 2.2. rész) lépések sem tökéletesek. Bár az egyes mondatról sok esetben hibás specifikus megállapítást tesz a rendszer, ettől még igaz az, hogy a bővítmények lényegességéről és az egyes igei szerkezetek jellegzetességéről szóló általános állítások megfogalmazásához biztos alapot ad. Megerősíthetjük azt az ismert tényt, hogy a statisztikus alapú általános állítások igazságára az alkalmazott eljárásban előforduló a ritka hibák nincsenek számottevő hatással (Teubert, 2005; Kilgarriff et al., 2004).

3.2.6. Illusztratív példák

A közzétett korpuszok lehetőséget adnak a különféle stílusrétegű szövegek bővítményszerkezetének összehasonlítására. Látni fogjuk, hogy a különböző stílusrégetű szövegek bővítményszerkezetükben is markánsan különböznek. Alább a Magyar Nemzet és az Index fórum korpuszból nyert, az 'ad -t' keretre vonatkozó adatokat elemzem. A 10. ábrán láthatók a tárgyként megjelenő lényeges szavak, itt a közös elemeket jelöltem meg. Ezek azok a szavak, keretek, melyek stílusrétegtől függetlenül lényegesek. A 11. ábrán látható ugyanez az összehasonlítás, de itt az eltérő szavak vannak kiemelve. Valóban, az 'ad -nAk OTTHON-t' és az 'ad -rÓl HÍR-t' sajtónyelvbe illő keretek, szemben az 'ad IGAZ-t' és az 'ad TIPP-t' kollokviális, hétköznapi, beszélt nyelvi jellegével.

10. ábra. 'ad -t' keret a Magyar Nemzet *(fent)* és Index fórum *(lent)* korpuszban: közös elemek

Második példánkban a *'köt vmit vmihez'* szerkezet jellegzetes tárgyragos (12. ábra) és jellegzetes *'-hOz'* ragos (13. ábra) névszóit látjuk. A kapcsolódó korpuszpéldák halmaza a két esetben természetesen azonos. Mindkét lekérdezésből látszik, hogy a *'köt vmit vmihez'* szerkezet nagyon jellegzetes megjelenése a *'köti az ebet a karóhoz'* szólás. Ez – valamint a hasonló módon vizsgálható számtalan egyéb szerkezet (pl. *'mosolyt fakaszt'*, *'a gyanú árnyéka sem vetődik rá'*, *'üsse kő'* – is alátámasztja a korpuszvezérelt lexikográfiának azon a fontos megfigyeléseit, miszerint egyrészt a többelemű lexikai egységek a nyelvnek kiemelten fontos építőelemei, másrészt az ún. metaforikus, vagy átvitt jelentést sokszor gyakrabban használjuk, mint a konkrét, esetleg történetileg is korábbról adatolható jelentést (Hanks, 2005).

Az idiómák és szólások azonosítása után megvizsgálva az eredményeket, és az egyes bővítményi helyeken megjelenő szavakból szemantikai csoportokat képezve (vö: 3.2.4.

DOI:10.15774/PPKE.ITK.2011.004

11. ábra. ‘ad -t’ keret a Magyar Nemzet *(fent)* és Index fórum *(lent)* korpuszban: eltérések

rész) feltérképezhetjük a különféle igei szerkezeteket, illetve a szerkezetek jelentéselemeit. A *köt vmit vmihez* szerkezet esetében a szó szerinti jelentés (‘*kutyát fához*’) gyakoriságát jóval meghaladja az a metaforikus jelentés, mikor valamilyen „jutalmat” (‘*támogatás*’, ‘*folyósítás*’, ‘*felvétel*’, ‘*engedélyezés*’) valamilyen „feltételhez” (‘*feltétel*’, ‘*határidő*’, ‘*megfizetés*’, ‘*teljesítés*’, ‘*vizsga*’) kötünk. További jellemző szerkezet a ‘*szerződést/megállapodást köt*’ (itt a ‘*-hOz*’ ragos bővítmény célhatározói szerepű), valamint a ‘*vmilyen árfolyamot egy másik árfolyamhoz köt*’ szerkezet, amiben szintén megjelenik a szó szerinti és a metaforikus jelentésben is meglévő „kénytelen együtt maradni” jelentéskomponens (Sass és Pajzs, 2010b).

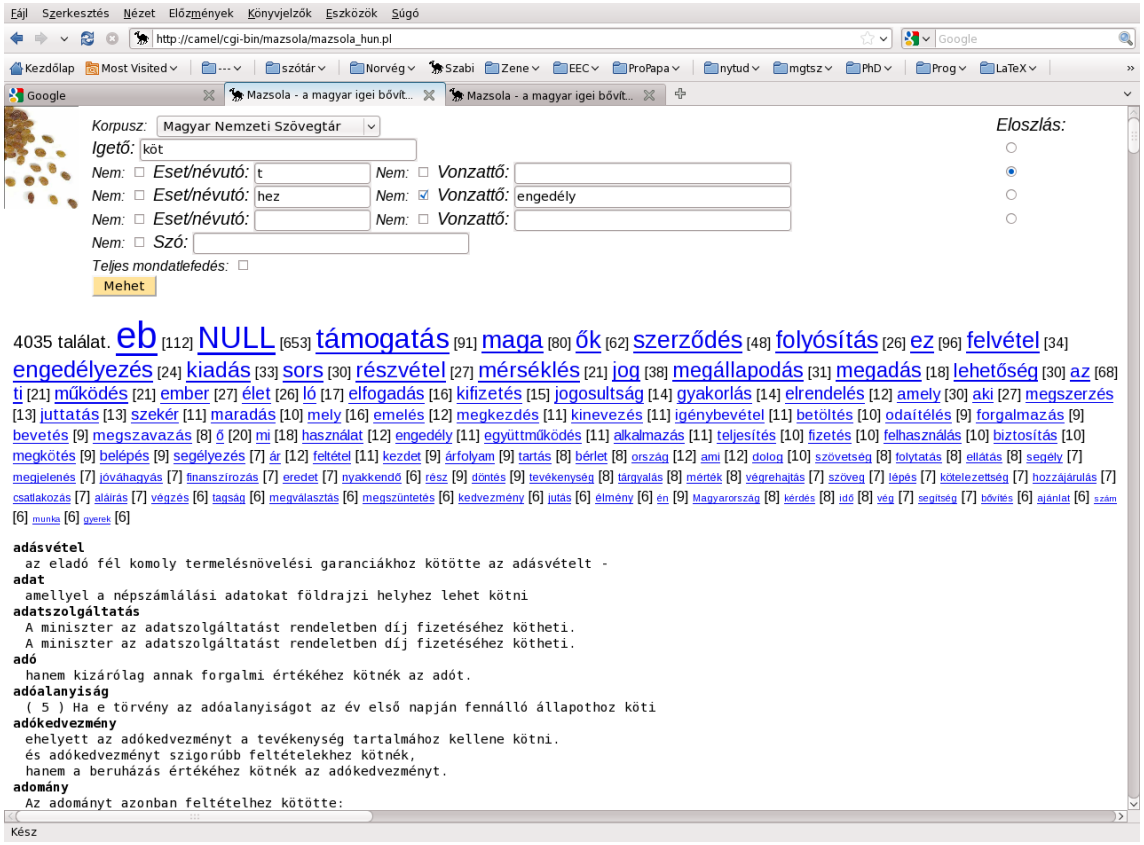
3.2.7. Összefoglalás

A Mazsola egy önálló nyelvészeti kutatóeszköz, mely felépítésében pontosan illeszkedik a kidolgozott tagmondat-reprezentációhoz (ld. 2.1. rész), ilyen módodon reprezentált korpusszal dolgozik, és szofisztikált ugyanakkor hatékony keresési lehetőséget biztosít az igei bővítménykeretek terében. Elérhető a `http://corpus.nytud-hu/mazsola` címen (ideiglenes felhasználói név: `vendeg`; jelszó: `mazsola`). Segítségével egy igei bővítménykeret adott bővítményi helyén megjelenő jellegzetes szavakra kérdezhetünk, illetve kereshetünk rá.

A Mazsola korpuszlekérdezőről szóló **3. tézis** a 111. oldalon olvasható.

Két típusú – kompozicionális és idiomatikus – lényeges igei szerkezet, szókapcsolat van, a Mazsola mindkét típust szolgáltatja (vö: 3.2.4. rész). Fontos, hogy a második típusba tartozó komplex igék (és szólások) kifejezetten gyakoriak a nyelvben (a ‘*hány*’ igét tartalmazó mondatoknak például 8%-a (!) a ‘*fittyet hány*’ szerkezetet tartalmazza). Ezért egy nyelvhasználónak, nyelvtanulónak és egy nyelvtechnológiai (pl.: gépi fordító) rendszernek ugyanúgy, a puszta szó szerinti jelentés ismerete nem elegendő, minden lényeges szerkezetet ismernie kell. A lényeges szerkezeteket kinyerő algoritmusról lesz szó a következő részben.

DOI:10.15774/PPKE.ITK.2011.004



12. ábra. A Mazsola felülete: a *'köt vmit vmihez'* (kivéve *'engedély'*) szerkezet, a benne előforduló jellegzetes tárgyragos szavak listájával. (A NULL – ld. a 38. oldalon is – a tárgyas ragozású igével rendelkező, de explicit tárgyat nem tartalmazó példamondatok tárgyát jelöli.)

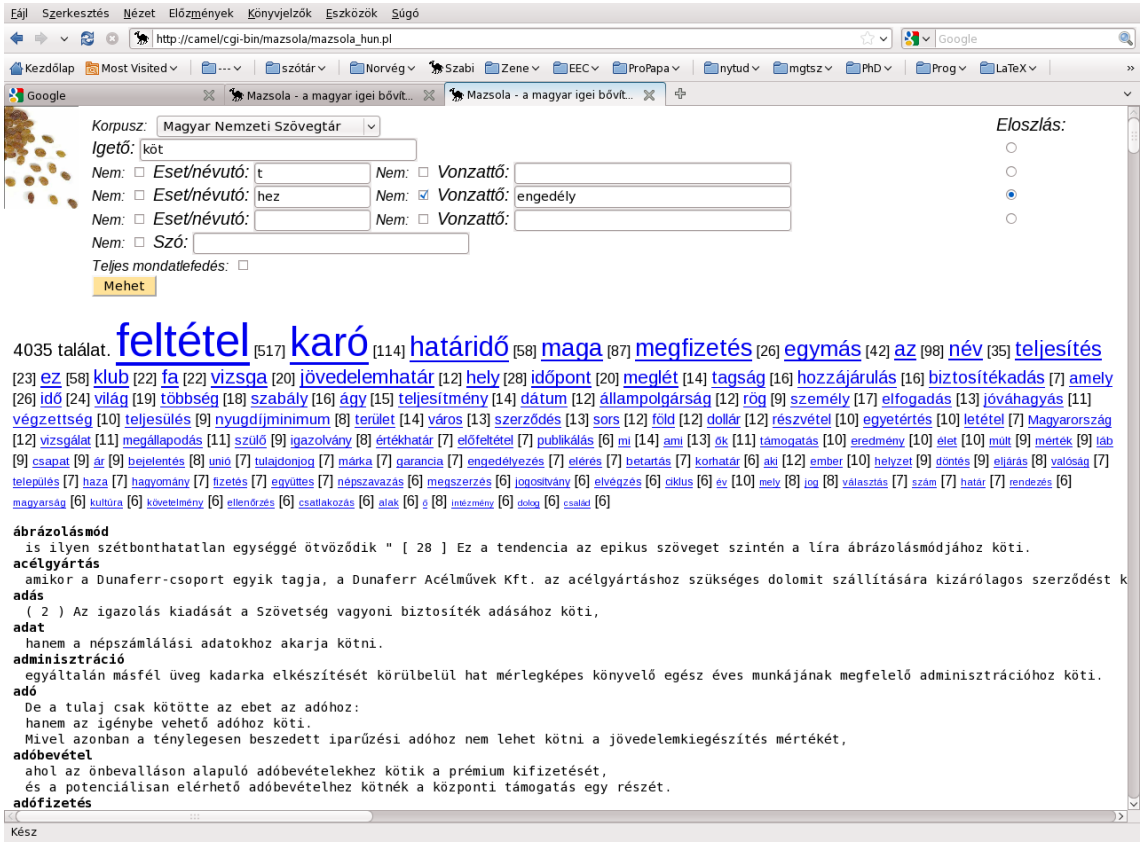
3.3. A jellegzetes igei szerkezeteket kinyerő algoritmus

Az előző fejezetben láttuk, hogy a Mazsola kutatóeszköz képes egyfajta módon összegezni a korpuszból leszűrhető információt. Konkrétan arra képes, hogy megmutassa egy bővítménykeret adott bővítményi helyén megjelenő leggyakoribb, legjellemzőbb szavakat.

Ennél jelentősebb kérdés, hogy egy igének *egyáltalán* mik a jellemző szószerkezetei és vonzatai, az ige mellett a különböző bővítmények milyen kombinációkban szoktak előfordulni. Hogy ezekben a kombinációkban mikor szükséges a tartalmi elem, azaz a konkrét szó, és mikor csak a viszonyjelölő, azaz az esetrag vagy a névutó? Az 50. oldalon említettük, hogy a Mazsola kézi használatával rekurzív módon egy ige összes szerkezetét feltérképezhetjük. A kérdés most az, hogy hogyan lehet egy ige *összes* jellemző szerkezetét összegyűjteni és ezáltal képet rajzolni magáról a nyelvről? Valamint: hogy hogyan lehet ezt az összegyűjtést *automatikusan* elvégezni?

A nehézséget az okozza, hogy sok nyelvhez hasonlóan a magyarban is ugyanazokkal a nyelvi eszközökkel (a magyarban éppen esetragokkal és névutókkal) jelöljük az

DOI:10.15774/PPKE.ITK.2011.004



13. ábra. A Mazsola felülete: a '*köt vmit vmihez*' (kivéve '*engedély*') szerkezet, a benne előforduló jellegzetes '*-hOz*' ragos szavak listájával. (Az ábrán látható a 'Nem' jelölőnégyzet használata is: a '*-hOz*' ragos bővítmények közül a fenti módon zárhatjuk ki például az '*engedély*' szót.)

ige összes bővítményét, függetlenül attól, hogy vonzatok vagy szabad határozók, és függetlenül attól is, hogy LKB-k vagy LSzB-k.

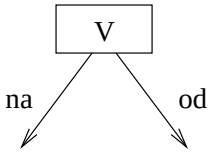
Az algoritmus lényege az lesz, hogy automatikusan felismeri, hogy egyrészt mely bővítmények tartoznak szorosan az igei szerkezethez; másrészt hogy mikor szerves része a szerkezetnek a tartalmi elem, és mikor csak a viszonyjelölő. Azaz például a 'húz HASZON-t -bÓl' esetében felfedezi, hogy az ige mellett egy lexikálisan kötött – LKB – tárgy és egy lexikálisan szabad – LSzB – '-bÓl' esetragos bővítmény alkotja a szerkezetet.

3.3.1. Az algoritmus működése

Kiindulópont

A most bemutatandó algoritmushoz az ötletet egy igei vonzatkereteket gyűjtő módszer adta (Zeman és Sarkar, 2000), e cikkben leírt megközelítés későbbi alkalmazásával, vagy folyományával az irodalomban nem találkozunk. Az igei vonzatkeretek

DOI:10.15774/PPKE.ITK.2011.004



14. ábra. Két LSzB-t – két viszonyjelölőt –tartalmazó vonzatkeretet ábrázoló függőségi fa. Ez a cseh nyelvű vonzatkeret az eredeti (Zeman és Sarkar, 2000) cikkből való, a magyartól eltérően a csehben a viszonyjelölők elöljárók.

formailag úgy karakterizálhatók, hogy bennük LKB-k nincsenek, csak LSzB-ket tartalmaznak, a 14. ábrán látható függőségi fának felelnek meg.

E módszer során a korpuszmondatokból nyert hosszabb (több bővítménnyel bíró) bővítménykeretektől indulva, először különféle statisztikai vizsgálatokkal megállapították az egyes keretekről, hogy elfogadhatók-e igei vonzatkeretként. Ha egy adott keret nem volt elfogadható, akkor törölték a listáról, választottak egy egy pozícióval rövidebb listán szereplő keretet, és annak gyakorisági értékéhez hozzáadták az eredeti keret gyakorisági értékét. A módszer során tehát egyfajta kumulatív gyakoriságot számolnak, így a végső eredmény minden igei szerkezethez egy kumulatív gyakorisági mérőszám. Elemzik, hogy milyen módon lehet kiválasztani az ilyen eggyel rövidebb *successor* kereteket, majd megjegyzik, hogy „végül kiderült, hogy a véletlenszerű kiválasztás nagyobb pontosságot eredményezett" („we eventually discovered … that a random selection resulted in better accuracy") (Zeman és Sarkar, 2000). Ennek nyomán én is is ezt a véletlenszerű kiválasztást veszem át, amennyiben több successor-jelölt is van; a statisztikai vizsgálatok helyett pedig egy egyszerű 5-ös gyakorisági küszöböt alkalmazok.

A 22. oldalon található 3. definíció alapján az igei vonzatkeretek az igei szerkezeteknek egy részhalmazát képezik, így a 2.1. részben ismertetett modellünkben nyilvánvalóan reprezentálhatók. Ennek következtében, ha a rendelkezésre álló fenti módszert minden általunk kezelendő ige szerkezet összegyűjtésére szeretnénk alkalmazni, akkor azt kell kidolgozni, hogy hogyan *terjesszük ki* erre a nagyobb halmazra. Valamilyen módon tehát alkalmassá kell tenni arra, hogy ne csak a vonzatkereteket, hanem az általában vett igei szerkezeteknek megfelelő bonyolultabb adatstruktúrát is kezelni tudja.

Az ötlet gyökere egyszerűen az, hogy a korpuszmondatokból nyert bővítménykeretekben *nemcsak a viszonyjelölőt, hanem a tartalmi elemet is eltároljuk*, azaz teljes mondatvázakat tartunk nyilván. Ahhoz, hogy az eljárást valóban képessé tegyük az összes fajta igei szerkezet kezelésére, néhány technikai kérdést kell még megoldani. Újra kell definiálni a *kerethossz* fogalmát; valamint, tudva, hogy mondatvázakat, azaz csak LKB-t tartalmazó struktúrákat tárolunk, valamilyen módon biztosítani kell, hogy az igei szerkezetekben *LSzB-k is megjelenhessenek* (és így pl. megkaphassunk a kívánt komplex igéket, pl.: '*részt vesz vmiben*'-t a '*részt vesz csatározásban*' és hasonlók alapján). (A részleteket alább tárgyaljuk „Az algoritmus lépései" részben.)

A létrejövő gyakoriságra épülő lexikai kinyerő algoritmus tehát összesíti az adott igét tartalmazó mondatvázakat, és automatikusan előállítja az igéhez tartozó jellegzetes

DOI:10.15774/PPKE.ITK.2011.004

igei szerkezetek listáját. Alapötlete a következőképpen is megfogalmazható: induljunk ki a teljes korpuszreprezentációból, és hagyjuk el azokat a bővítményeket, melyek nem részei a szerkezetnek, illetve (a viszonyjelölőt megtartva) azokat a tartalmi elemeket, melyek nem részei a szerkezetnek (hanem csak éppen, esetlegesen egy vonzati helyet töltenek ki), és így a korpusz igei szerkezeteihez jutunk.

Az algoritmus bemenete

Ez a lexikai kinyerő eljárás tagmondatokra bontott, szintaktikailag részlegesen elemzett korpuszt vár bemenetként. A tagmondatok egy igét és annak bővítményeit kell, hogy tartalmazzák, a szintaktikai elemzésnek pedig meg kell határoznia a tagmondat igéjét, a bővítmények fejét valamint az ige és a bővítmények közötti viszonyjelölőket.

Egy dependenciaviszonyokkal részlegesen annotált korpuszra van szükség, ahol az ige és annak bővítményeként megjelenő névszói csoportok közötti egyszintű dependenciaviszonyok vannak megjelölve. Pontosan az a reprezentáció szükséges itt, amit a modell (2.1. rész) megad, és amit a 2.2. részben leírtak szerint állíthatunk elő.

Az algoritmus lépései

Ebben a részben részletesen bemutatjuk az algoritmus lépéseit. Az összes fajta igei szerkezetet kezelő lexikai kinyerő algoritmus a következő lépésekből áll: (1) gyakorisági listát készítünk a keretekből, (2) alkalmas módon kiegészítjük ezt a listát, (3) hossz szerint rendezzük, (4) majd egy speciális módon összegezzük („örököltetjük") a ritka keretekhez tartozó gyakorisági értékeket, végül (5) egy módosító/javító lépést hajtunk végre. Lássuk ezeket a lépéseket részletesen:

1. *Gyakorisági lista.* Előkészítő lépésként a tagmondatok modell szerinti reprezentációjából gyakorisági listát készítünk, azaz megszámoljuk, hogy melyik mondatváz (tagmondat-reprezentáció) hányszor fordul elő a korpuszban. Természetesen nem csak a teljesen azonos tagmondatok reprezentációja egyezik meg, hanem csak szórendben eltérő tagmondatoké, sőt az összes olyan tagmondaté is, melyekben a különbségre a reprezentáció érzéketlen, azaz a különbség nem a bővítmények viszonyjelölőiben vagy tartalmi elemeiben jelentkezik (6. táblázat).

2. *Kiegészítés.* A kezdeti keretlistát mondatvázak, azaz tartalmi elemekkel teljesen kitöltött (csak LKB-t tartalmazó) bővítménykeretek alkotják. Ezt a listát kiegészítjük a következőképpen. (1) Az összes mondatvázból töröljük az összes tartalmi elemet és az így kapott, csak LSzB-ket tartalmazó kereteket a listához adjuk. (2) A legfeljebb *két* bővítményt tartalmazó mondatvázakból *váltakozva töröljük* a tartalmi elemeket azaz először az egyiket töröljük és a másikat tartjuk meg, aztán az elsőt tartjuk meg és a másikat hagyjuk el.

   Példa: a '*A szaxofonos vállat vont.*' mondat mondatvázából ('`ige`=von `-t`=váll `-0`=szaxofonos') a váltakozva törlés után az alábbi három további bővítménykeret keletkezik:

DOI:10.15774/PPKE.ITK.2011.004

6. táblázat. Azonos reprezentációval (mondatvázzal) bíró tagmondatok. Az első két példa csak szórendjében tér el, a harmadik mondatból pedig azért kapjuk ugyanazt a reprezentációt, mert a bővítmények jelzői ill. az igeidő nem része a modellnek.

1. *'amely nagy sikert aratott szakmai körökben'*
2. *'amely szakmai körökben nagy sikert aratott'*
3. *'amely hazai körökben osztatlan sikert arat'*

A közös reprezentáció: 'ige=arat -bAn=kör -t=siker -0=amely'

'ige=von -t -0'
'ige=von -t -0=szaxofonos'
'ige=von -t=váll -0'

Így előállítjuk a tagmondatoknak megfelelő, elvben lehetséges összes igei szerkezetet. Erre az átalakításra azért van szükség, hogy a listában megjelenjenek az LSzB-t (azaz esetleges vonzatot) illetőleg LSzB-t és LKB-t vegyesen tartalmazó szerkezetek is. Ez az átalakítás teszi lehetővé, hogy végül a 2. ábrán (21. oldal) szereplőhöz hasonló 1 LKB + 1 LSzB típusú szerkezeteket – a komplex igéket – is eredményül kapjuk. (Az LSzB-t és LKB-t vegyesen tartalmazó szerkezetjelöltek közül csak a két bővítményt tartalmazóakat állítjuk elő, így az *'őrizetbe vesz vkit vmi miatt'* (1 LKB + 2 LSzB) típusú szerkezetek nem jelennek meg a kiegészített listán sem. Ezek a szerkezetek viszonylag ritkák, alapesetben nem foglalkozunk velük.)

A létrehozott – immár LSzB-t is tartalmazó – bővítménykereteket 0 gyakorisági értékkel vesszük hozzá a listához. Ha a példában említett mondatváz 2-szer fordul elő a korpuszban, akkor a kiegészített lista-részlet gyakoriságokkal együtt így néz ki:

2	'ige=von -t=váll -0=szaxofonos'	*'(a) szaxofonos vállat von'*
0	'ige=von -t -0'	*'vki vmit von'*
0	'ige=von -t -0=szaxofonos'	*'(a) szaxofonos vmit von'*
0	'ige=von -t=váll -0'	*'vki vállat von'*

A 31. oldalon említetteknek megfelelően az alanyt itt speciálisan kezeljük. Alanyi LSzB nincs, vagyis az alanyt csak akkor tartjuk nyilván, ha kötött. Ez lényegében azt jelenti, hogy feltételezzük, hogy minden igei szerkezet kiegészülhet alannyal. A fenti keretlista módosított végső változata tehát:

2	'ige=von -t=váll -0=szaxofonos'	*'(a) szaxofonos vállat von'*
0	'ige=von -t'	*'vmit von'*
0	'ige=von -t -0=szaxofonos'	*'(a) szaxofonos vmit von'*
0	'ige=von -t=váll'	*'vállat von'*

Ebben a példában természetesen az utolsó szerkezet (*'vállat von'*) a helyes, elvárt, kinyerendő szerkezet; és amint látni fogjuk az algoritmus által eredményül

DOI:10.15774/PPKE.ITK.2011.004

adott listán valóban ezt fogjuk nagy gyakorisági mérőszámmal, előkelő helyen megtalálni.

3. *Rendezés.* Ezután hossz szerint csökkenő sorba rendezzük az igei bővítménykeretek 2. lépés szerint kiegészített teljes listáját. Ehhez meg kell határoznunk a *kerethossz* fogalmát. Azt szeretnénk, hogy ez megfeleljen annak az intuitív jelentésnek, hogy az adott igei szerkezet (az igén kívül) *hány elemből* áll: így ebbe bele kell számolnunk a viszonyjelölőket és a tartalmi elemeket is. Egy szerkezet hosszát a benne található viszonyjelölők és tartalmi elemek összesített száma adja, másképp fogalmazva: az LSzB-k 1-et, az LKB-k pedig 2-t érnek. Kerethossz = LSzB-k száma + 2 · LKB-k száma. A 6. ábrán (33. oldal) látható szerkezet hossza tehát 3 (1 LKB + 1 LSzB), a 19. oldalon látható (4) szerkezeté 2 (1 LKB), a 14. ábrán láthatóé szintén 2 (2 LSzB), az 5 ábrán (32. oldal) szereplőé pedig 1.

   Így „eggyel rövidebb keretnek" minősül nemcsak az eggyel kevesebb LSzB-t tartalmazó (pl.: 'kér -t -tól' vs. 'kér -t') keret, hanem adott LKB helyett LSzB-t tartalmazó keret is (pl.: 'kér ELNÉZÉS-t -tól' vs. 'kér -t -tól'). Megjegyzendő, hogy az azonos kerethosszal rendelkező keretek egymáshoz viszonyított sorrendi helyzete a rendezett listán esetleges.

4. *Gyakoriság-öröklötetés.* Végighaladunk a keretek listáján a leghosszabbtól kezdve a rövidebbek felé, és a ritka kereteket – melyek gyakorisága 5 vagy annál kisebb – elhagyjuk a listáról. Az elhagyott kerethez tartozó gyakorisági értéket azonban megőrizzük, mégpedig úgy, hogy *hozzáadjuk* egy alkalmas rövidebb keret gyakorisági értékéhez. Az alkalmas keret tehát egyrészt az eredetinél rövidebb, másrészt *illeszkedik* az eredeti keretre, és az ilyen tulajdonságokkal bírók közül a lehető leghosszabb. Azaz eggyel (ha nincs ilyen, akkor kettővel, ha ilyen sincs, akkor hárommal stb.) rövidebb illeszkedő keretet keresünk a lista sorrendje szerint, és az első ilyen örökli (veszi át) az elhagyott keret gyakorisági értékét.

   18. definíció. *Illeszkedés.* A rövidebb keret akkor illeszkedik, ha bővítményeinek halmaza az eredeti keret bővítményeinek részhalmaza, és ahol az eredeti keret LKB-t tartalmaz, ott a rövidebb keretben nincs eltérő konkrét szó. Az '`ige=`von `-t`' 1 hosszúságú keret például illeszkedik az '`ige=`von `-t=`váll' 2 hosszúságú keretre; utóbbi pedig illeszkedik az '`ige=`von `-t=`váll `-0=`szaxofonos' 4 hosszúságú keretre.

   Előfordul, hogy több lehetséges rövidebb illeszkedő keret van, ekkor – ahogy erre korábban (az 56. oldalon) utaltunk – ezek között a választás esetleges: egyszerűen a listán előrébb szereplő keret örököl.

   E lépés eredményeképpen tehát a ritka kereteket „elfelejtjük", illetve rövidebb illeszkedő keretek formájában összegezzük a gyakoriságukat. Ha a korábbi példának megfelelően a '`ige=`von `-t=`váll `-0=`szaxofonos' keret gyakorisága csak 2, akkor ez törlődik, és adott esetben a '`ige=`von `-t=`váll' keret gyakoriságához adódik hozzá. Természetesen az eredeti mondatváz ennek a keretnek is megvalósulása, így jogosan képviseli azt; jogosan mondjuk, hogy helyesen jártunk el, mert az eredeti tagmondat valóban a '*vállat von*' szerkezetet tartalmazta, amit

DOI:10.15774/PPKE.ITK.2011.004

most sikerült kinyerni. Az algoritmusnak lényegi tulajdonsága, amit most láttunk: mindig megpróbálja a lehető legspecifikusabb ugyanakkor elegendően gyakori szerkezetet megőrizni.

5. *„Visszaellenőrzés”.* A véletlenszerű választás miatt előfordulhat, hogy egyes mondatvázakhoz tartozó gyakoriság „lejjebb öröklődik a listán a kelleténél”, azaz egy túl általános kerethez rendelődik, miközben specifikusabb keretek is megjelennek, illetve megmaradnak a listán. Ha egy szerkezet megvan a listán (azaz nem törlődött), akkor arra törekszünk, hogy az összes őt megillető gyakoriság ennél a specifikus szerkezetnél halmozódjon fel, hitelesen mutatva a szerkezet gyakoriságát.

```
for f in összes szerkezet listája hosszútól rövidig rendezve
  for x in f-nél rövidebb összes szerkezet
    for k in x mondatvázai
      if f illeszkedik k-ra: k-t áttesszük f-hez
```

15. ábra. A visszaellenőrzési algoritmus pszeudokódja.

Ezt a következőképpen érjük el: a megmaradó keretek listáján (a hosszútól a rövidig) még egyszer végighaladva ellenőrizzük illetve szükség esetén biztosítjuk, hogy az elhagyott mondatvázak gyakorisága mindig valóban a lehető legspecifikusabb megmaradó szerkezethez rendelődjön hozzá. Ehhez természetesen minden szerkezetnél nyilván kell tartani, hogy az ott előállt összesített gyakorisági érték mely része mely mondatvázból ered. Minden keret*hez* (f) megpróbálunk plusz gyakorisági értéket rendelni úgy, hogy megnézzük a nála rövidebb összes szerkezetet (x), és ha ott találunk olyan mondatvázat (k) melyre illeszkedik az aktuális keret, akkor az adott mondatváznak a gyakorisági értékét átvesszük, és hozzáadjuk a jelenlegi értékhez. A visszaellenőrzési algoritmus pszeudokódja a 15. ábrán látható.

Ezzel az algoritmus lépéseit áttekintettük. A fenti lépések lefutása után a megmaradó szerkezeteknek a (4. lépésben leírt módon számított és az 5. lépésben leírt módon korrigált) kumulatív gyakorisági mérőszám szerint rendezett listája adja az összegyűjtött igei szerkezeteket. Ebből láthatunk egy szemelvényt a 16. ábrán, mely a ‘*vet*’ ige gyakoribb szerkezeteit mutatja be.

Elemzés, magyarázat

A fenti példából a kívánt szerkezet (az ‘`ige=`von `-0 -t=`váll’ azaz a ‘*vállat von*’) fog nagy gyakorisági értékkel, elöl szerepelni a végső listában, a következők miatt. Gyakori, hogy a ‘*von*’ mellett a tárgy a ‘*váll*’ szó, az alanyként megjelenő szavak viszont sokkal variábilisabbak ezekben a mondatokban. Azaz a ‘`ige=`von `-0 -t=`váll’ szerkezet sokféle ritka alannyal szereplő mondatra illeszkedik, azok gyakoriságát összegzi; a ‘`ige=`von `-0=`szaxofonos `-t`’ jellegű szerkezetek viszont ritkák maradnak. Az

DOI:10.15774/PPKE.ITK.2011.004

vet -nAk VÉG-t [1463]
vet SZEM-A-rA -t [805]
vet -rA PILLANTÁS-t [708]
vet -t [703]
vet -rA -t [380]
vet PAPÍR-rA -t [377]
vet SZÁM-t -vAl [297]
vet -rA FÉNY-t [267]
vet -bA -t [252]

16. ábra. A '*vet*' igéhez tartozó szerkezetek. Szögletes zárójelben a szerkezethez tartozó, az algoritmus által szolgáltatott gyakorisági mérőszám szerepel.

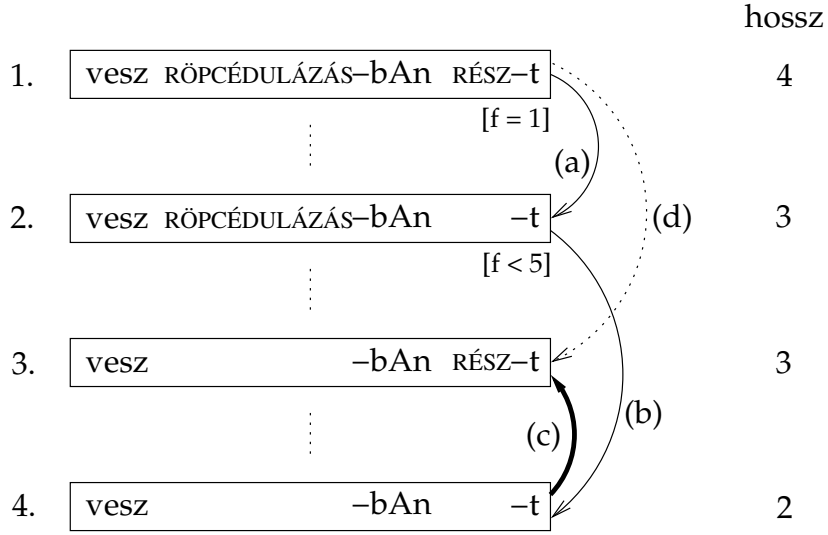
'`ige=`von `-0 -t`' pedig azért nem „nyelheti el" az összes ilyen mondatot, mert *két* egységgel rövidebb az „A szaxofonos vállat vont." jellegű mondatoknál, így azoktól közvetlenül nem tud gyakoriságot örökölni. Abban, hogy a gyakoriságok végül a helyükre kerülnek, fontos szerepe van a visszaellenőrzésnek, ezt a lépést ábra formájában mutatjuk be egy másik szerkezettel illusztrálva (17. ábra).

A 17. ábra azt is bemutatja, hogy milyen mechanizmus vezet ahhoz, hogy az algoritmus eredményeként végül megkapjuk a (vonzatos) komplex igéket. Az algoritmus működésének további megvilágítására nézzünk meg, hogy egy adott egyszerű (angol) bemenő listára milyen eredményt ad a módszer (18. ábra).

Még egyszer összefoglaljuk az újdonságokat, amik lehetővé tették, hogy egy egyszerűbb vonzatkeret-kinyerő algoritmus alapján egy sokkal általánosabb, igei szerkezetek teljes körének kinyerésére képes algoritmust alakítsunk ki: az alapötlet az volt, hogy a bővítménykeretekben *nemcsak a viszonyjelölőket, hanem a tartalmi elemeket is nyilvántartjuk*; a tartalmi elemeket is figyelembe véve meghatároztunk a keretekre egy alkalmas *hosszmértéket*; a váltakozva törlés segítségével biztosítottuk, hogy az igei szerkezetekben LSzB-k – azaz az igei szerkezetek között vonzatkeretek és komplex igék – is megjelenhessenek. Ezeken kívül jelentős még a visszaellenőrzési algoritmus, mely a gyakorisági értékek „javítását" végzi, és felelős a megbízható gyakorisági értékekért.

A függőségi elemzés terminológiáját használva úgy is fogalmazhatunk, hogy ez a módszer a korpuszból származó mondatvázakból jellemző, 1-mélységű függőségi fákat nyer ki, megfelelően kitöltött LSzB-kkel és LKB-kkal. Segítségével felfedezhetjük, hogy a modellen belül egyáltalán milyen típusú szerkezetek léteznek. Egyetértve a (Zarrieß és Kuhn, 2009) cikkel, ismét alátámaszthatjuk, hogy nem jogos (vö: 19. oldal), az az egyébként bevett (vö: 24. oldal) hozzáállás, hogy eleve csak bizonyos meghatározott típusú szerkezeteket vizsgálunk. Mivel a különböző típusú kifejezések átfedik egymást, a hozzájuk rendelt gyakorisági értékek torzulhatnak, eltérhetnek a valóságostól. Nem megfelelő például, ha például ige+tárgy párokat vizsgálva 'vesz -t' gyakoriságába beleszámoljuk a 'vesz RÉSZ-t -bAn' szerkezet előfordulásait is, mivel utóbbi egy teljesen önálló (jelentésű) szerkezet. Ezt a problémát módszerünk automatikusan megoldja.

DOI:10.15774/PPKE.ITK.2011.004



17. ábra. Az algoritmus működésének magyarázata. Tegyük fel, hogy ez a négy szerkezet (az 1. számú mondatváz és a belőle a váltakozva törlés során kialakított igei keretek) ebben a sorrendben szerepel a hosszúság szerint rendezett listán (a 2. és 3. keret sorrendje azonos hosszik miatt esetleges). A „jó" szerkezet nyilván a 3. számú komplex ige, azt szeretnénk, ha az 1. szerkezet gyakorisága [f=1], erre a szerkezetre öröklődne, ennél összegződne. Amiatt azonban, hogy a 2. szerkezet éppen előbb szerepel a listában, nem a (d) hanem az (a) nyíl mentén öröklődik ez a gyakoriság, mivel a listában *előrébb* szereplő illeszkedő szerkezet örököl. Ezután, mivel a 2. szerkezet gyakorisága még mindig túl alacsony, ez is törlődik, és az összegyűlt gyakorisági érték továbböröklődik (b) a még rövidebb 4. szerkezetre. A visszaellenőrzési lépésben aztán előáll a kívánatos helyzet: a 3. szerkezet – mely egyébként már nagyobb mennyiségű gyakoriságot összegyűjthetett az eredetileg *alatta* sorakozó egyéb 3 hosszúságú keretek „elől" – átveszi (c) az 1. mondatváznak megfelelő gyakoriságot a 4. szerkezettől, mivel 3. illeszkedik 1.-re.

Az algoritmus által szolgáltatott gyakorisági mérőszámok az adott igei szerkezetre illeszkedő korpuszmondatok összeszámlálásából adódnak. Az algoritmus minden szerkezethez egyértelműen hozzárendeli azokat a mondatokat, melyek egy-egy találattal gyarapítják gyakorisági mérőszámát, azaz minden szerkezet mérőszáma *más-más* mondatok összeszámlálásából adódik. Ha egy mondat több szerkezetre is illeszkedik, akkor az algoritmus véletlenszerűen *dönt*, hogy az adott mondatot melyik szerkezethez számítsa. Ez azt jelenti, hogy a 'vesz -bA -t' gyakorisági mérőszámába például *nem* számítanak bele a '-bA'-ragos LKB-t tartalmazó különféle szerkezetek ('vesz FIGYELEM-bA -t', 'vesz IGÉNY-bA -t', 'vesz ŐRIZET-bA -t', 'vesz KÉZ-bA -t', 'vesz CÉL-bA -t' stb.). E specifikus szerkezetek gyakorisági mérőszámainak összege éppen jelentősen meg is haladja az általános szerkezetét. A 'vesz FIGYELEM-bA -t' és a 'vesz -bA -t' gyakorisági mérőszáma *nem fed át*, az előbbi 5063 db rá illeszkedő mondat összeszámlálásából adódik, az utóbbi pedig 524 db *az előbbiektől különböző* mondat összeszámlálásából, melyekben nem a '*figyelem*' szó szerepel '*-ba/-be*' raggal. Természetesen ugyanígy igaz ez minden specifikusabb-általánosabb viszonyban lévő szerkezetre. Úgy is

DOI:10.15774/PPKE.ITK.2011.004

Input:

3 ‘ige=take into=account obj=measure’
3 ‘ige=take into=account obj=enterprise’
3 ‘ige=take into=account obj=development’
3 ‘ige=take into=account obj=requirement’
3 ‘ige=take into=account obj=change’
3 ‘ige=take into=consideration obj=future’
3 ‘ige=take into=consideration obj=information’
3 ‘ige=take into=consideration obj=refraction’
3 ‘ige=take into=consideration obj=rarity’
3 ‘ige=take into=consideration obj=preference’

Result:

15 ‘ige=take into=account obj’
15 ‘ige=take into=consideration obj’

18. ábra. A módszer működését bemutató angol példa. Amint látjuk, a ritka konkrét szavak kihullanak, az egyszerű bemeneti mondatvázlistából a megfelelő két igei szerkezetet (vonzatos komplex igét) kapjuk. (A sorok elején a megfelelő gyakorisági értékeket szerepelnek.)

mondhatjuk, hogy minden szerkezet rekurzívan „kihasítja” a maga részét a formailag az általánosabb szerkezethez tartozó mondatokból. A fentiek miatt a különböző bonyolultságú szerkezetek gyakorisága közvetlenül összehasonlíthatóvá válik.

Megemlíthető, hogy – mivel minden igei keretet összevet az összes nála rövidebb kerettel – az algoritmus az igei mondatvázak számában négyzetes futási idejű, ami elég nagy erőforrásigényt jelent, tekintve, hogy a leggyakoribb igéhez (a létigéhez) a 187 millió szavas Magyar Nemzeti Szövegtárból, nagyjából másfél millió (!) mondatváz tartozik. Hatékonysági szempontból nagy nyereség, ha ahelyett, hogy az összes korpuszmondattal egyben dolgoznánk, egyszerre csak egy ige mondatvázain futtatjuk az algoritmust. Ez minden további nélkül megtegető, mivel úgyis csak az azonos igét tartalmazó keretek illeszkedhetnek egymásra.

3.3.2. Az algoritmus kiértékelése

Kiértékelési módszerek és korábbi eredmények

A TSZK-kinyerés klasszikus kiértékelési módszere az n-best-listákat használja (Evert és Krenn, 2001). Ez a következő lépésekből áll:

1. előállítják a TSZK-jelöltek ranglistáját, azaz sorba rendezik őket a kinyerő eljárás (általában asszociációs mérték) által adott pontszám/mérőszám szerint – nyilván az a jó, ha az eljárás a valódi TSZK-kat sorolja a lista elejére;
2. e lista egy kezdőszeletében humán annotátorok megnézik, hogy hány valódi TSZK-t talált a kinyerő eljárás;

DOI:10.15774/PPKE.ITK.2011.004

3. a pontosság a valódi TSZK-k százalékos aránya lesz.

A state-of-the-art kiértékelő módszer pontosság-fedés (P-R) grafikonokkal dolgozik (Evert, 2005). Ekkor előzetesen manuális munkával megjelölik az összes valódi TSZK-t az TSZK-jelölteket tartalmazó listán, majd n-best lista kiértékelést végeznek $n := = 1 \dots c$ -re, ahol c a jelöltek száma, az eredményt pedig grafikonon ábrázolják. Ha az x tengelyen a fenti n szerepel, akkor pontosság grafikont kapunk, de szerepelhet az x tengelyen a (jelöltlistára vonatkozó) fedés is, ezek az ún. pontosság-fedés grafikonok. Utóbbiakat egy mérőszámban is össze lehet foglalni, ez a grafikon menti átlagos pontossági értéket megadó: *mean average precision* (MAP).

A több ezer tagból álló listáknak csak első 50-100 elemére kiterjedő vizsgálata nyilván nem ad hiteles képet a teljesítményről. Ezen a szakaszon általában nagy mértékű a pontosság ingadozása. Az n-best listák fő hátránya tehát az, hogy egy függvényt azzal akar jellemezni, hogy egy önkényesen kiválasztott pontján milyen értéket vesz fel. Ha nincs lehetőségünk az egész függvényt kiszámolni, nyilván azzal tehetjük biztosabbá a kiértékelést, hogy több mérést végzünk, több n-re kiszámoljuk a pontossági értéket, a tapasztalatok alaján lehetőleg a jelöltlisták 5-10%-áig érdemes elmenni (Evert, 2005). Az n-best listák másik hátránya, hogy nem adnak semmilyen fedési mérőszámot, a P-R grafikonok annyiban jobbak, hogy a jelöltlistára vonatkozó fedést is szolgáltatják. Ez a fedési érték azonban semmit nem mond a valódi fedésről, amire általában valóban kíváncsiak vagyunk, hogy ti. a nyelv összes TSZK-ja közül mekkora arányt képes megtalálni a módszer.

A fenti két kiértékelési módszer nem vethetők össze közvetlenül, de ökölszabályként kimondható, hogy az n-best listával kapott értéket a P-R grafikonok maximális értékével érdemes összehasonlítani. n-best listák esetén a szakirodalomban általában 50-60% körüli eredményekkel találkozunk. A már többször idézett cikkben a P-R grafikonok maximuma 55-65% között van, ami egy ennél valamivel kisebb MAP értéknek felel meg (Evert és Krenn, 2001). Az (Pecina, 2008) cikkben vizsgált 55 mérték közül a legjobb előforduló MAP érték: 69% (megjegyzendő, hogy 52%-os baseline mellett). Máshol a klasszikus χ^2 mértékkel 57%-os legjobb MAP értéket értek el (Ramisch et al., 2008), a cikkben szereplő grafikon tanúsága szerint ez nagyjából 65%-os maximális pontossági értéknek felel meg. Magyar vonatkozásban a *kölcsönös információ* (mutual information, MI) segítségével elért 54%-os 250-es n-best listán kimutatott eredményt említjük (Oravecz et al., 2004, 2005).

A nem kényelmetlenül hosszú jelöltlisták előállításának szokásos módszere, hogy csak bizonyos küszöbérték fölötti jelölteket vizsgálják, A statisztikai módszerek alacsony gyakorisági értékek mellett nem megbízhatóak, valamint hosszú listák esetén a manuális annotálás is kivihetetlenné válhat. A hátrány az lehet, hogy az TSZK-knak (type szinten) esetleg jelentős részét elveszítjük. Kimutatták, hogy bár valóban nagy lehet a ritka TSZK-k száma, de kinyerésükre (főleg a hapaxok esetében) a mai módszerekkel (asszociációs mértékek alkalmazásával) nincs esély, ezért a küszöbérték alkalmazása logikus döntés (Evert és Krenn, 2001).

A kiértékelési módszer mellett az összehasonlíthatóság végett adathalmazokról érdemes közölni a következő jellemzőket (Evert és Krenn, 2001): szintaktikailag homogén-e a jelöltlista, azaz mindig valamilyen konkrét szintaktikai viszony áll fenn az elemek

DOI:10.15774/PPKE.ITK.2011.004

között, vagy csak egyszerűen egymás közelében lévő szavakkal dolgozunk; milyen ez a fennálló szintaktikai viszony; a jelöltlistának hány százaléka valódi TSZK.

Az alkalmazott kiértékelési eljárás

Teszteléshez a 187 millió szavas Magyar Nemzeti Szövegtár (Váradi, 2002) megfelelően processzált változatát használtuk, a 2.2. részben leírtak szerint a modellnek megfelelően (2.1. rész) előkészítve.

Esetünkben szintaktikailag homogén szerkezetekkel foglalkozunk, abban a tág értelemben, hogy mindegyik illeszkedik a modell általános sémájára. Annotált jelöltlistán található valódi igei szerkezetek számáról nem tudunk adatot közölni az alábbiak miatt.

A tartalmi elemek tárolása miatt jelentős méretűre növekedik a lehetséges igeiszerkezet-jelöltek száma (Sass, 2006a). A Magyar Nemzeti Szövegtárban 4,368 millió féle pontosan *két* névszói csoport bővítményt tartalmazó mondatváz van (mint amilyennel az 1. ábrán (a 20. oldalon) találkoztunk). A váltakozva törlés (2. oldal) után egy 17,472 millió elemű listát kapunk. Ebbe nem számoltuk bele az esetleg kizárólag hosszabb mondatban megjelenő két bővítményi hellyel bíró szerkezetjelölteket. Egy ekkora lista kézi annotálása kivihetetlen. Következésképpen nem tudunk P-R grafikonokat készíteni vagy MAP értékeket számolni (Evert, 2005), csak az n-best listák módszerére hagyatkozhatunk (Evert és Krenn, 2001; Evert, 2005).

A kiértékelés során tehát az n-best lista módszert alkalmaztuk. A megbízhatóság növelése érdekében *két* független annotátorral dolgoztunk, és több n-re is elvégezzük a kiértékelést. Az igei szerkezeteket kinyerő eljárásunkat a Magyar Nemzeti Szövegtár 8000 leggyakoribb igéjére futtatunk le, az eredményként kapott szerkezetlista 50-es küszöbérték feletti része több mint 47000 elemű. A teljes listát egészében is vizsgáltuk, hogy képet kapjunk az általános teljesítményéről, aztán típusonként külön kiértékeléseket is végeztünk, hogy feltérképezzük a módszer erősségeit és gyengéit.

A típus (16. definíció a 32. oldalon) jelölését kicsit redundáns módon, a jobb olvashatóság kedvéért kiegészítjük az igei szerkezet hosszával, mostantól az eredeti típusjelölés előtt kettősponttal elválasztva feltüntetjük a hosszt is: `[01]` helyett `[1:01]`, `[11]` helyett pedig `[3:11]` lesz.

A típusok szerinti megoszlást az 7. táblázat tartalmazza. Szemben az adott (pl.: igetárgy) felépítésű TSZK-kra koncentráló vizsgálatokkal, itt e táblázatban térképeztük fel, hogy modellünkben egyáltalán milyen típusú igei szerkezetek fordulnak elő. Az igei szerkezetek „nullelemeként" megjelennek itt a `[0:00]` típusú puszta igék is, mint például a '*történik*'. Ezek definíció szerint nyilvánvalóan nem többszavas szerkezetek, a teljesség kedvéért azonban a kiértékelésbe belevesszük ezt a (jelentős számú) csoportot is.

A listából előzetesen kiszűrtük a következő jelölteket, melyek nyilvánvalóan nem idiomatikus igei szerkezetek:

– ha névmás volt a tartalmi elem, kivéve a visszaható igéknél megjelenő '*maga*' és '*egymás*' névmást;

DOI:10.15774/PPKE.ITK.2011.004

7. táblázat. Az eredménylista típusok szerinti megoszlása. Szürkével megjelöltük a legfeljebb két bővítményt tartalmazó szerkezeteket.

	LSzB: 0	1	2	3	4	5
LKB: 0	5006	10790	8509	2140	256	9
1	10647	9077	44	-	-	-
2	1148	160	2	1	-	-
3	91	1	-	-	-	-
4	20	-	-	-	-	-
5	3	-	-	-	-	-

– ha nagybetűs szó volt a tartalmi elem, ez a lépés lényegében a tulajdonnevek elhagyását jelenti;
– az egyértelműen valamelyik korábbi elemzési fázis hibája miatt rossz jelölteket, (pl. rossz igetőazonosítás, helytelen morfológiai elemzés) mivel nem az előzetes lépések, hanem csak a lexikai kinyerő eljárás teljesítményét akartuk mérni.

Az 7. táblázatban szürkével megjelölt legfeljebb két bővítményt tartalmazó típusokra végeztük el az n-best listás kiértékelést. Két független annotátorral megjelöltük a valódi igei szerkezeteket az első $n = 500$ jelölt között, majd típusonként az első $n = 100$ (illetve $n = 200$) jelölt között. Az annotáció során – az igei szerkezetek definícióját (22. oldal) tekintetbe véve – a következő kritériumok alapján tekintettünk egy jelöltet valódi igei szerkezetnek (Sass, 2009c):

1. nem tartalmaz LKB-t *vagy* az ige és az LKB(-k) által alkotott igei résznek van egy (legalább valamilyen mértékben) idiomatikus jelentése, és a szerkezetből nem hiányzik ehhez a jelentéshez elengedhetetlen elem;

2. *és* az adott egyszerű (nincs LKB) vagy komplex (van LKB) igének valóban van az igei szerkezetben lévő LSzB-k által megadott vonzatkerete, és ez a vonzatkeret teljes.

Ez alapján a puszta igék (`[0:00]` típus) megfelelnek, ha nem tárgyasak és megállnak pusztán alannyal.

A csak LSzB-t tartalmazó (`[n:0n]` típusú) szerkezeteknél a kritérium a vonzatkeret teljességét követeli meg (pl.: 'kér -t -tŐl'). Bár a modellben nem különülnek el automatikusan a vonzatok és a szabad határozók (vö: 1.4.3. rész), most kiértékeléskor szigorúbb kritériumot alkalmazunk: csak a vonzat LSzB-ket fogadjuk el helyesnek. Néhány esetben a spontán nyelvérzék nem volt elegendő a döntés meghozatalához, ilyenkor korpuszvizsgálattal segítettük a döntést.

LKB-t is tartalmazó (`[n:mk]` típusú), idiomatikus igei résszel (és teljes vonzatkerettel) bíró megfelelő szerkezetre példa: 'fér -hOz KÉTSÉG' vagy 'von KÉTSÉG-bA -t'. Nem egyértelmű esetben elfogadandó volt az a szerkezet, amelynek (az angol) fordítása speciális, a speciális fordítás ugyanis valamiféle idiomatikusságot mutat, legalább az

DOI:10.15774/PPKE.ITK.2011.004

egyik nyelven (pl.: 'kimond HATÁROZAT-t' = '*declare*'; 'ír VERS-t' = '*poetize*', ld. még az 1. táblázatot is a 33. oldalon). Az általában határozószóval kifejezett formailag szabad vonzat (pl.: 'érez MAGA-t VHOGYAN') hiánya nem számított hibának, mivel a határozószók eleve nem szerepelnek a reprezentációban.

Amint látjuk, itt egyfajta *szigorú* elfogadási kritériumot alkalmaztunk: a teljesség mellett csak a (legalább valamennyire) idiomatikus szerkezeteket fogadtuk el, ugyanakkor tudjuk, hogy például lexikográfiai szempontból bizonyos kompozicionális szerkezetek is jellemzőek, fontosak és ezért gyűjtendők lehetnek (vö: 44. oldal). Ha az ilyen szerkezeteket is elfogadjuk, természetesen a mostaninál magasabb pontossági értékeket kapunk.

Eredmények

Az eredményeket a 8. táblázat tartalmazza. A jelen dolgozatban tárgyalt feladat újszerűsége, miatt ezeket az eredményeket nem lehet korábbi eredményekkel közvetlenül összehasonlítani (vö: a 64. oldalon idézett százalékos értékekkel), mégis kimondhatjuk, hogy az eredmények általában véve nagyon is jók. Az annotátorok közötti egyetértés (8. táblázat, Cohen-κ) megfelelő, legtöbb esetben 0,6 fölött van, két alkalommal megközelíti a 0,8-at is. Megállapíthatjuk, hogy az annotációs kritériumunk (66. oldal) elfogadható alapot nyújt az annotátoroknak a szerkezetek megítélése során (Artstein és Poesio, 2008).

Az egyszerűbb típusokra jobb eredményeket kapunk, de kiemelendő a felismerési teljesítmény a vonzatos komplex igék (`[3:11]` típus) esetén is. Az n növelésével járó pontosságromlás ismert jelenség (Evert és Krenn, 2001). Az egyszerű gyakoriság sok esetben a jól teljesítő klasszikus mértékekkel (log-likelihood, t-test) majdnem egyező teljesítményt mutat, sőt esetleg az eltérés annyira kicsi, hogy nem is szignifikáns (Evert és Krenn, 2001). Ez egybevág azzal, hogy jelen algoritmus is a mélyben puszta gyakoriságokkal dolgozik. A nem túl mélyreható kiértékelés mellett is világosan látszik az algoritmus jó teljesítménye a modell által megfogható különféle igeiszerkezet-típusok kinyerésében. A pontossági értékek n szerinti ingadozása nem nagy. Ez meglepetés tekintve, hogy a listák első 1-2%-ára terjed csak ki a kiértékelés.

A most következő diszkusszióban a legfontosabb eredményeket (8. táblázat szürke hátterű részeit) kommentáljuk.

Az egyetlen vonzattal bíró egyszerű igék (`[1:01]` típus) esetén a legmagasabb az annotátorok közötti egyetértés. Itt érdemes külön vizsgálni azt az esetet, amikor ez a bizonyos LSzB a tárgy: az egyszerű tárgyas igék osztálya az a csoport, ahol módszerünk a legjobb (közel 100%-os) eredményt hozza, a nem tárgy egyvonzatos igéknél (pl.: 'hisz -bAn') a pontosság 80% körüli.

A vonzat nélküli komplex igéken belül (`[2:10]` típus) azt az esetet érdemes különválasztani, mikor az ige mellett csak egy kötött alany szerepel. Ezek a szerkezetek általában kompozicionálisak (pl.: a létige kopulaként jelenik meg bennük). Egy észt nyelvű „többszavas ige" szótárból ezt a típust eleve ki is hagyják (Kaalep és Muischnek, 2008), pedig számos példa van rá, hogy az ilyenfajta szerkezetek is hordozhatnak intézmé-

DOI:10.15774/PPKE.ITK.2011.004

8. táblázat. Eredmények. Átlagos pontossági értékek típus, és az n-best lista n-je szerint. A $\pm$ adatok a két független annotálásnak megfelelő két százalékos értéket jelölik ki. A legfontosabb adatok szürke háttérrel láthatók. Az annotátorok közötti egyetértést mérő Cohen-féle κ értéke szerepel az utolsó oszlopban; ez mindig a megfelelő sorban lévő utolsó értékhez tartozik. Az 'összesen' sorban a rangsorolt teljes lista első 500 elemét értékeljük ki. Ezen 500 szerkezet típusmegoszlása a következő: `[1:01]` 307 db; `[0:00]` 131 db; `[2:02]` 33 db; `[3:11]` 21 db; `[2:10]` 8 db.

típus	$n = 50$	100	150	200	500	Cohen-κ
`[0:00]`	83.0% ± 5.0%	82.0% ± 4.0%				0.53
`[1:01]`	94.0% ± 2.0%	92.0% ± 1.0%	92.0% ± 0.7%	91.8% ± 0.8%		0.77
tárgy	99.0% ± 1.0%	97.0% ± 1.0%	98.0% ± 0.7%	98.0% ± 0.5%		0.75
egyéb	79.0% ± 1.0%	79.5% ± 0.5%	78.7% ± 1.3%	79.8% ± 1.8%		0.68
`[2:10]`	58.0% ± 6.0%	44.0% ± 3.0%				0.64
alany	20.0% ± 6.0%	19.0% ± 6.0%				0.43
egyéb	83.0% ± 1.0%	80.5% ± 1.5%				0.33
`[2:02]`	77.0% ± 7.0%	66.5% ± 8.5%				0.63
`[3:11]`	94.0% ± 0.0%	88.5% ± 3.5%	87.0% ± 3.0%	83.3% ± 3.3%		0.59
`[4:20]`	51.0% ± 7.0%	39.0% ± 5.0%				0.50
összesen	94.0% ± 0.0%	93.5% ± 1.5%	89.3% ± 1.3%	89.5% ± 1.5%	88.9% ± 1.3%	0.65

DOI:10.15774/PPKE.ITK.2011.004

9. táblázat. A Magyar Nemzeti Szövegtárból kinyert első tíz `[3:11]` típusú valódi vonzatos komplex ige.

	ige	LKB	LSzB
1.	van	SZÓ	-rÓl
2.	tesz	LEHETŐ-vÁ	-t
3.	van	SZÜKSÉG	-rA
4.	vesz	ÉSZ-rA	-t
5.	kerül	SOR	-rA
6.	vesz	FIGYELEM-bA	-t
7.	hoz	LÉT-rA	-t
8.	tart	FONTOS-nAk	-t
9.	vesz	RÉSZ-t	-bAn
10.	vesz	TUDOMÁS-Ul	-t

nyesült, idiomatikus jelentést (pl.: 'megélénkül SZÉL'). A nem alanyos `[2:10]` típusú kereteken sokkal jobb a teljesítmény (pl.: 'jön LÉT-rA'), az annotátorok közötti egyetértés viszont itt a legalacsonyabb.

A jelen dolgozatban a `[3:11]` típusba tartozó tipikus vonzatos komplex igék állnak a figyelem középpontjában. A 7. táblázat szerint az ilyen szerkezetek száma és ennek kapcsán jelentősége nagy, és ebben a típusban nagyon sok idiomatikus szerkezet is van. A 8. táblázat bemutatja, hogy az algoritmus meglehetősen jó teljesítményt nyújt itt (közepes egyetértési értékek mellett). A Magyar Nemzeti Szövegtárból származó első tíz valódi vonzatos komplex ige a 9. táblázatban látható.

A két LKB-t tartalmazó `[4:20]` típusnál a pontossági érték alacsonyabb, a sok kompozicionális szerkezet mellett számos intézményesült szerkezetet is találunk: 'vesz IDŐ-t IGÉNY-bA', 'fokozódik SZÉL VIHAROS-vÁ', 'tesz ELÉG-t KÖTELEZETTSÉG-A-nAk'. Ebben a kategóriában már megjelennek a szólások is: 'hajt MALOM-A-rA VÍZ-t'. A bemutatott modell és algoritmus jelentősége éppen az ilyenfajta hosszabb – itt az ige mellett még 4 (!) elemet tartalmazó – igei szerkezetek megtalálásában rejlik. (Érződik, hogy az algoritmus gyakoriságokra épül: a gyakori igei szerkezetek kompozicionális specializációi kapnak magasabb pontszámot, a helyes találatok a lista alsóbb részén vannak. Jövőbeni megközelítés lehet, hogy a már megtalált rövidebb szerkezetek specializációit előre kiszűrjük.)

A még hosszabb/bonyolultabb típusoknál teljes (sokszor a hivatalos stílusrétegbe tartozó) mondatvázakat találunk, melyek általában kompozicionálisak: 'tesz -t LEHETŐ-vÁ TÖRVÉNY' (`[5:21]`). Intézményesült kifejezések is megjelennek, ahol az egyes pozíciókra nagyon kis szóosztályból választhatunk megfelelő tartalmi elemet. Ilyen a következő példa alanyi pozíciója: 'vesz -t ŐRIZET-bA HATÓSÁG'. Azonban a leghosszabb szerkezetek között is találunk iintézményesült szerkezetet: 'elfogad TARTÓZKODÁS·mellett JAVASLAT-t SZAVAZAT-vAl ORSZÁGGYŰLÉS' (`[8:40]`).

Specialitásként említjük, hogy kifejezetten rigid szerkezeteket is megtalálunk a mód-

DOI:10.15774/PPKE.ITK.2011.004

szerrel: az '*annak idején*' formula a `[3:11]` típusú 'van IDŐ-A-n -nAk' formájában jelenik meg. Valamint vonzatos melléknevek és főnevek, azaz elméletileg nem igeközpontú szerkezetek is előkerülnek, melyek mellett általában létigei kopula vagy kiüresedett ige (support verb) szerepel: 'van -rA KÍVÁNCSI', 'van -rA BÜSZKE', 'van -vAl TELE' illetve 'születik -rÓl DÖNTÉS'.

3.3.3. Összefoglalás – az algoritmus jelentősége

E fejezetben mutattuk be a jellegzetes igei szerkezeteket kinyerő algoritmust, mely a dolgozat legfontosabb új tudományos eredménye. Az algoritmus lényegi tulajdonsága, hogy gyakorisági alapon „kitalálja", hogy hány bővítmény szerves része egy igei szerkezetnek, valamint, hogy adott bővítmény esetén csak a viszonyjelölő (esetrag) fontos, vagy a bővítmény fejét adó tartalmi elem is. Az eljárás egy korábbi igei vonzatkeretekre kidolgozott eljárás általánosítása.

A módszer az alábbi két szempont szerint hoz újat. Egyrészt alkalmazkodik az igei szerkezet elemszámához, azaz kettő illetve *több* elemű kifejezéseket ugyanolyan jogon eredményez. Másrészt akalmazkodik ahhoz, hogy bizonyos szerkezeteknek csak a függőségi viszony inherens része, mások pedig a konkrét lexikai elemet is megkötik, azaz LSzB-k és LKB-t – *akár vegyesen* – tartalmazó kifejezéseket ugyanolyan jogon eredményez.

E dolgozatban figyelmünk középpontjában az LKB-t és LSzB-t egyaránt tartalmazó vonzatos komplex igék állnak (9. táblázat). Pontosan ez az a típus, mely kétarcúsága miatt sem a klasszikus vonzatkeretek, sem a klasszikus többszavas kifejezések közé nem tartozik. Mivel afféle határterületre esnek, sokszor ki is kerülnek a kutatások látóköréből (ld. a 19 oldalon a (4) példánál írtakat is). Az algoritmus legfontosabb tulajdonsága – és ebben rejlik jelentősége –, hogy egységes keretben kezelve a vonzatkereteket és a komplex igéket, képes az igék vonzataival és kollokátumaival *egyszerre* foglalkozni, megragadva az ilyenfajta összetett szerkezeteket, melyek igei kollokációk és vonzatkeretek egyszerre.

A kiértékelés megmutatta, hogy az algoritmus pontossága megfelelő, a fontos típusok esetében 80% fölötti arányban eredményez idiomatikus értelmű igei szerkezeteket.

Az algoritmusról szóló **4. tézist** a 112. oldalon fogalmazom meg.

Bár a fentiekben a szigorúbb, idiomatikusságot is megkövetelő kritérium szerint végeztük el az algoritmus kiértékelését, a továbbiakban, mikor az igei szerkezetekből valóban elkészítünk egy egynyelvű szótárt (4.2. rész), látni fogjuk, hogy ha egy lexikai adatbázist vagy szótárt akarunk készíteni az igei szerkezetekből, össze akarjuk gyűjteni az összes jellegzeteset, akkor engedhetünk a megfelelőségi kritériumból (66. oldal), és nem feltétlenül kell megkövetelnünk az idiomatikusságot. Azaz – a 44. oldalon írtaknak megfelelően – kompozicionális szerkezetek is lehetnek olyan jellegzetesek, lexikográfiailag fontosak, jellemzőek a nyelvre, hogy egy szótár anyagába bekerülhessenek.

DOI:10.15774/PPKE.ITK.2011.004

4. fejezet

Alkalmazások

Ebben a részben az előző fejezetekben ismertetett eredmények különböző alkalmazásait tárgyaljuk. Először röviden szólunk a Mazsola korpuszlekérdező felhasználási lehetőségeiről, majd egy hosszabb fejezet következik a jellegzetes igei szerkezeteket kinyerő algoritmus alkalmazásáról egy speciális magyar egynyelvű szótár készítése során.

4.1. A Mazsola közvetlen felhasználása

A Mazsolára tekinthetünk úgy mint egy a Magyar Nemzeti Szövegtárhoz készült alternatív korpuszlekérdező felületre, így haszonnal alkalmazható a magyar nyelv oktatása során, vagy a magyar nyelvet is érintő lexikográfiai munkák készítése során. Az eredeti (`http://mnsz.nytud.hu`) felületen hivatkozhatunk a szavak morfológiai jellemzőinek tetszőleges részletére, és a találatokat szövegkörnyezettel együtt kapjuk meg. A Mazsolában (`http://corpus.nytud.hu/mazsola`) ezzel szemben egy lekérdezéssel érhetünk el különböző szórendi variánsokat, és közvetlenül vizsgálhatjuk az igéket és a mellettük lévő bővítményeket. A kontextus – a modellnek megfelelően – itt mindig egy tagmondat. A két felület közös regisztrációval szabadon hozzáférhető, használható.

4.1.1. Lexikai adatbázisok manuális építése

A Mazsola hatékonyan használható arra, hogy segítségével számos lekérdezést manuálisan lefuttassunk, és az eredmények feldolgozása, elemzése után az igei bővítménykeretekről egy lexikai adatbázist építsünk. Két konkrét projektben vettük hasznát.

A magyar igei WordNet adatbázis (Kuti et al., 2007) építése során két szempontból volt hasznos a Mazsola. Egyrészt fontos, hogy egy igének hány jelentését tudjuk elkülöníteni, ugyanis annyi különböző synsetbe fog bekerülni az adott ige. Az igék egyes jelentéseinek elkülönítésében pedig segítenek a bővítménykeretek, ugyanis a különböző bővítménykeretek sok esetben az ige különböző jelentéseivel járnak együtt (Briscoe

DOI:10.15774/PPKE.ITK.2011.004

és Carroll, 1997): csak bizonyos bővítmények jelen léte esetén van új jelentése az igének, és tekinthető ezáltal egy másik ige szinonimájának. A WordNet-es synseteket ki is egészítették az odaillő, szinonim többszavas egységekkel, komplex igékkel, ezeket az eszköz közvetlenül szolgáltatja.

A másik projekt a Webfordítás (`http://www.webforditas.hu`) magyar-angol gépi fordítórendszer lexikai adatbázisának építése volt. A gépi fordítás szemszögéből nézve a fő kérdés az, hogy melyek azok a szavak, amelyek adott ige melletti adott bővítményként megjelenve az igei szerkezet speciális fordítását követelik meg. Ezek legtöbbször éppen a Mazsola által megadott lényeges szavak. Az ilyen szavakat tartalmazó igei szerkezeteket – lényegében megintcsak a komplex igékről beszélünk – összegyűjtöttük, emberi erővel lefordítottuk, és így fordítással együtt kerültek be a gépi fordító adatbázisába, ahonnan az aktuálisan lefordítandó mondat hívja elő a hozzá legjobban illeszkedő szerkezetet. Ennek köszönhető, hogy a Webfordítás portálon (`http://www.webforditas.hu`) található magyar–angol gépi fordító rendszer helyesen tudja fordítani az olyan idiomatikus igei szerkezeteket tartalmazó mondatokat, mint például '*Fülön csípték a tolvajt.*' vagy '*Csípi a szememet a füst.*' A Webfordítás által visszaadott kiváló minőségű fordítások a 10. táblázatban láthatók.

10. táblázat. A Webfordítás.hu portál által lefordított két példamondat. A kifogástalan fordítások előállítására a Mazsola segítségével kézzel készített, beépített lexikai adatbázis teszi képessé a rendszert.

magyar mondat	angol fordítás
'Fülön csípték a tolvajt.'	*'They caught the thief.'*
'Csípi a szememet a füst.'	*'The smoke stings my eye.'*

4.1.2. Elméleti nyelvészeti jelentősége

Az említett gyakorlati szempontok mellett helye van egy ilyen eszköznek az elméleti nyelvészeti kutatásban is. Nyilvánvaló, hogy a szigorú igen/nem grammatikalitási döntések sok esetben vita tárgyát képezik, az anyanyelvi beszélők véleménye itt sokszor nagy mértékben eltér. (Sampson, 2007) alapjában elhibázottnak tartja a grammatikus vs. nem grammatikus elkülönítést, a nyelvet, a megnyilatkozások rendszerét egy nyílt mezőn kialakuló úthálózathoz hasonlítja, szerinte a nyelvészetnek lényegében azt kellene leírnia, hogy melyik „útvonal" mennyire szokásos, azaz gyakorisági állításokat kellene megfogalmaznia. Szerinte szabálytalan vagy akár érthetetlennek is tűnő szerkezetekről sem mondhatjuk ki, hogy nem grammatikusak, ha valaki használta, és beszélőpartnere pedig megértette őket.

Ezzel egybecsengő felvetés szerint a nyelvtan feladata nem kizárólag az, hogy a grammatikus és a nem grammatikus mondatokat elkülönítse egymástól, hanem arra kell magyarázatot találnia, hogy bizonyos - adott esetben nem grammatikus - megnyilatkozások miért jelennek meg, és bizonyos adott esetben grammatikusak miért nem

DOI:10.15774/PPKE.ITK.2011.004

(Stefanowitsch, 2006). Eszerint tehát nem a grammatikalitás, hanem a megjelenés, illetve meg-nem-jelenés az elsődleges. Pontosan ennek - a megjelenésnek és a meg-nem-jelenésnek - a kvantitatív vizsgálatára alkalmas a Mazsola eszköz.

Levin (1993)-as művében jelenik meg az igék szemantikájának és viselkedésének kapcsolatáról szóló hipotézise, mely kimondja, hogy „az ige viselkedése, különösen az argumentumainak kifejez(őd)ése és értelmezése tekintetében nagy mértékben függ az ige jelentésétől." Ennek a hipotézisnek a vizsgálatára érdemes a bővítményszerkezetek hasonlóságán alapuló igeosztályokat felállítani és elemezni az így kialakuló igeosztályok szemantikus koherenciáját (Gábor és Héja, 2007) (Sass, 2007).

Néhány további lehetséges kutatási irányt vetek fel a (Sass, 2009a) cikkben, szó esik többek között szólások variációinak vizsgálatáról vagy az igék szinonimitásának és a bővítményszerkezetnek az összefüggéséről. Bemutatom, hogy a Jackendoff (2002, 173. oldal) által említett ún. igei konstrukciós idiómák, melyekben az ige a változó elem, a magyarban is tetten érhetők. Ezekben a szerkezetekben az az érdekes, hogy nem az ige határozza meg a bővítménykeretet, hanem a bővítménykeret – sokszor az igekötővel együtt – a konstrukció által adott, és az ige a variábilis, az igei helyre számos ige behelyettesíthető. A variábilis igei helyet alább V jelöli. Ilyen például az 'át|V <IDŐTARTAM>-t' ('*átmulatja az éjszakát*', '*átalussza a délelőttöt*'), vagy a 'ki|V MAGA-t' ('*kibeszéli magát*', '*kidühöngi magát*', '*kipanaszkodja magát*'). A felületen (vö: 8. ábra a 48. oldalon) az igető sort üresen hagyva főnevek, melléknevek, sőt igekötők vonzatosságát is vizsgálhatjuk a Mazsola segítségével.

4.2. A szótár

A kidolgozott jellegzetes igei szerkezeteket kinyerő algoritmus (3.3. rész) legfontosabb alkalmazása a *Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára* (Sass et al., 2010a). Az algoritmus a szótári anyaggyűjtést valósította meg egészében automatikusan.

E szótár alapegységei nem szavak, hanem a már jól ismert igei szerkezetek A szótári gyűjtés a lehetséges szerkezettípusok teljes spektrumát átfogja, az igei vonzatkeretektől a kollokációkon, intézményesült, idiomatikus kifejezéseken, komplex igéken át a szólásokig. Nemcsak egyes kiemelt jellemzőkre figyelünk, ahogy az a speciális szótárakban (vonzatszótárak, kollokációs szótárak, szólástárak stb.) szokásos, hanem a lexikográfiai hagyománnyal szemben az összes típussal egyetlen szótárban, egységes keretben foglalkozunk, ez lehetőséget ad az átmeneti esetek és a kombinációk bemutatására is. (Megjegyzendő, hogy a modern frazeológiai szótárak a kollokációk mellett figyelmet fordítanak a vonzatok gondos feltüntetésére is (Forgács, 2003; Bárdosi, 2009).) Amennyiben egy ige tipikus szerkezetei között csak vonzatkeret (pl.: '*bízik vmiben*') illetve csak szókapcsolat (pl.: '*csóválja a fejét*') található, akkor ezeket közöljük. A bővítményeket egységesen kezeljük, a szótárba minden jellegzetes szerkezet bekerül, függetlenül attól, hogy a benne szereplő bővítmény vonzat (pl.: '*hisz vmiben*') vagy szabad határozó (pl.: '*történik vmiben*'). Nem csak idiomatikus, hanem tipikus kompozicionális szerkezeteket is közlünk. A vonzatkeretek (pl.: '*végez vmit*') mellett – mivel

DOI:10.15774/PPKE.ITK.2011.004

gyakoriak és tipikusak – külön önállóan jelennek meg azok a szerkezetek, melyekben a vonzatkeret LSzB-jét egy az adott helyen gyakori, jellegzetes kötött szó „tölti be" (pl.: '*munkát végez*'). Mivel megnyilatkozásaink túlnyomó része igék köré épül, az igei szerkezetek révén a magyar nyelv egészéről ad átfogó képet a szótár.

A szótárban minden kellően gyakori, jellemző szerkezet helyet kap. A szótári anyaggyűjtés – a magyar lexikográfiában újszerű módon – automatikusan történt az igei szerkezeteket kinyerő algoritmus (3.3. rész) segítségével, reprezentatív korpusz, Magyar Nemzeti Szövegtár alapján. A szótáraknak nyilván a tipikust, a jellegzetest kell bemutatnia. Jelen esetben szigorúan a kinyerő algoritmus által közvetlenül szolgáltatott objektív *gyakorisági mérőszám*okból indultunk ki: azt, hogy mely szerkezet került be a szótárba kizárólag a szerkezet korpuszbeli gyakorisága határozta meg. A szótárba intuitív alapon nem vettünk fel szerkezeteket. A szerkezetek szintjén érvényesült a gyakorisági elv, azaz a szótár nem a leggyakoribb igék összes szerkezetét, hanem a korpuszban meglévő összes ige leggyakoribb szerkezeteit tartalmazza.

A szótár készítése során a sinclair-i szigorúan korpuszvezérelt megközelítést követjük (Tognini-Bonelli, 2001, illetve ld. még az 1.4.1. részt). Nincsenek előzetes elméleti feltételezéseink, azt fogadjuk el, amit a korpuszban találunk. Az intuícióval szemben a korpuszt tekintjük a nyelvet hitelesen reprezentáló entitásnak. A korpuszban nem elegendő számban előforduló szerkezetektől könyörtelenül megszabadulunk (Hanks, 2008), a szótárban csakis azok a szerkezetek jelennek meg, melyek a korpuszban kellő számban előfordulnak. A nyers szótári anyag automatikusan áll elő a korpusz alapján, ezt az anyagot a lexikográfus nem egészíti ki nyelvi intuíciója alapján hiányzónak vélt szerkezetekkel. A modern lexikográfiai felfogás szerint egy szótár esetében nem elég az, hogy egy elem valóban része a nyelvnek, az is szükséges, hogy megszokott eleme legyen (Hanks, 2008). Ezért nem próbáljuk lefedni az összes lehetséges jelentést és az összes lehetséges használatot (Hanks, 2001), csak a kellően gyakori nyelvi elemeket vesszük bele a szótárba, és ezen elemek mellett gyakorisági mérőszámot is feltüntetünk.

A szigorúan korpuszvezérelt megközelítés tehát úgy nyilvánul meg, hogy szótárunk csakis a felhasznált korpuszban meglévő szerkezeteket tartalmazza. Az automatikusan előállított szerkezetlistához a lexikográfusok *nem adtak hozzá* elemeket nyelvi intuíciójuk alapján, csak azért mert „odaillettek" volna, vagy mert a vélt „alapjelentést" képviselik. A szótár abban az értelemben *teljes*, hogy valóban tartalmazza az összes olyan korpuszbeli igei szerkezetet, melynek a gyakorisági mérőszáma egy meghatározott gyakoriság küszöbnél nagyobb. Az MNSZ 2002-ben készült el, ennek megfelelően a legújabb nyelvi fejlemények nem szerepelnek benne, így ezek szótárunkban sem jelenhetnek meg.

Bizonyos bővítmények vonzatok (pl.: 'hisz -bAn') vagy komplex igék kötött elemei (pl.: 'von VÁLL-t'); mások szabad határozók (pl.: 'történik -bAn') vagy az igével együtt speciális jelentést nem hordozó pusztán gyakori szavak (pl.: 'iszik SÖR-t'). Tiszta esetben az előbbiek *idiomatikus*, az utóbbiak pedig *kompozicionális* bővítmények; valójában azonban számos átmeneti eset létezik, ilyenkor nehéz besorolni a bővítményeket e két kategória valamelyikébe. A 'von VÁLL-t' típusú szerkezetet önállóan, „idiómaként" szokás kezelni, az 'iszik SÖR-t' típusút pedig az 'iszik -t' *alá* szokás besorolni. Auto-

DOI:10.15774/PPKE.ITK.2011.004

matikus módszer híján nem vállaltuk fel annak ódiumát, hogy manuálisan, intuitív alapon döntsünk a fenti két kategória tekintetében, inkább azt a megoldást választottuk, hogy *minden szerkezetet önállóan* kezelünk. Ez abban jelentkezik, hogy minden szerkezet külön példamondattal bír, és főként abban, hogy minden szerkezetnek a saját jogán van gyakorisági mérőszáma. A fő indok, ami miatt ezt a megoldást választottuk az, hogy nem tartjuk szerencsésnek, ha az önálló jelentéssel bíró idiomatikus szerkezetek *formai* alapon más szerkezetek alá sorolódnak. A 'vesz RÉSZ-t -n'-t külön akarjuk kezelni a 'vesz -t -n'-től, és a 'varr NYAK-A-bA -t' szerkezetet is a a 'varr -bA -t'-től. Éppen így, az elkülönítés révén derülhet ki, hogy bizonyos esetekben (ilyen a két imént idézett is), a formailag általánosabb keret jelentősen ritkább, mint a specifikusabb idiomatikus szerkezet, az előbbi szinte mesterséges csomópontként jelenik meg a szótárban

Minden szerkezet „azonos jogon", önállóan szerepel tehát a szótárban. Ez a felfogás azzal a jelentős előnnyel jár, hogy így az egyes szerkezetek – legyenek egyszerű vagy komplex igék – különböző tulajdonságait közvetlenül összevethetjük. Típusuk, bonyolultságuk, vonzat- ill. bővítményszerkezetük vagy gyakoriságuk közvetlenül összehasonlíthatóvá válik. Szerkezetileg és gyakoriság szempontjából eltér például a '*titkol vmit*' (egyszerű ige, gyakoribb) és a '*véka alá rejt vmit*' (komplex ige, szólás, ritkább) szerkezet, jelentésük viszont hasonló, és tárgyat megkövetelő vonzatkeretük is azonos.

Az, hogy az egyes igei szerkezeteket önállóan kezeljük, azzal az előnnyel is jár, hogy a szerkezetek önállóan *mozgathatókká* válnak. Ez ad lehetőséget arra, hogy az adott szerkezethez tartozó szótári anyagot (minden „szócikkrészletet") szigorúan *csak egyszer* írjunk meg, és azt használjuk fel a szótár különféle pontjain. Azaz, hogy ezekből a részletekből utólag (automatikusan) szerkesszük össze a szótárt. Ez is hozzájárul a szótár *egységességéhez*, nincsenek egymásnak ellentmondó előfordulások, megszűnnek az abból eredő problémák, hogy a szótár különböző részein más-más lexikográfus dolgozik. A mozgathatóságból következik a szótár könnyű *kiterjeszthetősége* is: a szótári munkálatok bármikor folytathatók, a szótár kiegészíthető a ritkább szerkezetekkel. A mostani szótári anyagot teljes egészében, egy az egyben felhasználjuk, a gyakorisági küszöböt csökkentjük, és az ezáltal bekerülő szerkezeteket feldolgozzuk, majd automatikusan fésüljük össze a kettőt.

A dolgozat elején, a célkitűzésban említetttük (14. oldal), hogy a létrehozandó szótár nem tartalmaz definíciókat, *definíció nélküli* szótárnak (meaningless dictionary) (Janssen, 2008) nevezhető. Bár a definíció, a jelentés megadása a szótárak egyik legfontosabb jellemzője, van haszna az effajta szótáraknak is (saját fordítás):

> „A legtöbb felhasználó csak alapvető információkat keres a szótárakban, mint például, hogy létezik-e egy adott szó vagy kifejezés, vagy hogy hogyan kell helyesen írni. Ilyen célokra a definíció nélküli szótárak jóval hatékonyabbak és könnyebb őket előállítani." (Janssen, 2008)

Látni fogjuk, szótárunk ezeken túlmutató célokra is alkalmasnak tűnik, ugyanakkor a jelentés megjelenítéséről sem mondtunk le teljesen: az igei szerkezetek jelentését alkalmasan választott korpuszpélda világítja meg.

DOI:10.15774/PPKE.ITK.2011.004

A 4.2.1. részben bemutatom, hogyan jutunk el a puszta szövegtől a nyers szótárig tisztán automatikus úton, nyelvtechnológiai eszközök alkalmazásával; utána az utófeldolgozás automatikus (78. oldal) és manuális (80. oldal) részét ismertetem. Ezt követi egy szemelvény a szótár végső formájából (81. oldal), majd a különféle mutatók (82. oldal), végül pedig a szótár lehetséges alkalmazásairól szólok (85. oldal).

4.2.1. A szövegtől a szótárig

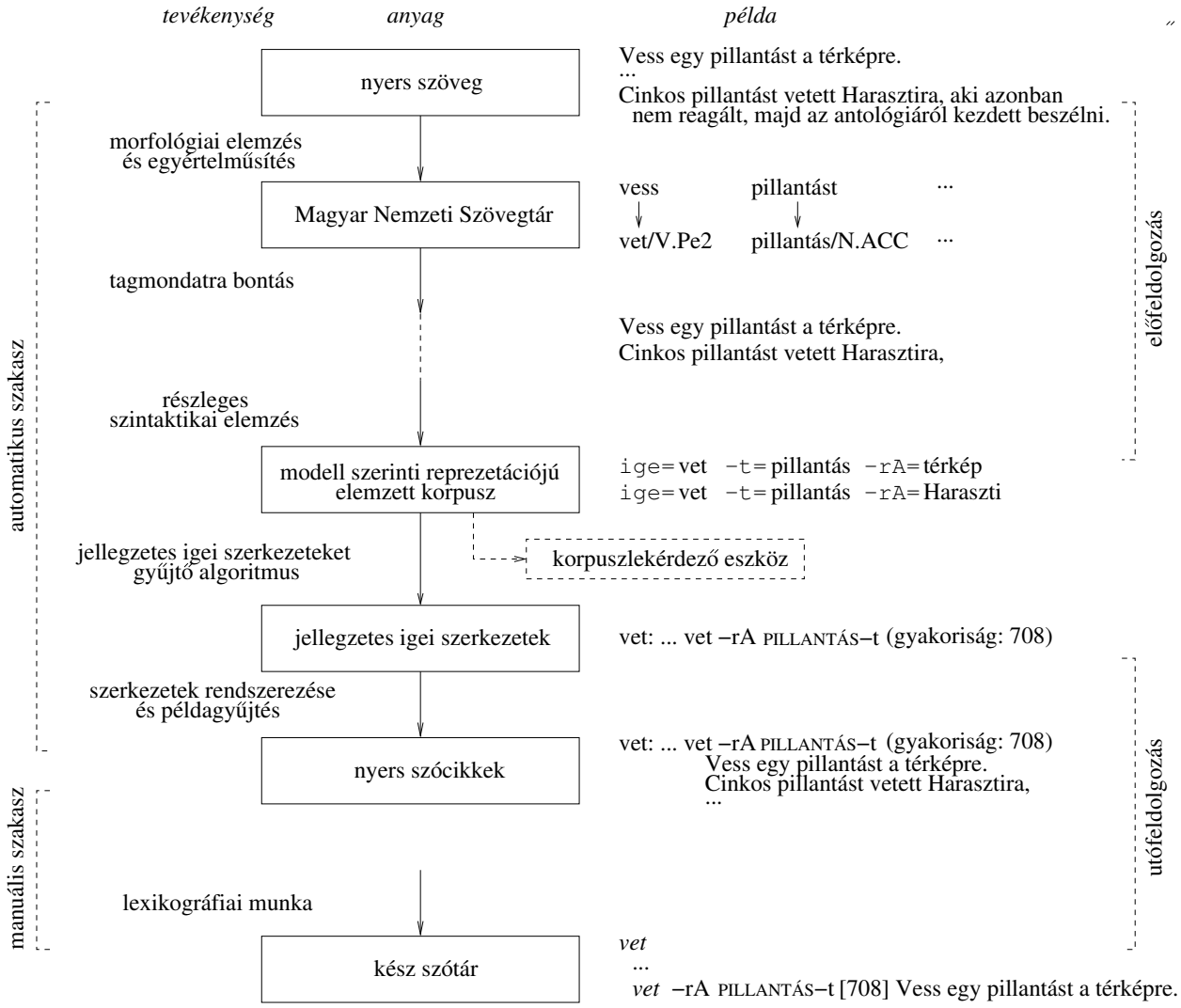
A teljes szótárkészítési folyamat a 19. ábrán tekinthető át. A nyers szövegtől (fent) a kész szótárig (lent) haladunk. Az első („automatikus") szakaszban tisztán automatikus eszközök (ld. *tevékenység* oszlop) használatával állítjuk elő a nyers szótárat emberi beavatkozás nélkül. Az ábra jobb oldalán példával illusztráljuk, hogy hogyan képzelhetjük el az adott lépésben a nyelvi-szótári anyag kinézetét, állapotát. A szótárkészítés legfontosabb lépése a jellegzetes igei szerkezeteket összegyűjtő algoritmus (3.3. rész) futtatása. A szótár felől szemlélve az ezt megelőző lépésekre előfeldolgozásként, az ezt követő lépésekre utófeldolgozásként tekintünk (ld. az ábra jobb szélét). A Magyar Nemzeti Szövegtárban meglévő morfológiai elemzésből ismert az egyes szavak szótöve (pl. '*vet*') és morfológiai kódja (pl. `V.Pe2`, azaz ige, felszólító mód, egyes szám második személy). A további előfeldolgozó lépések (2.2. rész) eredményeként előáll a szintaktikailag részben elemzett, modell szerinti reprezentációjú korpusz, melyen az algoritmust futtathatjuk. Ezt követik az automatikus (rendszerezés és példagyűjtés) és manuális (lexikográfiai munka) utófeldolgozási lépések. Utóbbi során az automatikus szakasz hibáit javítjuk, az esetleges hibás igei szerkezeteket elhagyjuk, és alkalmas példamondatokat választunk az egyes szerkezetekhez, így készül el a végleges szótár.

Fontos kiemelni, hogy az előfeldolgozásban szereplő klasszikus nyelvelemző modulokkal ellentétben az igei szerkezeteket gyűjtő algoritmus már egy *valódi* specifikus lexikográfusi részfeladatot – az anyaggyűjtést – váltja ki, amit hagyományosan manuálisan, korpuszlekérdező eszközökkel, konkordanciák vizsgálatával végeznek. Az igei szerkezetek összegyűjtéséhez szükséges számos korpuszlekérdezés kézi lefuttatása, és az eredmények kézi rendszerezése meglehetősen időigényes lenne, hibalehetőségeket rejt magában, a szótárba bekerülő szerkezetek meghatározása pedig a lexikográfusi intuícióra lenne bízva. Az algoritmus révén az automatikus anyaggyűjtés kiküszöböli ezeket a problémákat: a lexikográfus keze alá dolgozva összegzi a korpuszban található információt.

Az alkalmazott eljárás lényege, hogy képes automatikusan megállapítani, hogy a szerkezetek bővítményi pozíciójában megjelenő konkrét szó kellően gyakori-e ahhoz, hogy LKB-ként, „saját jogán" feltüntessük. Ha igen, akkor az adott kötött szóval kiegészített szerkezetet *teljes jogú, önálló* szerkezetként kezeli. Például bár a 'vet SZEM-A-rA -t' és a 'vet -rA PILLANTÁS-t' szerkezet egyaránt egy '-rA' ragos bővítményt és egy tárgyat tartalmaz, az algoritmus automatikusan állapítja meg, hogy az elsőben a '-rA'-ragos bővítmény LKB és a tárgy LSzB, a másodikban pedig fordítva. Az algoritmus dönti el, fedezi fel tehát, hogy *mik* a korpuszban meglévő szerkezetek.

Kimenetként a korpuszban fellelhető igei szerkezetek listája áll elő, gyakorisági mérőszámmal kiegészítve. A szótár készítése során alapvető a korpuszvezérelt gyakorisági

DOI:10.15774/PPKE.ITK.2011.004



19. ábra. A szótárkészítési folyamat. Áttekinthetjük az automatikus és manuális lépéseket, valamint az előfeldolgozás (2.2. rész), és az utófeldolgozás (4.2.2. és 4.2.3. rész) lépéseit. A dolgozatban kiindulópontunk a már elemzett Magyar Nemzeti Szövegtár, ezért a morfológiai elemzés nem képezi témánk szorosan vett tárgyát, csak a teljesség kedvéért szerepel.

DOI:10.15774/PPKE.ITK.2011.004

szempont, csak azok a szerkezetek kerülnek be, melyek a korpuszban kellő gyakorisággal előfordulnak. Meghatároztunk egy – a szótár tervezett méretének megfelelő – gyakorisági küszöböt, ezen egységes 250-es gyakorisági küszöbérték fölötti szerkezeteket választottuk ki, ezek alkották a nyers szócikkeket.

Így 2347 ige 6854 szerkezete alkotja a nyers szótárat az automatikus szakasz végén. Ezek típus (a jelölést ld. a 65. oldalon) szerinti megoszlása a 11. táblázatban látható.

11. táblázat. A nyers szótár igei szerkezeteinek megoszlása.

típus	példa	db	%
`[1:01]`	‘foglalkozik -vAl’	2808	41%
`[2:02]`	‘ad -t -nAk’	1166	17%
`[2:10]`	‘von VÁLL-t’	1138	17%
`[3:11]`	‘húz HASZON-t -bÓl’	923	13%
`[0:00]`	‘történik’	631	9%
egyéb	‘hajt MALOM-A-rA VÍZ-t’	188	3%
		6854	100%

Ahhoz, hogy ebből egy kiadásra kész szótár legyen, el kell végezni az utófeldolgozás lépéseit. Ezeket ismertetjük az alábbiakban.

4.2.2. Utófeldolgozás: automatikus lépések

Névmástörlés

Úgy ítéltük meg, hogy a *névmásoknak* LKB-ként általában nincs szerepük, az ilyen szerkezeteket nem érdemes önállóként kezelni (pl.: ‘mond AZ-t’). Ezért az igei szerkezetekből az LKB-ként megjelenő személyes, mutató és vonatkozó névmásokat töröltük. A névmás viszonyjelölőjét természetesen megtartottuk, az elhagyás után egybeeső szerkezeteket összevontuk (gyakorisági mérőszámaikat összeadtuk). A névmások közül fontos kivételt képezett a ‘*maga*’ és az ‘*egymás*’, ezeknek jellegzetes szerepe van számos szerkezet (például a ‘von -t MAGA·után’ vagy a ‘el|választ -t EGYMÁS-tÓl’) esetében, ezeket megtartottuk.

A szerkezetek automatikus rendszerezése

A jellegzetes igei szerkezetek között számos olyan van, amely egy másik szerkezet specializációjának tekinthető. (Formálisan éppen akkor specializációja A-nak B, ha A illeszkedik B-re, ld. a 18. definíciót az 59. oldalon). Az ‘*arat*’ igének jellegzetes szerkezete az ‘*arat -t*’ és – ennek specializációja – az ‘*arat győzelem-t*’ is; hasonlóan a ‘*vesz rész-t -bAn*’ specializációja az egyszerű ‘*vesz -t*’ szerkezetnek. Úgy érezzük, hogy a specifikusabb keret az általánosabb „alá” tartozik. Ez az elv azonban sokszor nem ad egyértelmű útmutatást, mert formai alapon a ‘*-t -nAk*’ keret a ‘*-t*’ és a ‘*-nAk*’ alá is tartozhat.

DOI:10.15774/PPKE.ITK.2011.004

A kérdés az, hogy hogyan jelenítsük meg a szótárban a bonyolult specializációs viszonyokat, miközben a gyakorisági szempontra is tekintettel vagyunk. Nem lenne szerencsés, ha a '*vesz rész-t -bAn*' szerkezetet a '*vesz -t -bAn*' szerkezet alá rendelnénk, mert az előbbi nagyon gyakori önálló komplex ige, az utóbbi szerkezet pedig lényegében önmagában nem is létezik.

Az általunk követett és javasolt megoldás szerint az azonos igéhez tartozó szerkezeteket egyszerűen csökkenő gyakorisági sorrendbe tesszük, kiegészítve azzal, hogy bizonyos feltételek teljesülése esetén egyes szerkezeteket mások alá rendelünk. A feltétel a következő: a specializált („alárendelendő") szerkezet gyakorisága kisebb mint az általános szerkezeté, valamint egy bővítményi helyen LSzB helyett LKB-t tartalmaznak *és/vagy* LKB alannyal bővebbek. A cél az, hogy azok a kifejezések, ahol csak az adott szerkezetben használt gyakori szavak jelennek meg, az általános keretük alá tartozzanak, a komplex igék viszont önálló, felső szintű szerkezetként szerepeljenek. Abban bízunk, hogy az előbbiek ritkábbak az általános keretüknél, az utóbbiak viszont gyakoribbak az általános keretüknél, amint, ezt fent a '*vesz rész-t -bAn*' kapcsán említettük. Az esetek jelentős részében ez az összefüggés megállja a helyét, ilyenkor a kitűzött cél teljesül. Amikor ez nincs így, választhatnánk az a megoldást, hogy a lexikográfus felülbírálja az automatikus rendszer döntését, ettől azonban eltekintettünk, hogy minél kevésbé támaszkodjunk a szótárban a lexikográfusi intuícióra.

A feltételnek megfelelő alárendelt szerkezeteket önálló egységként jelenítjük meg a szótárban, beljebb szedéssel jelezve, hogy az általánosabb szerkezet alá tartoznak:

alkalmaz -t [3209]
  alkalmaz MÓDSZER-t [278]

Fontos hangsúlyozni, hogy a fentiek csak a megjelenítést érintik, a szerkezetek önállóságát és az önálló gyakorisági mérőszámokat nem. Továbbra is érvényes, hogy a fenti két szerkezet két különálló, önálló egységet képez és saját jogán rendelkezik gyakorisági mérőszámmal, a specifikusabb szerkezet gyakorisági mérőszáma az általánosabbéval *nem fed át*, abba nem számít bele, azaz jelen esetben a 278 a 3209-en felül értendő.

Az automatikus rendszerezés eredményeként a komplex igék – gyakoriságuk révén – általában a felső szinten maradnak (pl.: 'fel|tesz KÉRDÉS-t', 'helyez KILÁTÁS-bA -t', 'játszik -bAn SZEREP-t', 'jön LÉT-rA'); azok a szerkezetek pedig, melyekben a jellegzetes kötött szó nem jár külön speciális jelentéssel, általában az alsó szintre sorolódnak (pl.: 'fel|emel KÉZ-A-t', 'fizet DÍJ-t', 'iszik SÖR-t').

Példagyűjtés

Az automatikus szakasz utolsó lépéseként példákat gyűjtünk az egyes szerkezetekhez. Minden szerkezethez olyan példa(tag)mondatokat rendelünk, amelyekre a szerkezet illeszkedik (ld. a 18. definíciót az 59. oldalon). Ilyen példákat egyszerűen találhatunk a modell szerint reprezentált korpuszunkban, melyből maguk a szerkezetek is származnak, csak automatikusan illeszteni kell az adott szerkezetet a korpusz tagmondataira. Az a cél, hogy a lexikográfus alkalmas példamondatot választhasson, ezért a 20 leg-

DOI:10.15774/PPKE.ITK.2011.004

gyakoribb olyan példamondatot kínáljuk fel, amelyekben pontosan azok a bővítmények vannak, amelyek a szerkezetben; illetve puszta igei szerkezet esetén a hosszabb példamondatok érdekében bővítményeket tartalmazó mondatokat is megengedünk. A példagyűjtés a Mazsola (3.2. rész) korpuszlekérdező eszköz automatikus használatával valósul meg.

4.2.3. Utófeldolgozás: manuális lexikográfiai munka

Fontos kiemelni, hogy szótárkészítési eljárásunk során a nyers szócikkek (igék köré rendezett igei szerkezetek) teljesen automatikusan állnak elő (ld. a 77. oldalon található 19. ábrán az automatikus szakaszt). Ez fejlettebb megközelítést képvisel a ma szokásos szótárírási eljárásnál, ahol a korpuszkezelés és szócikkek szerkesztése két elkülönülő alrendszert alkot, a szótáríró először (1) lefuttatja a szükséges lekérdezéseket egy korpuszlekérdező eszköz segítségével; (2) megtervezi a szócikket egy DWS-ben (vö: 11. oldal); (3) manuálisan kiválasztja és átmásolja (copy-paste) a szócikkhez szükséges nyelvi adatokat a lekérdezőből a DWS-be; (4) elkészíti a szócikk végleges formáját. Komolyan véve a korpuszvezérelt megközelítést bizonyos feladatokat lexikai kinyerő eszközünkre bízunk. Ez a szócikkhez szükséges minden információt automatikusan nyer ki a korpuszból, azaz a fenti 4 lépésből 3-at elvégez, a szótáríróra a szócikk végleges formájának előállítását hagyva. A nyers szócikkek minden adatot tartalmaznak, ami a szerkesztéshez szükséges, a lexikográfusnak nem kell a korpusz adatait elemeznie és rendszereznie, és megszűnik az adatok átmásolásából adódó hibalehetőség is.

A nyers szócikkek alkalmas XML formátumban állnak elő, a lexikográfus tetszőleges XML szerkesztővel végezheti a manuális lexikográfiai munkát. A „manuális" itt azt jelenti, hogy „nem automatikus", azaz hogy a lexikográfusnak kell szellemi munkával egyedi döntéseket meghoznia a szócikkek szerkesztése során. A szerkesztési lépések technikailag a lehető legegyszerűbbek, általában csak XML attribútumok értékét kell beírni vagy megváltoztatni, az XML fájl részleteit nem kell áthelyezni, a szótár végső formáját automatikusan generáljuk az XML attribútumokba írt utasítások (pl.: `DEL` = „Törlendő") alapján.

A szerkesztés során a lexikográfus feladata, hogy eldöntse, hogy az adott szerkezet valóban létezik-e (vagy csak valamilyen automatikus lépés hibás működése folytán jelenik meg) és a hibás szerkezeteket törölje, valamint hogy alkalmas példamondatot válasszon. Ezekről a feladatokról lesz szó az alábbiakban. A lexikográfiai munkát Pajzs Júlia és Kiss Margit végezte.

A nyers szótár minősége

A szigorúan korpuszvezérelt megközelítés nem engedi meg, hogy a lexikográfus saját nyelvi intuíciója alapján hozzáadjon vagy töröljön hiányzónak vagy fölöslegesnek vélt szerkezeteket. Azonban mivel az automatikus eszközök nem tökéletesek, előfordul, hogy hibás, nem létező igei szerkezetek jelennek meg, ezeket természetesen szükséges törölni. Érdemi munka annak eldöntése, hogy a program által felkínált szerkezetek valóban létező, helyes szerkezetek-e (Sass és Pajzs, 2010b, 19-20. oldal).

DOI:10.15774/PPKE.ITK.2011.004

12. táblázat. A szótárkészítés automatikus szakaszának kiértékelése. A lexikográfusok a 6854 igei szerkezet közül 6266-at fogadtak el jónak, 346 igei szerkezetet hibásnak ítéltek, illetve 121 esetben egy igéhez tartozó valamennyi szerkezetet (összesen 242-t) hibásnak ítéltek. Utóbbi esetben általában az igetőazonosítás volt rossz.

igei szerkezetek száma	6854
elfogadott igei szerkezetek száma	6266
pontosság	**91,4**%

A lexikai kinyerő eljárás és a teljes automatikus szakasz teljesítményének fontos minőségi mutatója, hogy a lexikográfusok a kinyert igei szerkezetek mekkora hányadát találták végül elfogadhatónak (12. táblázat). Bár a szótárba a 44. oldalon tárgyalt megfontolások miatt a kinyert kompozicionális szerkezetek is bekerültek mégis a 3.3.2. részben bemutatott pontossági értékekkel (vö: a 68. oldalon lévő a 8. táblázat *összesen* sorával) nagyjából egyező értéket kaptunk. Ennek oka a kinyerő algoritmuson kívüli automatikus lépések hibáinak összesített hatása lehet. Összességében elmondhatjuk, hogy az automatikusan előállított nyers szócikkek jó minőségűek, a lexikográfusok viszonylag ritkán találkoznak hibás szerkezettel.

Példaválasztás

A lexikográfusok feladata volt, hogy az automatikusan felkínált példák közül kiválasszák a legjobbat, mely végül a szótárba került. A példaválasztás szempontrendszerét Kilgarriff et al. (2008) nyomán Kiss Margit dolgozta ki (Sass és Pajzs, 2010b, 20-21. oldal). A Mazsola lekérdezőnek (3.2. rész) a példaválasztáskor is nagy hasznát vettük. Segítségével a szótáríró bármikor ellenőrizhette az igei szerkezeteket, és a felajánlott korpuszpéldákat. Amennyiben egyik felkínált példamondat sem volt megfelelő, lehetőség volt arra, hogy a Mazsola manuális használtatával további pédákat keressen a korpuszban, és egy megfelelőt illesszen be a példák közé.

Megjegyzendő, hogy a szótárak példáiban általában nem jelzik, hogy melyik bővítmény LKB és LSzB. Ez az információ a gépi feldolgozásra szánt valamint az aktív tanulói szótárakban mindenképpen hasznos. Ez az információ szótárunkban azáltal van explicitté téve, hogy a szerkezetben – melyhez az adott példamondat tartozik – egyértelműen látszódnak a lexikálisan kötött és lexikálisan szabad bővítmények.

4.2.4. A szótár végső formája

A kész szótár hagyományos betűrendes szótári részből, valamint öt különböző mutatóból áll. A szótári rész szócikkeinek szerkezetét a 20. ábrán látható példa mutatja be, az igei szerkezeteket a hagyományos szótári megjelenítéshez hasonlóan igék köré csoportosítva prezentáljuk.

A példaszócikk XML alakjának részlete a 21. ábrán látható.

DOI:10.15774/PPKE.ITK.2011.004

4. Alkalmazások

vet (15728)
vet -nAk VÉG-t [1463] *vessen véget az erőszaknak*
vet SZEM-A-rA -t [805] *hasonló diszkriminációkat vetnek az albán hatóságok szemére*
vet -rA PILLANTÁS-t [708] *vess egy pillantást a térképre*
vet -t [703] *vetem a magot*
vet -rA -t [380] *a humanista könyveket máglyára vetették*
vet PAPÍR-rA -t [377] *vesse papírra az új problémákat*
vet -rA FÉNY-t [267] *ez rossz fényt vet az edzők nevelőmunkájára*
vet SZÁM-t -vAl [297] *vessünk számot eddigi politikánkkal*
vet -bA -t [252] *a tó vizébe vetette magát*

20. ábra. Példaszócikk a kész szótárból. Az alapigét követi a Mazsola lekérdező által szolgáltatott előfordulási száma. Ezután a gyakorisági mérőszám csökkenő sorrendjében következnek a tipikus szerkezetek, megjelenítésük követi a reprezentáció megjelenítésénél leírtakat (ld. a 30. oldalon lévő 4. ábrán szereplő *b)* formát.) Szögletes zárójelben a szerkezet gyakorisági mérőszámát láthatjuk. Látjuk, hogy a szócikkbe valóban csak az említett 250-es küszöbértéknél gyakoribb szerkezetek kerültek be. Ezt követi a példamondat. Mivel a példatagmondatok sok esetben kisbetűvel kezdődnek, illetve vesszővel végződnek, az egységesség kedvéért a szótárban a példákat kisbetűsítve és a végső írásjelet elhagyva közöljük a szótárban. A 'vet SZEM-A-rA -t' és a 'vet -rA PILLANTÁS-t' komplex igék nagy gyakoriságuknak köszönhetően az első szinten jelennek meg, nem rendelődnek a 'vet -rA -t' szerkezet alá (vö: 4.2.2. rész). A dolgozaton végigvonuló, többször is említett (pl.: 5. oldal, 76. oldal) példa jól illusztrálja, hogy a komplex igék milyen változatos formában jelennek meg.

A szótári rész segítségével összevethetjük az egy igéhez tartozó bővítménykeretek gyakorisági viszonyait. A szótárban az '*óv*' és a '*tanul*' esetén is '-t', '-t -tÓl', '-tÓl' gyakorisági sorrendben szerepel ez a három keret. Ez utalhat arra, hogy ezen igék mellett tárgyi és egy opcionális '*-tÓl*'-ragos vonzat szerepel, és nem elhanyagolható gyakoriságúak azok a mondatok, ahol a tárgyat (elliptikusan) elhagyjuk.

4.2.5. Mutatók a szótárban

A mutatók nem kiegészítő funkciót látnak el, hanem szerves részét képezik a szótárnak. Minden mutató a saját szempontja szerint rendezve, csoportosítva mutatja be a teljes anyagot, lehetővé téve a szerkezetek e szempont szerinti összevetését. Fontos megjegyezni, hogy a mutatók az XML alakból emberi beavatkozás nélkül automatikusan generálhatók.

A *gyakoriság szerinti mutató*ban láthatjuk például, hogy a leggyakoribb LKB-t tartalmazó szerkezet a 'van -rÓl SZÓ', vagy hogy eltérő bonyolultságuk ellenére nagyjából azonos gyakoriságúak a 'tart FONTOS-nAk -t', 'be|számol -rÓl', 'tervez -t' és 'él' szerkezetek. Ez a lista akkor hasznos, ha egy adott szerkezettel nagyjából azonos gyakoriságú (vagy ritkább, gyakoribb stb.) egyéb szerkezeteket keresünk.

DOI:10.15774/PPKE.ITK.2011.004


```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE fdvc SYSTEM "fdvc.dtd">
<?xml-stylesheet type="text/xsl" href="fdvc_plain.xsl.xml"?>

<fdvc>

<entry remark="OK">
<verb lemma="vet" freq="15728"/>

<pattern freq="1463">
<frame><p c="-nAk"/><p c="-t" l="vég"/></frame>
<type str="3:11" len="3" fixed="1" free="1"/>
<cits>
  <cit selected="yes">vessen véget az erőszaknak</cit>
  <cit>...</cit>
</cits>
</pattern>

<pattern freq="805">
<frame><p c="-rA" l="szem-A"/><p c="-t"/></frame>
<type str="3:11" len="3" fixed="1" free="1"/>
<cits>
  <cit selected="yes">hasonló diszkriminációkat vetnek
                      az albán hatóságok szemére</cit>
  <cit>...</cit>
</cits>
</pattern>

<pattern freq="708">
<frame><p c="-rA"/><p c="-t" l="pillantás"/></frame>
<type str="3:11" len="3" fixed="1" free="1"/>
<cits>
  <cit type="sentence" selected="yes">vess egy pillantást a térképre</cit>
  <cit>...</cit>
</cits>
</pattern>

...
```

21. ábra. A 20. ábrán látható példaszócikk XML alakjának részlete. Az ige (`<verb>`) a szócikk elején egyszer jelenik meg, a bővítménykeret (`<frame>`) a reprezentációnak megfelelően bővítményeket (`<p>`), és azon belül viszonyjelölőket (`c`) és tartalmi elemeket (`l`) tartalmaz. A szerkezetek típusát is feltüntetjük (`<type>`). Az automatikusan felkínált példák (`<cit>`) között a kiválasztott példát `selected="yes"` jelöli.

A *keretek szerinti mutató* az igék mellett megjelenő bővítményi kombinációkat listázza. Segítségével azt vizsgálhatjuk, hogy milyen különféle igék társulnak egy adott kerettel (pl.: '-bÓl -t' vagy 'MAGA-bAn -t'), csoportokat képezhetünk olyan igékből, amelyek több szerkezetben is jellemzően előfordulnak.

A *koz@kötött szavak szerinti mutató*ban az LKB-ként megjelenő kötött szavak szerint csoportosítva látjuk a szerkezeteket. Szótárunk alapvetően az *igék* viselkedésének feltérképezésére vállalkozik, e mutató segítségével viszont éppen fordított irányú vizsgálatot végezhetünk: az LKB-ként megjelenő *névszók* viselkedéséről kaphatunk (vázlatos) képet, köszönhetően annak a döntésünknek, hogy minden tipikus szerkezetet szerepeltetünk a szótárban, idiomatikusakat és kompozicionálisakat egyaránt. E mutatóból kiderül például, hogy a '*szerződés*' szóval legjellemzőbben együtt járó ige a '*köt*', az '*aláír*' és a '*megköt*'.

DOI:10.15774/PPKE.ITK.2011.004

4. Alkalmazások

Szótárunkban az *igekötőket önálló elemnek* tekintjük. Az igéket a morfológiai elemző alkalmazásával automatikusan választjuk szét igekötőre és alapigére (pl.: 'szétválaszt' → 'szét|választ'). A klasszikus szótári gyakorlat szerint az igekötős igéket (pl.: '*összevon*') külön egységként, külön lexémaként kezelik, a komplex igéket (pl.: '*kétségbe von*') pedig az alapige ('*von*') alatt tárgyalják. (Így az előbbi az *ö*, az utóbbi pedig a *v* betűhöz kerül a betűrendben.) Attól eltekintve, hogy az előbbit egybeírjuk, az utóbbit pedig külön, ezeknek a szerkezeteknek a felépítése valójában nagyon hasonló. Hasznos tehát, ha ezeket a szerkezeteket egy helyen, együtt láthatjuk.

Ezért – amellett, hogy a szótári részben megtartottuk a hagyományos betűrendet – létrehoztuk az ún. *alapige szerinti mutató*t, mely a javasolt csoportosítást valósítja meg: az egy alapigéhez tartozó összes szerkezetet az alapigénél tünteti fel. Így szótárunkban mindkét módon megtalálhatjuk a keresett szerkezeteket. Az alapige szerinti mutató segítségével azt vizsgálhatjuk, hogy milyen igekötőkkel jár egy alapige, és hogyan viszonyul ez a bővítményekhez (pl.: 'át|csap -bA', 'le|csap -rA', 'be|csap -t'). Megváltoztatja-e az igekötő bővítménykeretet (pl.: 'ad -hOz -t' vs. 'hozzá|ad -hOz -t')? Mely igék járnak szinte mindig igekötő nélkül (pl.: '*aggódik*'), illetve szinte mindig igekötővel (pl.: '*ki/megfejt*', '*be/el/lehuny*')?

Ezt egészíti ki az *igekötős keretek szerinti mutató*, melynek segítségével egy másik szempontból tanulmányozható az igekötők viselkedése. Ez a mutató abban különbözik a keretek szerinti mutatótól, hogy itt a bővítménykerethez az igekötőt is hozzávesszük önálló elemként. A magyarban az igekötők sok esetben az igétől függetlenül kapcsolatban állnak a bővítményekkel; bizonyos igekötők együtt járnak bizonyos esetragokkal, másképp fogalmazva az igekötő és az esetrag együtt egy szerkezetet alkot (pl.: 'bele -bA', 'fel -rA', 'ki -bÓl' stb.). Ennek a mutatónak a segítségével az ilyenfajta szerkezeteket tanulmányozhatjuk.

Akkor jó egy szótár, ha a többelemű egységeket bármely elemükből kiindulva könnyen meg lehet találni benne. Ezt általában kereszthivatkozásokkal és/vagy az elemek többszöri feltüntetésével szokták megoldani. Szótárunk a szerkezeteket minden részben külön feltünteti, ezek a bejegyzések felfoghatók a szótári rész megfelelő címszavára utaló kereszthivatkozásként is. A plusz információt mindig a kikeresett szerkezet *környékén* lévő egyéb szerkezetek hordozzák.

A szótár alkalmas arra is, hogy a szerkezetek építőköveit külön-külön vizsgálat tárgyává tegyük. A 'le|von -bÓl KÖVETKEZTETÉS-t' szerkezet elemeiről, részeiről például a következők szerint tudhatunk meg további információt: a 'le' igekötős szerkezeteket a szótári részben és az igekötős keretek szerinti mutatóban; a 'von' alapige szerkezeteit az alapige szerinti mutatóban; a '-bÓl', '-t', '-bÓl -t' kereteket a keretek szerinti mutatóban; a 'KÖVETKEZTETÉS' szót tartalmazó szerkezeteket a kötött szavak szerinti mutatóban; a 'le|von' ige szerkezeteit a szótári részben; az egyetlen 'le -bÓl -t' igekötős kerettel bíró szerkezetet az igekötős keretek szerinti mutatóban; a hasonló (923-as) gyakoriságú szerkezeteket pedig a gyakoriság szerinti mutatóban találjuk meg.

DOI:10.15774/PPKE.ITK.2011.004

4.2.6. A szótár felhasználása

Dolgozatom legfontosabb eredménye a jellegzetes igei szerkezeteket gyűjtő algoritmus (3.3. rész). Ezt az eljárást alkalmaztuk a szótár készítése során, most pedig a szótár felhasználási lehetőségeiről szólunk, azaz az alkalmazás alkalmazásáról.

A szótárt elsősorban a nyelvész szakmának szánjuk. Korpuszalapú elméleti nyelvészeti kutatásban nyelvi adatok hiteles forrásaként, illetve a szerkezetek többszempontú összevetése révén hasznosítható. Gyakorisági adatokat szolgáltat nyelvi adatokra épülő (pl. pszicholingvisztikai) kísérletekhez. Lexikális erőforrásként jelenhet meg a nyelvtechnológia számos területén az információ visszakereséstől a gépi fordításig; valamint hasznos segédeszköz lehet más lexikográfiai munkák készítésekor: korpuszból nyert autentikus adatokat foglal össze, manuálisan ellenőrizve és javítva, alkalmas korpuszpéldákkal kiegészítve.

A nyelvünk összefüggéseire kíváncsi, anyanyelv iránt érdeklődő nagyközönséget is megszólítjuk. A szótári rész segítségével az ige-névszó, a kötött szavak szerinti mutató segítségével pedig a névszó-ige kollokációs kapcsolatokat is számba vehetjük, kideríthetjük, hogy az adott névszó mely igékkel milyen kifejezéseket alkot. Megtudjuk: mi adott igéhez a szokásos bővítmény ill. mi adott bővítményhez a szokásos ige.

A fordítói munka során kollokációs szótárként alkalmazható, és hasznos lehet olyan nyelvtanároknak, kutatóknak is, akik magyar nyelvtanítási célú tananyagot készítenek, magyar mint idegen nyelv oktatása és az anyanyelvi nevelés terén egyaránt.

A fentieken túl kiemelendő, hogy szótárunk révén a haladó magyarul tanulók egyfajta speciális tanulói szótárt kapnak a kezükbe, mely a legkülönbözőbb típusú gyakori igei szerkezetek bemutatása révén elősegíti az „idiomatikus", a magyar nyelvre jellemző nyelvhasználatot, a nemcsak nyelvtanilag helyes, hanem magyarul *megszokott* kifejezésmódot (Hanks, 2008). Annak a döntésünknek, hogy nem csak idiomatikus szerkezeteket, hanem kompozicionális kollokációkat is közlünk, nagy előnye, hogy képet kapunk a névszók kollokációs viselkedéséről is, hogy adott szó mely igének a bővítménye szokott lenni.

Képzeljük magunkat egy magyarul tanuló angol anyanyelvű helyébe. Hogyan is mondjuk magyarul, hogy '*meet the requirements*'? Tegyük fel, hogy tudjuk, hogy a '*követelmény*' szót kell használni, de mi a hozzá illeszkedő ige? A válasz: '*megfelel*', és az is kiderül, hogy az angol tárgyas kifejezéstől eltérően a '*követelmény*' szót '*-nak/-nek*' raggal kell használnunk. Hasonlóan találhatjuk meg a '*make a contract*' kifejezés kapcsán a '*szerződés*' szóhoz a '*köt*' igét, kiegészítve azzal az információval, hogy a szerződéskötésben szereplő másik fél '*-val/-vel*' ragos bővítményként jelenhet meg a magyarban. Az '*ajándék*' szónál megtalálható 'kap AJÁNDÉK-bA -t' szerkezetből pedig azt tudhatjuk meg, hogy itt a '*-ba/-be*' ragot kell használnunk eltérően az angol '*as a gift*' formától. A jó szótár *ötletet ad* arra, hogy ténylegesen hogyan mondják az adott dolgot az adott nyelven (vö: 107. oldal). Ennek megvalósításához a jelen szótár készítésekor alkalmazott korpuszvezérelt módszertan vihet legközelebb.

A szótárhasználatot kiegészítheti (az azonos szövegkorpuszra épülő) Mazsola (3.2. rész) korpuszlekérdező eszköz használata. Segítségével a magyar igék bővítményszer-

DOI:10.15774/PPKE.ITK.2011.004

kezetét vizsgálhatjuk, egyes bővítményi helyeken megjelenő jellegzetes szavakra tudunk rákérdezni; fontos azonban látni, hogy ez az eszköz *nem* tartalmazza a jellegzetes igei szerkezeteket összegyűjtő lépést. Ha az a kérdés, hogy mely szerkezetek tipikusak, akkor a szótárhoz kell fordulnunk.

4.2.7. A szótárkészítés költségigénye

Megközelítésünk lényegi pontja, hogy az automatikus szakaszban (ld. a 19. ábra első részét a 77. oldalon) alkalmazott nyelvtechnológiai eszközök jelentős mennyiségű manuális munkát váltanak ki, így a szükséges lexikográfiai munka volumene nem túl nagy. Jelen szótár esetében, mely nagyjából 2200 ige 6200 igei szerkezetét tartalmazza (vö: 12. táblázat; 81. oldal) a szótári munkálatok hozzávetőleges munkaigénye – a Magyar Nemzeti Szövegtárat adottnak véve – a következőképpen alakult:

nyelvtechnológiai eszközök megvalósítása, fejlesztése	1 emberév
lexikográfiai munka	1 emberév

Az automatikus és a manuális szakaszra fordított idő nagyjából megegyezik, a lexikográfiai munkán belül nagyjából fele-fele idő szükséges az első változat elkészítéséhez, illetve az ellenőrzéshez. Valóban igaz tehát, hogy az ismertetett módszerrel készülő szótár – illetve esetleg jövőben készülő hasonló szótárak – költségigénye alacsony.

A manuális munka eredményeként XML formában előálló szótári részből (21. ábra) kiindulva a mutatók generálása tisztán automatikusan történt, ill. a szótár mint könyv majdnem nyomdakész – tördelés előtti – állapotú előállítása a LaTeX 2_ε szövegszedő rendszerrel szintén automatikusan valósult meg.

4.2.8. Összefoglalás

Ebben a fejezetben *Magyar igei szerkezetek* (Sass et al., 2010a) szótárt mutattuk be. A szótár a dolgozat gerincét képező jellegzetes igei szerkezeteket kinyerő algoritmus (3.3. rész) legfontosabb gyakorlati alkalmazása, a kutatás gyakorlati kicsatolása.

Szótárunk a leggyakoribb magyar igei szerkezeteket tartalmazza. Egynyelvű szótár explicit szótári értelmezések nélkül; a szerkezeteket, azok jelentését autentikus, korpuszból származó példák illusztrálják. Egyrészt *vonzatkeret*-szótár és *kollokációs* szótár egyszerre: az igék legjellemzőbb vonzatkereteit és legjellemzőbb névszói szókapcsolatait is tartalmazza, illetve azokat a szerkezeteket is, melyekben e két aspektus kombinációja jelenik meg (vö: 23. oldal). Másrészt *gyakorisági* szótár: kvantitatív információt szolgáltat a szerkezetek gyakorisági viszonyairól. Harmadrészt *összehasonlító* szótár: lehetőséget ad a magyar igei szerkezetek többszempontú összevetésére, a közöttük lévő kapcsolatok feltárására.

A szótár lexikográfiai szempontból több újdonságot tartalmaz. Alapegységei nem szavak, hanem *szószerkezetek*; az anyaggyűjtés korpuszvezérelt módon, *automatikusan* történik, a nyers szócikkek a lexikográfus nyelvi intuíciójától függetlenül automatikusan

DOI:10.15774/PPKE.ITK.2011.004

állnak elő; autentikus *korpuszpéldák* világítják meg a szerkezetek jelentését; valamint *gyakoriségi* mérőszámot is rendel a szerkezetekhez. A fentiek jól illeszkednek a modern szótárkészítési trendekbe, melyek szerint a szavak helyett egyre inkább a többszavas kifejezéseket állítjuk a középpontba, és különféle automatikus eljárásokkal próbáljuk csökkenteni a lexikográfia „rabszolgamunka" részét, így téve gyorsabbá és olcsóbbá a szótárkészítést (vö: 1.1. rész a 11. oldalon). A szótár a nagyfokú automatizáltság miatt viszonylag alacsony költséggel, gyorsan előállítható, hasznos segédeszköz lehet a nyelvészet számos területén és a nyelvoktatásban.

A szótár jelentőségét méltató **5. tézist** a 113. oldalon fogalmazom meg.

DOI:10.15774/PPKE.ITK.2011.004

DOI:10.15774/PPKE.ITK.2011.004

5. fejezet

Kiterjesztések

A 3.3. részben bemutattam a jellegzetes igei szerkezeteket kinyerő eljárást, amit aztán a gyakorlatban is alkalmaztam egy speciális szótár készítése során (4.2. rész). Most a módszer – és a mögötte rejlő modell (2.1. rész) – különféle, szerteágazó kiterjesztési lehetőségeit mutatom be, felvillantom, hogy a modell kínálkozó általánosítása mi mindenre teszi még alkalmassá ezt a megközelítést. Ebben a fejezetben befejezett, publikált eredmények, és folyamatban lévő kutatás is helyet kap.

5.1. Nyelvfüggetlenség

Az automatikus eljárásoknak külön jelentőséget ad, ha nyelvfüggetlenek. Ilyenkor kis munkabefektetéssel lehet egyéb nyelvekre az eredetihez hasonló eredményeket elérni segítségükkel. Az olvasóban talán már a modell leírása során felmerült, hogy az ismertetett, függőségi nyelvtanon alapuló modell valójában nem magyar-specifikus, kismértékű változtatással számos más nyelvre is alkalmazható lehet, azaz a megközelítésünk nyelvfüggetlen.

E fejezetben bemutatom, hogy módszer valóban nyelvfüggetlen, azaz számos nyelvre elő tudjuk állítani a modellnek megfelelő reprezentációt, és az adott nyelvű Mazsolát, valamint a kinyerő algoritmus futtatása után egy adott nyelvű igei szerkezeteket tartalmazó szótár előállítása is lehetséges. Azon túl, hogy hogyan hajtható végre az egész folyamat, azt vizsgáltam, hogy a létrehozott produktumok (a megfelelő nyelvű Mazsola korpuszlekérdező és a megfelelő nyelvű igeiszerkezet-szótár), ugyanolyan jellemzőkkel bírnak-e, és ugyanazokra a célokra használhatók-e fel, mint az eredeti magyar nyelvűek.

Az alkalmazott automatikus eszközök két részre oszthatók. A klasszikus nyelvelemző eszközök – a morfológiai elemző és egyértelműsítő, a tagmondatra bontó (2.2.1. rész), és a szintaktikai elemző (2.2.2. rész) – nyilvánvalóan nyelvfüggőek. Ezek azonban sok nyelvre már elkészültek, illetve várható, hogy az alapvető nyelvtechnológiai eszközkészlet részeként néhány éven belül számos nyelvre rendelkezésre fognak állni. A szintaktikailag elemzett korpuszra épülő további automatikus eszközökről pedig –

DOI:10.15774/PPKE.ITK.2011.004

kiemelendő a jellegzetes igei szerkezeteket gyűjtő algoritmus (3.3.1. rész) és a példagyűjtésben is használt korpuszlekérdező eszköz (3.2. rész) – az alábbiakban mutatjuk meg a nyelvfüggetlenséget.

A nyelvfüggetlenség tesztelésekor 4 nyelven: dán, szerb, francia és holland nyelven végeztem kísérleteket. A dán nyelv esetében végeztem részletes vizsgálatot (Sass, 2009d), a többi nyelvből való példák főként illusztrációként szolgálnak. A dán nyelvre vonatkozó vizsgálatban csak a Mazsola kialakításáig haladtam (ld. a 77. oldalon található 19. ábra felső részét: előfeldolgozás ill. a korpuszlekérdező eszköz), mivel itt valójában a reprezentáció kialakítása a kulcskérdés. Az előfeldolgozás után következő lépések már nem tartalmaznak nyelvfüggő elemeket, azaz ha birtokunkban van a reprezentáció, akkor az várható, hogy a csak a reprezentációtól függő további lépések, már nyelvtől függetlenül ugyanolyan módon fognak működni. A jellegzetes igei szerkezeteket gyűjtő algoritmus működésének egyetlen feltétele, hogy ilyen reprezentációjú bemenő korpuszt kapjon. Arra pedig, hogy a reprezentáció megfelelőségét vizsgáljuk, éppen a Mazsola korpuszlekérdező alkalmas: a kérdés az, hogy a kialakított dán nyelvű Mazsola ugyanazokat a tulajdonságokat mutatja-e, mint az eredeti magyar nyelvű változat.

Azért esett a választásom a dán nyelvre, mert szerkezete nagy mértékben eltér a magyartól. Ami nekünk most a legfontosabb, hogy a dán nyelvben másfajta nyelvi eszközöket használnak a bővítmények és az ige közötti viszony jelölésére. Egyszerűen fogalmazva: míg a dánban kötött a szórend és elöljárókat használ, addig a magyar szórendje szabadabb és gazdag esetrag-rendszerrel bír.

5.1.1. Modell és a reprezentáció megvalósítása

Nézzük a korábbi egyszerű magyar példánkat. Az '*A lány vállat vont.*' mondat reprezentációja a következő volt: '`ige=`von `-t=`váll `-0=`lány' (ld. a 4. ábrát a 30. oldalon).

Most kialakítjuk a dán '*26 personer kom på hospitalet.*' (26 ember került kórházba.) mondat reprezentációját. A modell (29. oldal) szerint a tagmondat bővítményeit egy tartalmi elem + viszonyjelölő pár reprezentálja.

A magyar és a dán mondat hasonló szerkezetű abban az értelemben, hogy ugyanúgy két bővítményt tartalmaz. A két nyelv számunkra érdekes szerkezeti különbsége – amint említettük – a viszonyjelölők milyenségében rejlik. A dán alanyt a mondatbeli sorrendi helye jelöli ki, a határozói bővítményt pedig egy elöljárószó; a magyarban mindkét bővítményi helyet esetrag jelöli ki.

Megtehetjük, hogy a viszonyjelölőket a nyelv tulajdonságainak megfelelően választjuk meg, így a magyar esetragok helyett a dánban az elöljárószókat fogjuk viszonyjelölőnek venni. Ezen kívül a dán alany és tárgy esetében egy speciális viszonyjelölővel dolgozunk: ez a *sorrendi megkötés*. Azt, hogy mi alany és mi tárgy – egyéb indueurópai nyelvekhez hasonlóan – a dánban is a sorrend dönti el: ige előtt alany, ige után tárgy lesz a jelöletlen főnévi (névszói) csoport. Ennek megfelelően a dánban viszonyjelölő lesz minden elöljárószó (*i*, *til*, *på* stb.) valamint az absztrakt `subj` (alany) és `dobj` (tárgy), melyeket sorrendi megkötés határoz meg.

DOI:10.15774/PPKE.ITK.2011.004

Ezek alapján a fenti példamondat reprezentációja a következő lesz:

'26 personer kom på hospitalet.'
'ige=komme subj=person på=hospital'

Nem meglepő, hogy a modell nyelvfüggetlen, hiszen lényegében csak arra támaszkodik, hogy van predikátum–argumentum struktúra a nyelvekben, azaz hogy vannak egy igéből és az ige bővítményeinek összességéből álló természetes egységek, és hogy az ige és adott bővítmény közötti (felszíni/szintaktikai) viszony valamilyen módon megragadható, leírható. A viszonyjelölőket egységesen kezelve a modell elvonatkoztat a konkrét nyelvspecifikus viszonyjelölők formai tulajdonságaitól, legyenek azok önálló szavak (pl.: dán elöljáró), kötött morfémák (pl.: magyar esetragok) vagy akár sorrendi megkötések.

Ahhoz, hogy előállíthassuk a reprezentációt, egy alkalmas dán korpuszból ki kell nyernünk a tagmondatokat, azonosítani kell az igéket és a bővítményeket, a tartalmi elemeket és a viszonyjelölőket.

Két lehetőség van. Egyrészt – a magyarhoz hasonlóan – megtehetjük, hogy egy morfológiailag elemzett korpuszból indulunk ki, és kifejlesztjük a 2.2. részben leírt szükséges nyelvfeldolgozó modulokat. Másrészt kiindulhatunk egy treebank-ból (szintaktikailag elemzett korpuszból), ekkor a feladat a reprezentációhoz szükséges információ kinyerése az általában gazdag annotációból. Ehhez a kísérlethez a második – kényelmesebb – lehetőséget választottam. Bár a szintaktikailag elemzett korpuszok általában két nagyságrenddel is kisebbek mint a csak morfológiailag elemzettek, az itt felvázolt tesztelési célra megfelelő ez a korpuszméret is. A választott korpusz a szabadon hozzáférhető és jól dokumentált, 300000 szavas Danish Dependency Treebank (dán függőségileg elemzett korpusz) (Trautner Kromann, 2003). A korpusz feldolgozása során a treebank szintaktikai fáit bejárjuk és a megfelelő relációkat feldolgozzuk, így a reprezentációhoz szükséges információ kinyerhető. A feldolgozás technikai részletei (Sass, 2009d) 264. oldalán olvashatók.

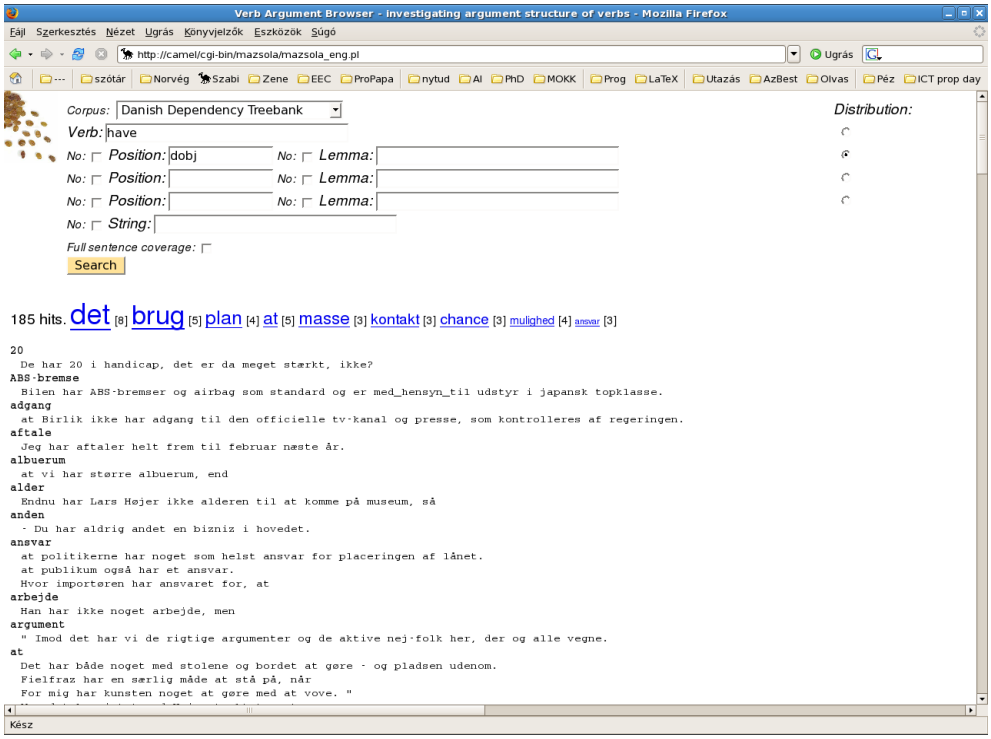
Ennek az alfejezetnek az üzenete az, hogy valóban kialakítható a reprezentáció dán nyelvre is. Ez egy viszonylag nyilvánvaló eljárás – meg kell határozni a viszonyjelölőket, majd egy korpuszt a megfelelő formára kell hozni –, a jelentősége viszont annál nagyobb, mivel mindössze a reprezentációra van szükség ahhoz, hogy a rá épülő további lépések automatikusan működhessenek.

5.1.2. Dán nyelvű Mazsola

A dán függőségileg elemzett korpusz alapján elkészítettük a dán nyelvű Mazsolát, mely szintén szabadon elérhető a `http://corpus.nytud.hu/vabd` címen (ideiglenes felhasználói név: `vendeg`; jelszó: `mazsola`). Az eredeti magyar változathoz (9. ábra a 49. oldalon) mindenben hasonló lekérdezőfelület és válaszképernyő a 22. ábrán látható.

DOI:10.15774/PPKE.ITK.2011.004

5. Kiterjesztések



22. ábra. A dán Mazsola válaszképernyője. A lekérdezőfelület alatt a ‘have `dobj`’ (‘birtokol vmit’) bővítménykeret jellegzetes tárgyait látjuk: ‘*brug*’ (használat), ‘*plan*’ (terv), ‘*masse*’ (tömeg), ‘*kontakt*’ (kapcsolat), ‘*chance*’ (esély), ‘*mulighed*’ (lehetőség). A sorrendi megkötés által meghatározott tárgy (direct object) viszonyjelölője a `dobj` kód.

Amint az 50. oldalon említettük, a Mazsola kétféle jellegzetes bővítményi kollokátum kinyerésére alkalmas: gyakori szavak „szó szerinti” jelentésben, valamint az igével együtt idiomatikus jelentésű komplex igét alkotó szavak.

Látjuk, hogy a Mazsolának ez a képessége dán nyelven is ugyanúgy működik, ilyen kicsi korpuszméret mellett is. Az első csoportot mindössze egy szó képviseli: a ‘*plan*’ (terv). Viszont már ebben a kis példában számos – a második csoportba tartozó – komplex igével találkozunk (13. táblázat).

További (gyakori) ige + elöljáró kombinációkat lekérdezve hasonló (vonzatos) komplex igéket kapunk (14. táblázat).

A fentiekhez hasonló vizsgálatot a szerb nyelvre is elvégeztem. Itt csak egy példát közlök illusztrációképpen, mely jól mutatja a szerb nyelvű Mazsola komplex igéket kinyerő képességét (23. ábra). Az ‘ići u·ŠKOLA’ (‘megy ISKOLA-bA’) és az ‘ići u·PRAVAC’ (‘megy IRÁNY-bA’) nyilvánvalóan „szó szerinti” jelentésben illik ebbe a keretbe. Az ‘ići u·PRILOG’ (szó szerint: ‘megy HASZON-bA’) viszont más eset, itt egy valódi komplex igével van dolgunk melynek a jelentése egyébként: ‘támogat’.

A fentiek alapján az várható, hogy a dán igei szerkezetek szótárát is el lehet készíteni a magyarra kidolgozott módszer szerint. Egy tényleges szótárhoz a használt korpusz nem biztosít elég anyagot, arra azonban alkalmas, hogy néhány mintaszócikken bemutassuk, hogy hogyan is nézne ki egy ilyen szótár.

DOI:10.15774/PPKE.ITK.2011.004

13. táblázat. A ‘have `dobj`’ (‘birtokol vmit’) bővítménykeret kapcsán kinyert komplex igék. Látjuk, hogy mindegyik példa egyben vonzatos komplex ige. A vonzatokat természetesen nem automatikusan, hanem a korpuszpéldák kézi vizsgálatával állapítottuk meg. (Az összetett – két szóból álló, de egy bővítményt kijelölő – ‘for·at’ elöljáró egybetartozását jelöljük a ponttal.)

kollokátum	dán komplex ige	magyar megfelelő
‘*brug*’	‘have BRUG for’	‘van -rA SZÜKSÉG-A’
‘*masse*’	‘have MASSE av’	‘van -bÓl TÖMEG-A-vAl’
‘*kontakt*’	‘have KONTAKT med’	‘van KAPCSOLAT-bAn -vAl’
‘*chance*’	‘have CHANCE for·at’	‘van -rA ESÉLY-A’
‘*mulighed*’	‘have MULIGHED for·at’	‘van -rA LEHETŐSÉG-A’

14. táblázat. Egyéb bővítménykeretek kapcsán kinyert dán komplex igék. A vonzatokat a 13. táblázathoz hasonlóan a korpuszpéldák kézi vizsgálatával állapítottuk meg.

keret	kollokátum	dán komplex ige	magyar megfelelő
‘være i’			
	‘*tvivl*’	‘være i·TVIVL om’	‘van ·felől KÉTSÉG-A’
	‘*forbindelse*’	‘være i·FORBINDELSE med’	‘van KAPCSOLAT-bAn -vAl’
‘være på’			
	‘*vej*’	‘være på·VEJ’	‘van ÚT-n’
	‘*besøg*’	‘være på·BESØG’	‘van LÁTOGATÓ-bAn’
‘få `dobj`’			
	‘*lov*’	‘få LOV til’	‘kap -rA ENGEDÉLY-t’

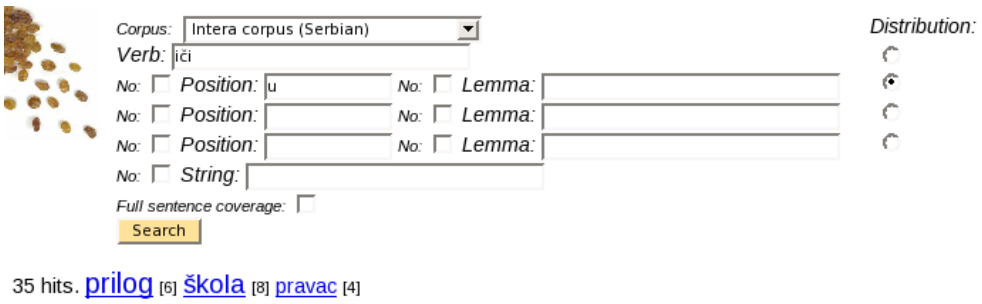
A szerkezeteket gyűjtő algoritmus lefuttatásakor a korpusz kis mérete miatt 5 helyett 2-es küszöböt alkalmaztunk (ld. a 3. lépést az 59. oldalon). Az eredményben azt tapasztaljuk, hogy bár komplex igék (ld. 13. és 14. táblázat) a kis korpuszméret miatt nem jönnek ki, a 24. ábrán látható két nyers szócikk megfelel az elvárásoknak.

5.1.3. Összefoglalás

A magyartól jelentősen különböző szerkezetű dán nyelv példáján megmutattam, hogy a dolgozat korábbi részeiben részletezett megközelítésem nyelvfüggetlen. A nyelvfüggetlenség demonstrálására egy dán nyelvű treebank-ből előállítottam az egységes reprezentációt. Az reprezentáció itt a lényegi pont, ha ezt – szükségképpen nyelvfüggő nyelvelemző eszközökkel – létrehoztuk, akkor a megfelelő nyelvű Mazsola korpuszlekérdező valamint a megfelelő nyelvű igei szerkezetek szótára szinte „gombnyomásra” áll elő.

Láttuk, hogy a dán Mazsola ugyanazokkal a hasznos tulajdonságokkal bír, mint az

DOI:10.15774/PPKE.ITK.2011.004



23. ábra. Egy példa a szerb Mazsolából. Az ‘ići u’ (‘megy -bA’) bővítménykeret jellegzetes kollokátumait látjuk: ‘*prilog*’ (haszon), ‘*škola*’ (iskola), ‘*pravac*’ (irány).

se
se [28] (‘néz’)
se på [9] (‘ránéz -rA’)

komme
komme [21] (‘jön’)
komme til [11] (‘jön -bA’)
komme i [11] (‘jön -bAn’)
komme på [9] (‘jön -rA’)
komme til·at [8] (‘fog csinálni vmit’)

24. ábra. Két automatikusan előállított, dán nyelvű, nyers szócikk. A kis korpuszméret ellenére a legjellegzetesebb szerkezetek helyesen megjelennek.

eredeti magyar változat: alkalmas a dán nyelv komplex igéinek és egyéb fontos igei szerkezeteinek összegyűjtésére. Így hasonlóan alkalmas korpuszvezérelt lexikográfiai munkálatok segédeszközéül, valamint a (korpuszvezérelt) nyelvoktatásban is alkalmazható.

Megmutattuk, hogy ez a reprezentáció várhatóan a nyelvek széles körére előállítható, mert a nyelveknek csak azt az alapvető tulajdonságát használja ki, hogy van benne prédikátum-argumentum struktúra. Módszerünk alkalmazásának feltétele tagmondatokra bontott, szintaktikailag megfelelően elemzett korpusz, *vagy* az ennek előállításához szükséges morfológiai elemző, tagmondatra bontó és szintaktikai elemző modul megléte.

A nyelvfüggetlenséget kimondó **6. tézis** a 113. oldalon található.

5.2. A modell általánosítása

5.2.1. Sorrendi megkötés mint viszonyjelölő

A dolgozat legnagyobb részében a magyar nyelvvel foglalkoztunk, és viszonyjelölő alatt a magyar esetragokat (és névutókat) értettük. Ha visszaidézzük az eredeti 7. definíciót (28. oldal), látjuk, hogy az ennél jóval általánosabban fogalmaz, és az 5.1. részben láttunk példákat egyéb viszonyjelölőkre. Tekintsünk át néhány lehetséges viszony-

DOI:10.15774/PPKE.ITK.2011.004

jelölő-típust (15. táblázat).

15. táblázat. Néhány viszonyjelölő-típus.

viszonyjelölő	példa
esetrag	magyar (‘*-bAn*’, ‘*-tÓl*’)
névutó	magyar (‘*alá*’, ‘*mögött*’)
elöljáró	angol (‘*into*’, ‘*for*’), dán (‘*til*’, ‘*på*’)
névutó + esetrag	magyar (‘*-n kívül*’)
elöljáró + esetrag	német (‘*zu*’ + részes), szerb (‘*u*’ + tárgy)
sorrendi megkötés	angol és dán (alany és tárgy)

Az nyilvánvaló, hogy bizonyos nyelvek ugyanúgy használnak elöljárószókat a mondatbeli szintaktikai helyek kijelölésére, mint ahogy mások ugyanerre a célra esetragokat vagy névutókat, vagy ezek kombinációit, ahogy ezt a magyar, a német és a szerb esetében láttuk. Ezek a látható viszonyjelölők általában lehetővé teszik, hogy az általuk megadott bővítmények eléggé szabad sorrendben helyezkedjenek el a mondatban.

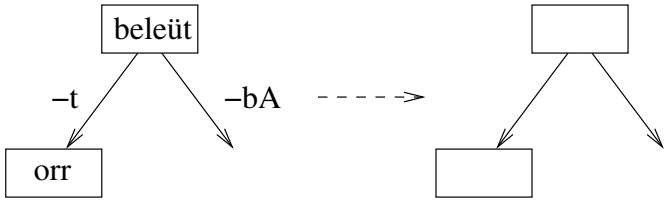
Bizonyos SVO nyelvekre jellemző, hogy az alanyt és a tárgyat morfológiailag semmi nem jelzi, csak a mondatbeli sorrendi helye: az ige előtti főnévi (névszói) csoport alanyként, az ige utáni tárgyként értelmeződik (vö: ‘*Et barn ser en voksen.*’ (Egy gyerek néz egy felnőttet.) vs. ‘*En voksen ser et barn.*’ (Egy felnőtt néz egy gyereket.)) Fogalmazhatunk úgy is, hogy itt egy speciális viszonyjelölővel van dolgunk: a nyelvi egységek *sorrendje* hordozza a szintaktikai szerepükre vonatkozó információt. A *sorrendi megkötés*t is jogosan tekintjük tehát viszonyjelölőnek – amint ezt a 90. oldalon meg is tettük –, függetlenül attól, hogy felszíni alakja nincs, ti. ez is egy olyan nyelvi elem, mely az ige és a bővítmény közötti szintaktikai/felszíni viszonyt jelöli (Megjegyzendő, hogy ezzel a többmorfémás kifejezések (ld. 21. oldal) fogalmát is kiterjesztettük olyan módon, hogy nyelvi elemként, „morfémaként” most már olyan elemek is alkothatják ezeket a kifejezéseket, melyeknek egyáltalán nincs felszíni alakja.)

5.2.2. A modell absztrakt leírása

Már a 29. oldalon említettük, hogy a modell szerinti reprezentáció 1-mélységű függőségi szerkezeteket tartalmaz, és mondhatjuk (61. oldal), hogy a jellegzetes igei szerkezeteket kinyerő algoritmus pedig ilyen 1-mélységű – lexikálisan megfelelően kitöltött vagy kitöltetlen – függőségi szerkezeteket nyer ki.

Ebből a megfogalmazásból adódik a következő általánosítási lehetőség. Ha az adatszerkezetet (reprezentációt) – elvonatkoztatva az eddig viszonyjelölőktől és tartalmi elemektől – kiterjesztjük az 1-mélységű fák általános osztályára, akkor várhatóan a kinyerő algoritmus erre a struktúrára is ugyanúgy alkalmazható lesz: az új reprezentációnak megfelelő jellegzetes szerkezeteket fog kinyerni. Az általánosítás miként a 25. ábrán látható.

DOI:10.15774/PPKE.ITK.2011.004



25. ábra. A reprezentáció általánosítása. A 2. ábrán (21. oldal) látott első szerkezet függőségi fája, és a neki megfelelő absztrakt modell szerinti struktúra, absztrakt függőségi fa (irányított gráf). A korábbi viszonyjelölőktől és tartalmi elemektől elvonatkoztatva csak az 1-mélységű fa struktúrát tartjuk meg, mint általános keretet. A korábbiakhoz hasonlóan továbbra is fontos, hogy bizonyos szerkezeteknek része másoknak pedig nem része a tartalmi elem, ahogy ezt az alsó szinten megjelenő téglalap illetve annak hiánya mutatja.

A továbbiakban a 25. ábrán látható szerkezetek részeinek megjelölésére az alapvető gráfelméleti fogalmakat is fogjuk használni, azaz ige helyett *gyökér*, viszonyjelölő helyett *él* vagy *címke*, tartalmi elem helyett pedig *csomópont* fog szerepelni. A bővítmény helyett a *jegy* szót fogjuk használni, ez a korábbiakhoz hasonlóan utalhat egy élre, vagy egy él és egy csomópont együttesére is. A 31. oldalon bevezetett LSzB és LKB fogalmaknak az *LSzJ (lexikálisan szabad jegy)*, és *LKJ (lexikálisan kötött jegy)* felel meg. Az előbbi jelenti az egy élt, az utóbbi pedig az él és csomópont együttesét.

Az általánosítás lényege az, hogy a gráfstruktúrát megőrizzük, de a konkrét élek és csomópontok tekintetében mostantól teljesen szabad kezet adunk. Mindössze annyi tehát a követelmény, hogy csak olyan entitásokat próbáljunk ebben a modellben reprezentálni, amelyek rendelkeznek az alábbi tulajdonságokkal: van bennük egy központi elem, ehhez alárendelt elemek kapcsolónak, a központi elem és egy alárendelt elem mindig valamilyen meghatározható viszonyban van, és bizonyos estekben csak a viszony érdekes, a konkrét alárendelt elem nem.

Ezt a modellt tekinthetjük egy lapos adatbázis-szerkezetnek is: az entitásokat olyan speciális jegyekkel írjuk le, melyeknél egyrészt érdekes, hogy adott példánynál a jegy megjelenik-e; ha pedig megjelenik, akkor két lehetőség van: vagy számít a konkrét értéke (ekkor természetesen fontos, hogy mi az), vagy pedig nem, ilyenkor csak az számít, hogy megjelenik a jegy.

Az általánosításnak az az értelme, hogy reményeink szerint a kinyerő algoritmus ugyanúgy fog dolgozni erre a struktúrára is, azaz ilyenfajta jellegzetes szerkezeteket fog kinyerni, következésképpen a szótárnak megfelelő adatbázis is előállítható lesz.

A modell fenti általánosításával azt engedjük meg, hogy *bármilyen* az absztrakt modellnek megfelelő viszonyokkal *bármilyen mértékben* annotált korpuszból kiindulhatunk, és e korpuszból kiindulva olyanfajta jellegzetes szerkezeteket tudunk kinyerni az algoritmussal, amilyen a konkrét reprezentáció meghatároz.

A további két fejezetben látunk példákat arra, hogy milyen különféle módokon lehet ezt az absztrakt modellt alkalmazni.

DOI:10.15774/PPKE.ITK.2011.004

5.3. Példák az absztrakt modell alkalmazására

Ebben a részben publikációval még alá nem támasztott folyamatban lévő kísérleteket mutatok be, melyek érzékeltetik az absztrakt modellben rejlő lehetőségeket.

5.3.1. Új bővítménytípusok

Névszói csoporttól különböző bővítmény

Elsőként megemlítjük a legegyszerűbb alkalmazást: a nem névszói csoportként megjelenő bővítmények kezelését.

A bővítmény definíciójakor (27. oldal) említettük, hogy a modell általánosításával tetszőleges bővítmény kezelhető lesz, nem csak a névszói csoport bővítmények. Az absztrakt modell egy közvetlen alkalmazása lehet, hogy a segédige–ige viszony kezelése. A '*10 hrivnyát kell leperkálni kilójáért.*' mondatban például a két igének megfelelően két (egymásba ágyazott) igei szerkezet van. A 'leperkál -ért -t' megfelel a már jól ismert eredeti modellnek, a másik szerkezet pedig a '`ige=`kell `fni=`leperkál' lesz, ha éppen `fni`-vel jelöljük a segédige-ige közötti bővítményi viszonyt. Az újdonság éppen ennek a bővítményi viszonynak a kezelése.

Ide tartozik a határozószók kezelése is: a jegyek közé felvehetjük a „határozószó" jegyet is, következésképpen megkapjuk azokat a szerkezeteket is, melyekben tipikus határozószó szerepel (pl.: '`ige=`akad `adv=`mindig', '`ige=`él `adv=`együtt', '`ige=`eltűnik `adv=`szinte').

Szemantikai információ használata

A dolgozatban végig felszíni – ti. a felszínen is megjelenő, morfológiailag látható – jegyekkel foglalkoztunk. Ezek a jegyek vannak a legtöbb korpuszban megjelölve, és a Magyar Nemzeti Szövegtárban is ezek álltak rendelkezésre, és ezek használatával is értékes nyelvi erőforrások készíthetők. Az előfeldolgozás során a bővítményeket e felszíni jegyek (esetragok/névutók/elöljárók) alapján kapcsoltuk az igékhez, azaz nem végeztünk semmiféle szemantikai elemzést, és a felhasznált korpuszok sem tartalmaztak szemantikai információt. Emiatt fordul elő, hogy a szótárban a 'lakik VHOL' szerkezet helyett a (gyakoribb) 'lakik -bAn' és a (sokkal ritkább) 'lakik -n' jelenik meg, illetve a fordított eset, mikor egy esetrag szempontjából egységes bővítmény számos különböző jelentést fed le, pl. 'nyer -vAl': '*pontozással*', '*lelkesedéssel*', '*kisgazdákkal*'. Az igei szerkezeteket olyan mértékig tudtuk elkülöníteni, amennyire a felszíni jegyekre épülő megközelítés ezt lehetővé teszi.

Nagy lehetőségeket rejt magában a szemantikai annotációt tartalmazó korpuszok felhasználása, illetve a korpuszok szemantikai információval való felcímkézése az előfeldolgozás során. Ide tartozna például a hely-, idő- és módhatározók automatikus felismerése, és bővítményi kategóriaként való kezelése, ami által a fenti 'lakik VHOL'

DOI:10.15774/PPKE.ITK.2011.004

probléma oldódna meg; valamint a különböző szemantikai kategóriák kezelése, és ezáltal szemantikus alapú szerkezetek (pl. ‘vág ÉLŐ-hOz ÉLETTELEN-t’) azonosítása.

5.3.2. Nem ige-központú szerkezetek

Valóban az ige a mondat központi eleme, de természetesen vannak kisebb egységek, melyek még mindig bonyolult belső szerkezettel rendelkeznek, és érdekes lehet a jellegzetes ilyen szerkezetek feltérképezése is. Adódó példa a főnévi csoport: ebben az alfejezetben tehát a központi elem nem az ige, hanem a főnév lesz.

Egy kísérletben a főnévi csoport jellemzőiként a következő jegyeket határoztam meg: jelző, főnév esete, főnév száma, főnév birtokos személyragja. Az egyes jegyek alapértelmezett értéke rendre: nincs jelző, alanyeset, egyesszám, nincs birtokos személyrag. Alapértelmezett érték esetén úgy tekintjük, hogy „a jegy nem jelenik meg”, éppen úgy ahogy a ‘-tÓl’ jegy nem jelent meg egy igei szerkezetben, ahol nem volt ‘*-tól/-től*’ ragos bővítmény. A ‘*kóbor kutyák*’ szerkezet reprezentációja tehát a következő: ‘`fn`=kutya `jelző`=kóbor `szám`=többes’ Az ilyen módon kialakított főnévicsoport-reprezentációra futtatva a kinyerő algoritmust a 16. táblázatban látható eredményeket kapjuk.

16. táblázat. Néhány automatikusan kinyert, jellegzetes főnévi szerkezet. A felsorolásban jópár idiomatikus értelmű szerkezetet találunk (‘*belső fül*’, ‘*szabad szemmel*’), és van sok olyan is, melyek egy nagyobb idiomatikus egység részét képezi (pl.: ‘*gyenge lábakon áll*’, ‘*száraz lábbal kel át*’, ‘*süket fülekre talál*’, ‘*saját szemével lát*’). Úgy tűnik, hogy sikerült megragadni a jellemző eseteket, és az egyesszám/többessszám jelentőségét is.

láb	fül	szem
‘*lába*’	‘*belső fül*’	‘*szemmel*’
‘*lábon*’	‘*süket fülekre*’	‘*szabad szemmel*’
‘*lábak*’	‘*füllel*’	‘*mai szemmel*’
‘*saját lábán*’	‘*nagy füle*’	‘*jó szemmel*’
‘*száraz lábbal*’	‘*emberi fül*’	‘*szemek*’
‘*hátsó lába*’		‘*emberi szem*’
‘*hátsó lábai*’		‘*magyar szemmel*’
‘*gyenge lábakon*’		‘*saját szemével*’

A szokásos, jellegzetes névszói csoportok ilyen tára választ adhatna arra a fordítói munka során gyarkan felmerülő kérdésre, hogy vajon adott főnevet milyen jelzőkkel használunk, illetve adott jelző megszokott-e adott főnév mellett.

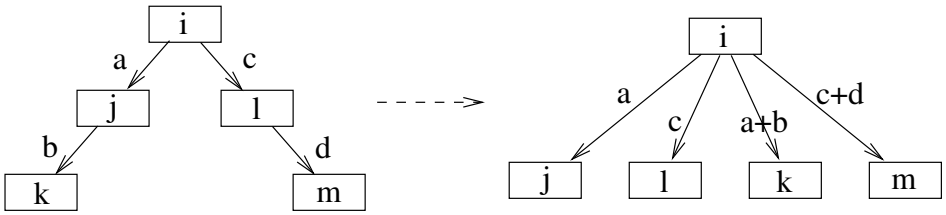
Figyeljük meg, hogy itt egészen másképp kezeljük az esetet, mint ahogy azt az igei szerkezeteknél tettük: a konkrét eset ott él volt, itt viszont csomópont. De erre az absztrakt modell lehetőséget nyújt, épp ez a rugalmasság a haszna. És fordítva: ahogy itt a főnév jellemzőit jegyként kezeltük, ugyanígy kezelhetnénk jegyként az igei szerkezetekben az ige különböző jellemzőit is (szám, személy, mód, idő), és akkor eredményül kaphatnánk olyan szerkezeteket, melyekre például az jellemző, hogy milyen időben vannak (‘*ez a hajó elment*’).

DOI:10.15774/PPKE.ITK.2011.004

5.3.3. Többszintű függőségi fák

A 16. táblázathoz fűzött megjegyzésben utaltunk rá, hogy a főnévi csoportokra kihegyezett módszerrel sokszor olyan főnévi csoportokat kapunk, melyek egy nagyobb (igei) szerkezet részét képezik (pl.: *'gyenge lábakon áll'*). Nyilván a legjobb lenne az egész igei szerkezetet megkapni a benne lévő főnévi csoporton belüli jellegzetességekkel együtt.

Másképp fogalmazva, sok igei szerkezetben nem csak a bővítmény esetragja és tartalmi eleme (a bővítmény névszói feje), hanem például a tartalmi elem jelzője, vagy száma stb. is jellegzetes. Az ilyen többszintű szerkezeteket többszintű függőségi szerkezetekket (n-mélységű függőségi fákkal) tudjuk kezelni a dolgozatban eddig mindvégig szereplő egyszintűek (1-mélységűek) helyett. A 26. ábrán láthatjuk, hogy hogyan vezethetjük vissza a kétszintű (vagy akár többszintű) függőségi szerkezeteket az egyszintűek esetére.



26. ábra. Kétszintű függőségi fa kezelése az absztrakt modellben. A két- vagy többszintű függőségi fákat *kisimítjuk*, azaz a gyökérhez minden csomópontot egy közvetlen éllel kapcsolunk hozzá, az él címkéje az eredeti többszintű szerkezetben a gyökértől a csomópontig vezető út élcímkéinek konkatenációja lesz. Így egy 1-mélységű struktúrát kapunk, amit az ismert eljárásokkal kezelhetünk.

A többszintű függőségi szerkezetek fenti kezelési módja lehetőséget ad arra, hogy tekintetbe vegyük a kinyerendő szerkezetek tetszőleges jegyét: igeidőt, igemódot, jellemző jelzőt, jellemző névelőt stb.

Ilyen kísérletet egy holland korpuszon (Macken et al., 2007) végeztem. Köszönet Héja Enikőnek, aki a teljes szintaktikai elemzést létrehozta az Alpino parser (Bouma et al., 2001) segítségével, és ez alapján kialakította a részletes reprezentációt. A kapott szerkezetek közül néhány a 17. táblázatban látható.

Létezik egy hasonló módszer, mely többvszavas kifejezéseket nyer ki függőségileg elemzett korpuszból (Martens és Vandeghinste, 2010). E cikk szerzői ragaszkodnak a klasszikus függőségi felfogáshoz, hogy ti. a csomópontokban felszíni elemek, szavak legyenek. A teljes függőségi elemzés használata miatt nagyon sok elemű, nagyon specifikus szerkezeteket kapnak, szemben az általam leírt megközelítéssel, ahol csak a fontosnak vélt, előre definiált jegyeket használjuk, és ezáltal valóban a jellegzetes szerkezeteket kapjuk meg.

Látjuk, hogy az absztrakt modell teljesen szabad kezet ad a tekintetben, hogy hogyan alakítjuk ki a reprezentációt. Mindig adaptálhatjuk azokhoz a szerkezetekhez, ame-

DOI:10.15774/PPKE.ITK.2011.004

17. táblázat. Néhány automatikusan kinyert, jellegzetes többszintű holland szerkezet és magyar megfelelője. A kétszintű élek két tagját a 26. ábrán látható jelölésnek megfelelően '+' jel kapcsolja össze. A 3. szerkezet a 97. oldalon említett főnévi igenév bővítményt is ('inf') példázza; a 4. szerkezet pedig egy szép idiomatikus vonzatos komplex ige, melynek magyar megfelelője egészen másképp hangzik, mint az eredeti.

1.	holland szerkezet	'ige=speel obj=rol obj+ADJ=belangrijk'
	magyar megfelelő	'*jelentős szerepet játszik*'
2.	holland szerkezet	'ige=bewaar in=verpakking in+ADJ=oorspronkelijk'
	magyar megfelelő	'*eredeti csomagolásban tárol*'
3.	holland szerkezet	'ige=kan subj=bloed_suiker_waarde subj+ADJ=hoog inf'
	magyar megfelelő	'*magas vércukorérték tud vmit csinálni*'
4.	holland szerkezet	'ige=breng tot=einde tot+ADJ=goed obj'
	magyar megfelelő	'*sikerre („jó befejezésig") visz vmit*'

lyekre éppen kíváncsiak vagyunk. Fontos, hogy hogyan alakítjuk ki a reprezentációt, nem biztos, hogy egy függőségileg elemzett korpuszban megtalálható összes információ szükséges a jellegzetes szerkezetek kinyeréséhez.

5.4. Párhuzamos igei szerkezetek kinyerése

Amint látni fogjuk, valójában ez is egy példa az absztrakt modell alkalmazására, de jelentősége folytán külön fejezetben tárgyaljuk. Alább a (Sass, 2010d) cikkben ismertetett eredményeket tekintjük át; egy alternatív megközelítés található a (Héja és Sass, 2010) cikkben.

Most a megismert, egynyelvű korpuszra kifejlesztett jellegzetes igei szerkezeteket kinyerő eljárást alkalmazzuk párhuzamos korpuszra, a korpusz-reprezentáció alkalmas átalakításával, kétnyelvű, párhuzamos igei szerkezetek kinyerése céljából. A nyelvtechnológiai alkalmazások (pl.: a gépi fordítás) lexikális erőforrásainak tartalmaznia, ismernie kell ezeket a kifejezéseket, hogy magas nyelvi minőségű kimenetet adhassanak. Ezek a szerkezetek ugyanakkor sok esetben más nyelvre lefordítva teljesen más formát mutatnak. Bár a szükséges elemző lépések során alkalmazott egyszerű közelítő módszerek, valamint a feladat nehézsége miatt a kinyerés pontossága nem kiemelkedő, jelen fejezetből világos lesz, hogy az algoritmus képes különféle, akár aszimmetrikus, párhuzamos szerkezetek feltérképezésére is.

Hasonló célt tűz ki egy korábbi munka (Bojar és Hajič, 2005). Szintén függőségileg elemzett párhuzamos korpuszon dolgoznak, és párhuzamos szerkezeteket nyernek ki, de az igei vonzatkeretekkel foglalkozó kutatási vonulatba illeszkedve az ő látóterükbe csak az igei vonzatkeretek, azaz a csak LSzB-ket tartalmazó igei szerkezetek kerülnek.

DOI:10.15774/PPKE.ITK.2011.004

A mi megközelítésünk középpontjában viszont – amint már megszokhattuk – éppen azok a szerkezetek állnak, melyek nem csak vonzatkeretek, hanem ugyanakkor többszavas kifejezések is. A '*kilátásba helyez vmit*', '*részt vesz vmiben*' típusú vonzatos komplex igékben *négy* egységet különíthetünk el: az igét, a vonzatot (magyarban esetrag képviseli), a komplex ige névszói elemét, valamint e névszói elem esetragját.

A nyelvekre általában jellemző, hogy a komplex igék névszói elemét és a vonzatot *ugyanazokkal* a nyelvi eszközökkel kapcsolják az igéhez, legyen az esetrag, névutó, elöljáró, igei partikula vagy akár sorrendi megkötés (ld. 5.2.1. rész). Emiatt ezek a „négyelemű kollokációk" speciális kezelést igényelnek: az őket megcélzó lexikai kinyerő eljárásnak fel kell ismernie, hogy az adott bővítményi elem lexikálisan kötött módon a komplex ige része-e ('*kilátásba*', '*részt*'), vagy pedig vonzat, mely esetben a konkrét szó nem része a szerkezetnek, csupán a viszonyjelölő ('*vmit*', '*vmiben*').

Nyilvánvalónak tűnik, hogy ezek a szerkezetek csak a vonzatukkal együtt teljesek, csak teljes formájukban tudnak hozzájárulni nyelvtechnológiai alkalmazások teljesítményének javításához, például tipikusan egy gépi fordítóban használt lexikai adatbázis elemeként. Mégis a korábbi kutatásokra jellemző, hogy elfogadják helyes eredménynek a hiányos szerkezeteket is. A kollokációkutatók sokszor megfeledkeztek arról, hogy a kollokációknak vonzatuk is lehet, amint ez az (Evert és Krenn, 2001) cikkben idézett '*zur Verfügung stellen*' (rendelkezésre bocsát) szerkezet esetében is kitűnik (vö: 19. oldal). Ebben a cikkben csak a elöljáró+főnév+ige típusú szerkezeteket vizsgálták, ennek megfelelően a fenti szerkezet inherens részét képező *tárgy* megtévesztő módon elmarad. Siepmann (Siepmann, 2005, 416. oldal) is hangsúlyozza: „az igei kollokációk és a vonzatok szorosan összefüggnek, számos ige+főnév kollokáció a vonzatok adott disztribúcióját kívánja meg . . . a vonzatuktól megfosztott ige+főnév kombinációk nem tekinthetők teljes értékű szerkezetnek".

Visszatérve a gépi fordítás példájára, gondolhatnánk, hogy a tárgy elmaradása nem is jelent nagy problémát, mert amit az egyik nyelv tárggyal fejez ki, azt „nyilván" a másik is ugyanúgy tárggyal jeleníti meg. Ez azonban egyáltalán *nem mindig igaz,* és még kevésbé igaz az egyéb esetragokra/elöljárókra, melyek változatos mintázatokban felelhetnek meg egymásnak két nyelv viszonylatában.

A jellegzetes igei szerkezeteket kinyerő eljárás segítségével eddig végig egynyelvű szerkezeteket nyertünk ki egynyelvű korpuszból. Egy gépi fordításban közvetlenül hasznosítható kétnyelvű lexikai adatbázis vagy szótár összeállításához azonban kétnyelvű, párhuzamos igei szerkezetekre van szükség. (Ezt természetesen emberi erővel is elő lehet állítani, az automatikusan kinyert szerkezetek kézi fordításával, amint ezt a 71. oldalon kezdődő 4.1.1. részben láttuk.) Most azt vizsgáljuk, hogy hogyan adaptálható a 3.3. részben leírt eljárás párhuzamos korpuszra. Azaz arra a feladatra, hogy bemenetként párhuzamos korpuszt dolgozzon fel, eredményként pedig párhuzamos igei szerkezeteket (igei szerkezeteket és a fordításukat) szolgáltasson. Mivel az algoritmus az igei szerkezetek teljes spektrumát lefedi, azt várjuk, hogy szükség esetén képes lesz párba állítani *különböző* felépítésű szerkezeteket is: képes lesz megragadni azokat az eseteket is, amikor az egyik nyelv egyszerű igét használ ugyanarra, amit a másik nyelv komplex ige segítségével ír körül.

DOI:10.15774/PPKE.ITK.2011.004

5.4.1. A módszer alkalmazása párhuzamos korpuszra

Jelen munkálathoz a Dutch Parallel Corpus (Holland Párhuzamos Korpusz) (Macken et al., 2007) francia-holland részét használtuk. Ez egy könnyen hozzáférhető, morfológiailag elemzett korpusz, mely 3,2 millió holland és 3,6 millió francia tokent tartalmaz. A nyelvválasztás lehetőséget ad arra, hogy az eredetileg magyar nyelvre használt algoritmus nyelvfüggetlenségét (ld. 5.1. rész) is újból alátámasszuk.

A korpusz feldolgozása során egyszerű eszközökkel elvégeztük az előfeldolgozási lépéseket (vö: 2.2. rész), az eredeti modell szerinti „hagyományos" reprezentációt hoztuk létre mindkét nyelvre. (Ez a reprezentáció sokkal egyszerűbb, mint az ugyanennak a korpusznak a holland részére kialakított részletes, többszintű fákat tartalmazó reprezentáció, amiről a 99. oldalon volt szó.)

Egyszerű, szabályalapú tagmondatra bontó módszerünk a következő szabályokat tartalmazta. A mondathatáron kívül tagmondathatárt jelentett a kötőszó, az alárendelt tagmondatot bevezető holland *te* ill. francia *pour*, a vonatkozó névmás és bizonyos írásjelek (vessző, kettőspont és pontosvessző) is, amennyiben a legutóbbi tagmondathatár óta szerepelt a mondatban ige (hasonlóan a 2.2.1. részben leírt a magyar nyelvű tagmondatra bontóhoz). A részleges szintaktikai elemzést szintén egyszerű szabályok használatával valósítottuk meg. A tagmondatokban lévő főnevek (illetve a reflexív igék miatt a holland *zich* és a francia *se*) lettek a bővítményi tartalmi elemek, az elöljárók pedig a viszonyjelölők. A francia '*à*' elöljáró + '*le*' névelő összevonásából keletkező '*au*' szócska szótövét a korpuszban lévő '*au*'-ról '*à*'-ra javítottuk, így egységesen kaptuk meg az összes '*à*' elöljárós bővítményt; hasonlóan jártunk el a '*de*' + '*le*' = '*du*' esetében is. Ha nem találtunk a fej előtt elöljárót, akkor az ige előtt alanyként, az ige után pedig tárgyként kezeltük a szóban forgó bővítményt.

Az így előállított két elemzett „félkorpuszból" a következő módon alakítottuk ki a kétnyelvű bemeneti korpuszt, azaz a szükséges párhuzamos reprezentációt:

1. Az igét tartalmazó tagmondatokat fordítási egységenként sorra egymáshoz rendeltük (a fordítási egység első holland tagmondatához a megfelelő fordítási egység első francia tagmondatát stb.). Ha a fordítási egység nem azonos számú tagmondatot tartalmazott, akkor a fennmaradó(ka)t figyelmen kívül hagytuk.

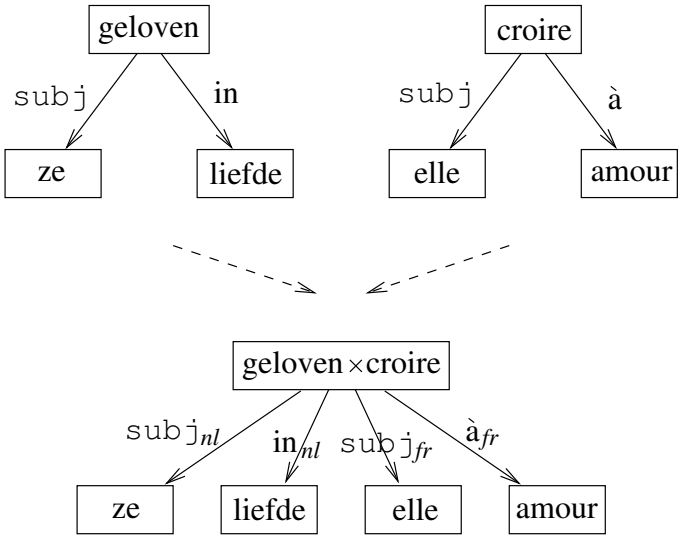
2. Az egymáshoz rendelt tagmondatok holland ill. francia igéjéből egy igepárt hoztunk létre (pl.: *gaan*×*aller* 'megy'), ez játssza majd az eredeti eljárásban az ige szerepét.

3. A tagmondatpárban található bővítményi csoportokat (mindkét nyelvűeket) egy halmazként soroltuk fel az igepár mellett, az egyes bővítményeket a megfelelő nyelv kódjával megjelölve.

A fenti lépések során egyfajta metakorpuszt alakítottunk tehát ki, mely párhuzamos tagmondatokból áll, a két eredeti tagmondat igéje egy metaigét alkot, a bővítmények pedig egy egyesített halmazként – de megtartva azt az információt, hogy eredetileg melyik nyelvből származnak – állnak a metaige mellett. A reprezentációt a 27. ábrán látható példa szemlélteti.

DOI:10.15774/PPKE.ITK.2011.004

holland tagmondat: *'Ze geloofde in de grote liefde'*
francia tagmondat: *'Elle croyait au grand amour'*
magyar fordítás: 'Hitt a nagy szerelemben'



reprezentáció:
'ige=geloven×croire $\texttt{in}_{nl}$=liefde $\texttt{à}_{fr}$=amour $\texttt{subj}_{nl}$=ze $\texttt{subj}_{fr}$=elle'

27. ábra. Példa a kétnyelvű bemeneti korpuszból. Az ábra felső részén a holland és a francia mondat látható a magyar fordítással együtt. Középen a két mondatnak megfelelő függőségi fa, illetve az ebből képzett párhuzamos reprezentáció függőségi fája látható. Alul a végső párhuzamos reprezentációt közlöm a szokásos formában. Az igepárt '×' jel kapcsolja össze, az elöljárókat alsóindex sorolja a megfelelő eredeti nyelvhez.

Ezek után az így kialakított kétnyelvű reprezentációra közvetlenül futtattuk az eredeti algoritmust. Mindössze két apróbb szükséges változtatást tettünk meg:

– Az algoritmus eredetileg két bővítményi helyet kezelt (ld. 2. lépés az 57. oldalon), ezt most *négyre* bővítettük, hogy megkaphassuk azokat a párhuzamos szerkezeteket is, melyben mindkét nyelvben 2-2 (tehát párhuzamos szerkezetenként összesen négy) lényeges bővítmény van.

– A három és négy pozíciót tartalmazó keretek közül a vonzatos komplex ige formájúak hosszához hozzáadtunk egy 0,2-es értéket. Így ezeknek a szerkezeteknek az esélyét megnöveltük, hogy az algoritmus 3. lépésében (59. oldal) a kiesőktől gyakoriságot örökölhessenek. E heurisztika hatására a végső listában több vonzatos komplex igét kaptunk.

DOI:10.15774/PPKE.ITK.2011.004

párhuzamos szerkezet:	‘geven×donner $\texttt{obj}_{nl}$ aan_{nl} $\texttt{obj}_{fr}$ à_{fr}’
holland szerkezet:	‘geven obj aan’
francia szerkezet:	‘donner obj à’
magyar megfelelő:	‘*ad vmit vkinek*’
párhuzamos szerkezet:	geloven×croire in_{nl} à_{fr}
holland szerkezet:	‘geloven in’
francia szerkezet:	‘croire à’
magyar megfelelő:	‘*hisz vmiben*’

28. ábra. Példák egyszerű vonzatot tartalmazó szerkezetekre. A párhuzamos szerkezetekből egyszerűen levezethetők a holland és francia szerkezetek, így a párhuzamos szerkezet közvetlenül megmutatja az adott igével használandó megfelelő elöljárót.

5.4.2. Kiértékelés

A bemeneti kétnyelvű metakorpuszban 20-szor vagy annál többször előforduló 1356 igepárra futtattuk az algoritmust. Számos egy vagy két egyszerű vonzatot tartalmazó szerkezet is került az eredménylistára (ld. 28. ábra), a kiértékelést csak a legizgalmasabb részre, a (leggyakoribb) vonzatos komplex igékre korlátoztam.

Összesen 67 olyan, legalább 15-ös gyakorisági értékkel bíró szerkezetet kaptunk, melyben vonzati pozíció és lexikálisan kötött bővítményi pozíció is volt. Az alábbi szempontok alapján fogadtam el egy párhuzamos szerkezetet helyesnek:

- Ami értelmes, az helyesnek számít, függetlenül attól, hogy idiomatikus-e a jelentése vagy sem. (Említettük (44. oldal), hogy például lexikográfiai szempontból fontosak lehetnek kompozicionális szerkezetek is.)
- A holland ‘*van*’ ill. francia ‘*de*’ általában az elemzés által egyáltalán nem kezelt birtokos szerkezetek miatt jelent meg. Ezeket nem vettük figyelembe, nem befolyásolták a szerkezetek helyességét.
- Az alany és a tárgy megállapítása nem tökéletes, ezért az alany és a tárgyat egymás helyett is elfogadtuk.
- Helyesnek fogadtuk el a szerkezetet akkor is, ha határozószó hiányzott belőle, mivel az elemzés nem kezelte a határozószókat.
- A hiányos szerkezetek nem jók, a helyességhez szükséges minden lényeges elem megléte (vö: a korábbi kiértékelési kritérium a 66. oldalon).

A fenti szempontok miatt 9 szerkezet egy másik szerkezettel egybeesett. A kapott 58 szerkezetnek a kiértékelése a 18. táblázatban látható.

Az eredmény természetesen jócskán elmarad a (Sass, 2009c) cikkben közölt, egynyelvű, magyar korpuszon $n = 50$ esetén mért 94 százalékos pontossági értéktől (vö: 8. táblázat alsó sora a 68. oldalon) Jelen feladat nyilvánvalóan jóval nehezebb: sokkal több elemet kell helyesen megtalálni, hogy a kapott párhuzamos szerkezet valóban teljes legyen. A 34 helyes vonzatos komplex-ige szerkezetet a 19. táblázat tartalmazza.

DOI:10.15774/PPKE.ITK.2011.004

18. táblázat. A kinyert holland–francia párhuzamos szerkezetek kiértékelése. A kapott 58 szerkezetből a kiértékelés során 34 bizonyult helyesnek, ez 58,6 százalékos pontosságot jelent.

párhuzamos igei szerkezetek száma	58
helyes párhuzamos igei szerkezetek száma	34
pontosság	**58,6**%

5.4.3. Aszimmetrikus példák

A bevezető végén elővételeztük, hogy az algoritmusunk várhatóan leghasznosabb tulajdonsága az lesz, hogy olyan párhuzamos szerkezetek felfedezésére is képes, ahol a két nyelv teljesen más felépítésű szerkezetet használ az adott jelentés kifejezésére. Ezeket a párhuzamos szerkezeteket *aszimmetrikus*nak nevezzük.

19. definíció. *Gyenge aszimmetria.* Gyengén vagy „tartalmilag" aszimmetrikus egy párhuzamos szerkezet, ha ugyanannyi LSzB és LKB szerepel benne, de a bővítmények nem az alapértelmezett módon megfelelnek egymásnak: tárgynak nem tárgy felel meg, vagy a tartalmi elemeknek illetve a viszonyjelölőknek nem a szokásos fordítása szerepel.

20. definíció. *Erős aszimmetria.* Erősen vagy „formailag" aszimmetrikus egy párhuzamos szerkezet, ha a bővítmények közvetlenül nem feleltethetők meg egymásnak, vagy a bővítmények száma nem is egyezik a két nyelvben.

A 19. táblázatban aszimmetrikusként megjelölt szerkezetek közül a legérdekesebb a következő három:

- A 18. sorszámú szerkezet klasszikus példája az egyszerű és komplex ige megfelelésének: a '*részt vesz*' fogalmát a holland nyelv a magyarhoz hasonlóan komplex igével ('*nemen deel*') fejezi ki, a francia pedig a korpusz tanúsága szerint általában egy szóval ('*participer*').
- A 22. sorszámú szerkezet aszimmetriáját az (is) okozza, hogy a francia tárgy a hollandban nem tárgynak, hanem '*op*' elöljárós bővítménynek felel meg.
- A legbonyolultabb a 16. sorszámú szerkezet: itt a francia részen vonzatos reflexív igével ('*appliquer se à*') találkozunk, a hollandban pedig egy vonzatos létigés komplex igével ('*zijn van·toepassing op*').

Az eredmények jól mutatják az ismert tényt, hogy a különböző nyelvek egyes nyelvi elemei csak ritkán fedik le pontosan egymást, csak ritkán felelnek meg pontosan egymásnak (Atkins és Rundell, 2008, 467. oldal): sokszor van példa arra, hogy az egymás fordításának vélt szavak csak bizonyos környezetben fordításai egymásnak, vagy bizonyos környezetben nem fordításai egymásnak. Másképp fogalmazva a nyelvi elemek (például igék vagy elöljárók), a kifejezések (és jelentések) különböző részhalmazait fedik le, és két nyelv viszonylatában ezek a részhalmazok szinte soha nem esnek

DOI:10.15774/PPKE.ITK.2011.004

19. táblázat. A kinyert 34 helyes vonzatos komplex ige. A második és harmadik oszlopban a párhuzamos szerkezetből levezetett holland illetve francia szerkezet olvasható. A negyedik oszlopban a párhuzamos szerkezet gyakorisági értéke található.

#	holland szerkezet	francia szerkezet	gyak	magyar megfelelő	megjegyzés
1.	gaan om	agir SE de (1)	114	'szó van vmiről'	
2.	zijn obj	agir SE de (2)	69	'vmi van'	
3.	houden REKENING met (1)	tenir COMPTE de	40	'számításba vesz vmit'	
4.	hebben obj	avoir BESOIN de	39	'szükség van vmire'	holland határozószó ('*nodig*') hiányzik
5.	bestaan uit	composer SE de	35	'áll vmiből'	*aszimmetrikus*
6.	stellen te·BESCHIKKING van	mettre à·DISPOSITION de	31	'rendelkezésére bocsát'	a tárgy már nem fért bele a 4 pozícióba
7.	spelen ROL in	jouer RÔLE dans	30	'szerepet játszik vmiben'	
8.	bedoeld in·ARTIKEL	viser obj à·ARTICLE	30	'hivatkozik paragrafusban'	
9.	doen BEROEP op	faire APPEL à	29	'fellebbez vkihez'	
10.	betreffen obj	agir SE de (3)	27	kb. 'illeti'	
11.	zijn STAD−subj obj	être VILLE−subj obj	26	'a város vmilyen'	
12.	vermelden in·ARTIKEL	viser obj à·ARTICLE	24	'említ paragrafusban'	
13.	maken DEEL van	faire PARTIE de	24	'részét képezi vminek'	
14.	gaan over	agir SE de (4)	24	'szó van vmiről'	
15.	zien AFBEELDING	voir FIGURER de	23	'lásd az ábrát'	
16.	zijn van·TOEPASSING op	appliquer SE à (1)	22	'érvényes, vonatkozik vmire'	*aszimmetrikus*
17.	gelden voor	appliquer SE à (2)	22	'érvényes, vonatkozik vmire'	*aszimmetrikus*
18.	nemen DEEL aan	participer à	21	'részt vesz vmiben'	*aszimmetrikus*
19.	richten ZICH tot	adresser SE à	19	'megcéloz, megszólít vkit'	
20.	kennen VOORDEEL	octroyer AVANTAGE de	19	'megvan az előnye vminek'	
21.	houden REKENING met (2)	prendre en	19	'számításba vesz vmit'	ti. en·COMPTE/CONSIDÉRATION
22.	hebben BETREKKING op	concerner obj	19	'vonatkozik vmire'	*aszimmetrikus*
23.	zijn op·ZOEK naar	être à·RECHERCHE de	18	'keres vmit'	
24.	heten	appeler SE obj	18	'hívják vhogy'	
25.	hebben EFFECT op	avoir EFFET sur	18	'(vmilyen) hatása van vmire'	
26.	zijn in·BELGIË	être en·BELGIQUE de	17	'van Belgiumban'	
27.	vergaderen	réunir SE de	17	'találkozót tart, összeül'	
28.	zijn obj	être obj à·FOI	16	'egyszerre van'	'à la fois' = ugyanakkor + holland határozószó
29.	stoppen	arrêter SE de	16	'befejeződik'	
30.	liggen aan·BASIS van	être à·BASE de	16	'vminek az alapja'	
31.	branden	allumer SE de	16	'ég (pl. lámpa)'	
32.	bedragen EURO	élever SE à	16	'(vm euró) össszeget tesz ki'	*aszimmetrikus* (hiányzik a francia 'euro')
33.	zijn obj	faire OBJET de	15	'vmi tárgyát képezi'	
34.	spelen ROL	jouer RÔLE de	15	'szerepet játszik'	'vmiben' nélküli változat (vö: 7.)

DOI:10.15774/PPKE.ITK.2011.004

pontosan egybe, az átfedés mértéke széles határok között változik. Mikor egy párhuzamos szerkezetben egy tartalmas szónak nem a szokásos fordítása van jelen, máris egy gyengén aszimmetrikus szerkezettel van dolgunk.

A párhuzamos szerkezetek szépen megadják az igék egy-egy „jelentését" (vö: 1.4.7. rész a 24. oldalon), pontosabban azt, hogy adott környezetben, az adott elemek mellé éppen melyik ige illik. A szerkezet többi része sok esetben „szó szerinti" fordítás, és pontosan az ige az, amely kifejezésről kifejezésre más-más, nem kikövetkeztethető, megtanulandó, idiomatikus. Így van ez a 9. és a 13. szerkezet (19. táblázat) esetében, mikor a 'csinál' jelentésű francia '*faire*' az egyik kifejezésben a hasonló jelentésű holland '*doen*'-nal áll párban, máskor pedig a szintén hasonló jelentésű '*maken*'-nel, de nem felcserélhető módon. Hasonlóan viselkednek az elöljárók is, gyakran kevéssé megjósolható módon. A nagyjából '*-on/-en/-ön*' vagy '*-ra/re*' szerepű elöljárók közül valamikor az '*op–à*' (16. szerkezet), máskor pedig az '*aan–à*' (18. szerkezet) áll párban, ugyanakkor az '*op*'-nak a '*sur*' is megfelelhet (25. szerkezet).

5.4.4. Összefoglalás

Az eredetileg egynyelvű korpuszra kidolgozott módszert sikerrel alkalmaztuk párhuzamos korpuszra, a módszer korpuszvezérelt módon, *kétnyelvű*, *párhuzamos* igei szerkezetek hasznos gyűjteményét képes előállítani. Más szóval képes hozzárendelni a másik nyelvű megfelelőt az egyes szerkezetekhez. Lényeges tulajdonsága, hogy felfedezi és párba állítja az aszimmetrikus, formailag egymásra nem hasonlító, de egymásnak megfelelő, egymás fordításaiként kezelendő igei szerkezeteket.

A nyelvenkénti 3-3,5 millió szavas korpusz ilyen feladatra kicsinek számít, ezért viszonylag alacsony a kapott szerkezetek száma. A párhuzamos korpuszok előállítási költsége magas, ezért a közeljövőben maximum ennél egy nagyságrenddel nagyobb párhuzamos korpuszokra számíthatunk. Ezek használata azonban már jelentősen növelhetné a kinyerhető párhuzamos szerkezetek mennyiségét.

Amint a fentiekben láttuk, rendre egyszerű közelítő módszereket alkalmaztunk az előkészítő, elemző lépések során. Az e lépések során előforduló különféle hibáktól, hiányosságoktól függetlenül egyértelművé vált a módszer képessége az egymásnak megfelelő igei szerkezetek közvetlen megragadására. Az elemzési lépések fejlesztése nagy mértékben javíthatna a végső eredmény minőségén, de az a mostani vizsgálatból így is látszik, hogy maga az algoritmus megfelel a kívánt célnak.

Említettük (85. oldal), hogy a szótárnak az lehet az egyik jó tulajdonsága, azzal segítheti legjobban a nyelvhasználatban a szótárhasználót, ha ötletet ad arra, hogy egy kívánt dolgot hogyan szoktak mondani a másik nyelven. Ennek a követelménynek az algoritmus által felépített párhuzamos igeiszerkezet-tár megfelel. Sok esetben nem mondhatjuk, hogy a kinyert holland és francia kifejezés jelentése azonos, az viszont igaz, hogy ha egy helyzetben az egyik nyelvben az egyik kifejezés használatos, akkor hasonló helyzetben a másik nyelven a párját használhatjuk.

A algoritmus párhuzamos igei szerkezetek kinyerésére való adaptálást **a 7. (utolsó) tézis** tartalmazza, ez a most következő összefoglaló végén található a 114. oldalon.

DOI:10.15774/PPKE.ITK.2011.004

DOI:10.15774/PPKE.ITK.2011.004

6. fejezet

Összefoglalás: új tudományos eredmények

A dolgozat jellegzetes igei szerkezetek korpuszból való kinyerésével foglalkozik. Elsősorban azokra az igei szerkezetekre koncentrál, melyek egyszerre többszavas kifejezések és vonzatkeretek, azaz a vonzattal rendelkező komplex igékre. Ilyen például a '*hasznot húz vmiből*', az '*igényt tart vmire*' vagy az '*lehetővé tesz vmit*'. Ezek a szerkezetek lexikálisan szabad bővítményt, LSzB-t ('*vmiből*', '*vmire*', '*vmit*'), és lexikálisan között bővítményt, LKB-t ('*hasznot*', '*igényt*', '*lehetővé*') is tartalmaznak.

Az első feladat az volt, hogy kidolgozzak egy olyan modellt magyar nyelvre, mely az igei szerkezetek összes típusát – különös tekintettel a fent említett típusra – ábrázolni képes. Erre egy speciális függőségi elemzés alapú gráf volt a legalkalmasabb.

A modell kialakításával a 27. oldalon kezdődő 2.1. részben foglalkozom, az új eredményeket a következőképpen foglalhatjuk össze:

1. tézis.
Kidolgoztam magyar nyelvre egy olyan modellt, mely képes a tagmondatok, illetve a bennük rejlő formailag nagy mértékben különböző igei szerkezetek egységes reprezentálására. A reprezentáció alapegysége a tagmondat, mely egy központi ige és a hozzá tartozó bővítmények összességét jelenti. A bővítményeket legfontosabb tartalmi elemükkel (névszói csoport bővítmény esetén a bővítményt képviselő csoport feje) és a bővítményt az igéhez kapcsoló függőségi viszonnyal (névszói csoport bővítmény esetén az esetrag vagy névutó) jellemzem. Összefoglalva:

tagmondat = ige + bővítmények halmaza

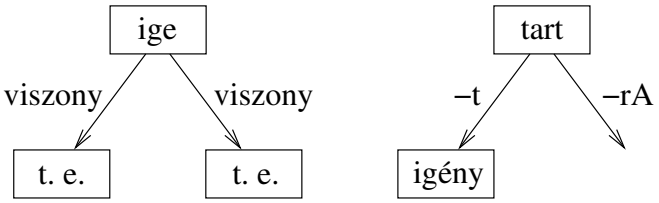
bővítmény = viszonyjelölő + tartalmi elem

A tézishez kapcsolódó publikáció:
(Sass, 2009c), (Sass, 2009a), (Sass, 2008), (Sass, 2005)

DOI:10.15774/PPKE.ITK.2011.004

6. Összefoglalás: új tudományos eredmények

A modell legszemléletesebben 1-mélységű függőségi fával ábrázolható, melynek az ige a gyökere, az élek a viszonyjelölők, a csomópontok pedig a tartalmi elemek. A 29. ábrán látható a modellnek megfelelő általános függőségi fa, és az egyik fenti szerkezet konkrét reprezentációja.



29. ábra. A modell megjelenítése függőségi fával. Bal oldalon a modellnek megfelelő általános függőségi fa látható viszonyjelölőkkel és tartalmi elemekkel (t. e.), jobb oldalon pedig egy konkrét szerkezet, az '*igényt tart vmire*' reprezentációja. Az LSzB-hez (esetünkben ez a '*vmire*') tartozó tartalmi elem nem része a szerkezetnek.

— • —

A következő kérdés nyilván az, hogy hogyan alakítható ki egy korpusznak a fenti modell szerinti reprezentációja. Természetesen előállítható ez a forma egy függőségileg elemzett korpuszból (treebank-ból), vagy függőségi elvű szintaktikai elemző felhasználásával. Megfelelő méretű függőségileg elemzett korpusz, illetve függőségi elemző magyar nyelvre nem állt rendelkezésre. Dolgozatomnak nem célja egy magyar függőségi elemző kialakítása (ez egy önálló dolgozat tárgya lehetne), a további kutatáshoz egy nagy méretű korpusz megfelelő minőségű reprezentációjára volt szükségem.

Reprezentatív magyar nyelvű korpuszként a 187 millió szavas Magyar Nemzeti Szövegtárat választottam, és azt vizsgáltam meg, hogy közelítő módszerrel, szabályalapú megközelítéssel, egyszerű szabályokkal elő lehet-e állítani a szükséges reprezentációt. Kiderült, hogy a tagmondatra bontás és a szükséges részleges szintaktikai elemzés (lényegében igeazonosítás és névszói csoport bővítmények azonosítása) is megfelelő minőségben megoldható így.

A korpusz feldolgozását a 34. oldalon kezdődő 2.2. részben tárgyalom, a fejezet tanulságát a következő tézis mondja ki:

2. tézis.
Megmutattam, hogy morfoszintaktikailag annotált korpuszból szabályalapú tagmondatra bontással és szabályalapú részleges szintaktikai elemzéssel, viszonylag egyszerű szabályrendszerrel megbízható modell szerinti reprezentációjú korpusz állítható elő.

A tézishez kapcsolódó publikáció:
(Sass, 2006b), (Sass, 2005)

DOI:10.15774/PPKE.ITK.2011.004

Természetesen a jövőben egy valódi függőségi elemző felhasználásával a reprezentáció minősége javítható, de mostani állapotában is elegendően jó ahhoz, hogy a további kutatásnak alapanyaga lehessen.

— • —

Az így létrehozott reprezentáció önmagában értékes erőforrás. Mint speciális korpusz különböző olyan lekérdezésekre ad lehetőséget, melyek egy korpuszlekérdezőnél nem megszokottak: elvonatkoztathatunk a szórendtől, az igei szerkezeteket az adott korpuszmondatban épp megjelenő szórendjüktől függetlenül egységesen vizsgálhatjuk. Ezért készítettem el a Mazsola elnevezésű korpuszlekérdező rendszert, melynek segítségével az igék, illetve igei keretek mellett megjelenő jellegzetes bővítményeket vizsgálhatjuk. Megjeleníti a lekérdezésben megjelölt bővítményi helyen megjelenő tipikus szavakat, és a hozzájuk tartozó megfelelő korpuszpéldákat is.

A rendszer alapvetően kétféle tipikus bővítményt szolgáltat. Egyrészt a „szó szerinti" értelmű szavakat, melyek sok esetben szemantikailag egységes csoportot alkotnak; ilyenek például az '*eszik vmit*' tárgyi bővítményeként megjelenő különféle ételek ('*kenyér*', '*hús*', '*hal*', '*leves*' stb.). Másrészt viszont az idiomatikus, komplex igék, vagy szólások elemét alkotó szavakat; ilyen a szintén az '*eszik vmit*' lekérdezés eredményében szereplő '*kása*', mely nem azért kerül a jellegzetes szavak közé, mert manapság olyan tipikus étel lenne, hanem pontosan a '*nem eszik olyan forrón a kását*' szólás miatt.

A Mazsola korpuszlekérdezőt a 47. oldalon kezdődő 3.2. részben ismertetem, jellemzőit az alábbi tézisben fogalmazom meg:

3. tézis.
Létrehoztam a Mazsola elnevezésű speciális korpuszlekérdező eszközt. Segítségével feltérképezhetjük az igék bővítményszerkezetét, megállapíthatjuk igék, illetve igei keretek lényeges bővítményeit, beleértve a komplex igéket is. Hasznos segédeszköz a korpuszalapú nyelvészeti kutatásban, lexikai adatbázisok kézi építésekor, és igei szerkezetekre való példák keresésekor.

A tézishez kapcsolódó publikáció:
(Sass és Pajzs, 2010b) (Sass, 2009b) (Sass, 2008) (Sass, 2006b)

A rendszer tetszőleges modell szerinti reprezentációjú korpuszra alkalmazható. A Magyar Nemzeti Szövegtár anyagát tartalmazó eredeti magyar változat keresőfelülete szabadon elérhető a `http://corpus.nytud.hu/mazsola` internetes címen, ki is próbálható a `vendeg` ideiglenes felhasználói névvel és a hozzá tartozó `mazsola` ideiglenes jelszóval. Százmillió szavas korpuszméret mellett a lekérdezések feldolgozási ideje mindössze néhány másodperc.

— • —

DOI:10.15774/PPKE.ITK.2011.004

6. Összefoglalás: új tudományos eredmények

A mai korpuszok elérték azt a méretet, mikor a kézi lekérdezők mellett szükség van olyan eszközökre is, melyek automatikusan összegzik a korpuszból kinyerhető információt. A Mazsola ebből a szempontból a kézi lekérdezőnek felel meg, képes konkrét igei keret konkrét bővítményi helyén megjelenő tipikus szavakat bemutatni.

Dolgozatom legfontosabb eredménye az az automatikus módszer, mely ennél egy nagyon fontos lépéssel tovább megy: képes arra, hogy korpusz alapján meghatározza, hogy *egyáltalán* mik egy ige jellegzetes bővítménykeretei, azaz automatikusan megállapítani, hogy „mi mindent érdemes" a Mazsolától kérdezni, és mintegy ezeket a lekérdezéseket „le is futtatja". Ezáltal az egyes igékhez tartozó jellegzetes igei szerkezeteket tudjuk számba venni.

Az algoritmus részletes bemutatása és kiértékelése az 54. oldalon kezdődő 3.3. részben található, lényegét a következő tézis foglalja össze:

> **4. tézis.**
> **Kidolgoztam egy lexikai kinyerő eljárást, mely a mondatvázak gyakoriságainak speciális összegzésére épül. Ez az eljárás alkalmas arra, hogy a modell (1. tézis) szerinti reprezentációval bíró korpuszból a különféle bonyolultságú, jellegzetes igei szerkezeteket kinyerje.**
>
> A tézishez kapcsolódó publikáció:
> (Sass, 2010d), (Sass és Pajzs, 2010b), (Sass, 2009c)

A módszer újdonsága, hogy egyrészt alkalmazkodik az igei szerkezet elemszámához, azaz kettő illetve több elemű kifejezéseket egyaránt eredményez; másrészt képes felfedezni, hogy az ige mellett egy adott fontos bővítmény esetén csak a viszony (LSzB) vagy a konkrét tartalmi elem is (LKB) lényeges: LSzB-ket és LKB-kat – akár vegyesen – tartalmazó igei szerkezeteket egyaránt szolgáltat. Az utóbbi csoportba tartoznak az 1. tézisnél említett '*hasznot húz vmiből*', '*igényt tart vmire*' és '*lehetővé tesz vmit*' vonzatos komplex igék.

— • —

Az algoritmus által szolgáltatott, igei szerkezeteket tartalmazó lista felhasználásával egy igei szerkezeteket tartalmazó szótár készíthető el. Az igei szerkezeteket az igék köré rendezve automatikusan előállított nyers szócikkekhez jutunk. Ahhoz, hogy ebből egy kiadható szótár álljon elő manuális lexikográfiai munkára van szükség. A lexikográfiai munkaigény alacsony, a munka az ellenőrzésre és példaválasztásra korlátozódik, a szótár gyorsan és kis költségvetéssel előállítható. A szótár vonzatkeretszótár, kollokációs szótár és gyakorisági szótár egyszerre, valamint a szofisztikált mutatók révén lehetővé teszi az igei szerkezetek összevetését számos szempont szerint.

A szótárkészítés lépéseit, magát a szótárt, és lehetséges felhasználásait a 73. oldalon kezdődő 4.2. részben tárgyalom, jelentőségét az alábbi tézis fogalmazza meg:

DOI:10.15774/PPKE.ITK.2011.004

5. tézis.
Létrehoztam egy új típusú szótárt, melynek alapelemei nem szavak, hanem szószerkezetek: az igei szerkezetek. A puszta szövegtől a nyers szócikkekig tisztán automatikus nyelvfeldolgozó eszközökkel jutottam el, melyek közül kiemelendő a jellegzetes igei szerkezeteket kinyerő algoritmus (4. tézis), mely a szótári anyaggyűjtést automatizálja. Megmutattam, hogy ez a lexikai kinyerő eljárás jól alkalmazható a szótárkészítésben: az elkészült szótár valóban a nyelvre jellemző vonzatokat és igei kifejezéseket tartalmazza. Olyan tanulói szótár jött így létre, mely a legfontosabb igei jelentéseket megvilágítja, elősegíti az „idiomatikus", a nemcsak nyelvtanilag helyes, hanem magyarul megszokott kifejezésmódot.

A tézishez kapcsolódó publikáció:
(Sass et al., 2010a) (Sass és Pajzs, 2010b) (Pajzs és Sass, 2010) (Sass és Pajzs, 2010c)

Hogyan használhatjuk a szótárt a nyelvtanulás támogatására, ha külföldiként magyarul akarunk megnyilatkozni? Segítségével feltérképezhetjük az ige–névszó kollokációkat: meghatározhatjuk az igékhez társítható névszókat, és (a kötött szavak szerinti mutató segítségével) a névszókhoz társítható igéket is. Ha angolként a magyarul akarunk megszólalni, és a '*meet the requirements*' megfelelőjét keressük, akkor a '*követelmény*' szónál meg fogjuk találni, hogy az ehhez illeszkedő ige a '*megfelel*', és nem a '*találkozik*' vagy valami hasonló.

A kész szótár (Sass et al., 2010a) hozzáférhető, megjelent a Tinta Könyvkiadó gondozásában.

— • —

Külön jelentőséget ad egy automatikus nyelvfeldolgozó eljárásnak, ha nyelvfüggetlen. A mi megközelítésünk nyelvfüggetlensége a reprezentáció előállíthatóságának nyelvfüggetlenségén múlik. A reprezentációra épülő eszközök, eljárások (a korábbi tézisekben ismertetett korpuszlekérdező, az igei szerkezeteket kinyerő eljárás, a szótárkészítés automatikus része) a reprezentáció automatikus folyományai. Mivel a reprezentáció lényegében csak arra támaszkodik, hogy van a nyelvekben prédikátum–argumentum struktúra, az várható, hogy a reprezentáció számos nyelvre előállítható. Ezt a sejtést a magyartól különböző szerkezetű dán és szerb, nyelvvel végzett kísérletek révén támasztottam alá.

A módszer nyelvfüggetlenségét a 89. oldalon kezdődő 5.1. részben tárgyalom, a fejezet eredményét a következő tézis tartalmazza:

6. tézis.
Megmutattam, hogy az 1. tézis szerinti egységes reprezentáció nyelvfüggetlen, számos nyelvre kialakítható. Ez lényegében azon múlik, hogy a nyelvek megnyilatkozásai felbonthatók igéből és az ige bővítményeiből

DOI:10.15774/PPKE.ITK.2011.004

álló egységekre (tagmondatokra), valamint megadható az egyes bővítmények és az ige közötti függőségi viszony. A korpuszlekérdező (3. tézis) elkészítése alig igényel plusz munkát, egyszerűen beilleszthetjük az új korpuszt az eddigiek közé. A 4. tézisben leírt algoritmus tetszőleges egységes reprezentációjú korpuszon ugyanúgy futtatható, ezáltal az igei szerkezetek gyűjtése nyelvfüggetlen módon megvalósítható. Végeredményben az erre épülő, az 5. tézisben bemutatott szótár is előállítható, korlátozott mennyiségű manuális lexikográfiai munka befektetésével.

A tézishez kapcsolódó publikáció:
(Sass, 2009d)

A jövőben a módszerrel az előző tézisben bemutatott magyar nyelvű szótárhoz hasonló nyelvtanulást segítő szótárak készülhetnek egyéb – hazánkban keresett – idegen nyelvekre is.

— • —

A modellt (1. tézis) többféle módon is kiterjeszthetjük, pontosabban többféle bonyolultabb struktúrát visszavezethetünk a 29. ábrán is látható 1-mélységű függőségi fa szerkezetre. A legizgalmasabb kérdés az, hogy elő tudunk-e állítani olyan reprezentációt, mely párhuzamos korpusz alapján készül, párhuzamos tagmondatokat, és ezáltal párhuzamos szerkezeteket (szerkezeteket és megfelelő fordításaikat) tartalmaz; de emellett megfelel az eredeti modellnek, következésképpen a kinyerő algoritmusunk futtatható rajta. Ezen a módon egy olyan eljárást nyernénk, mely a változatlan kinyerő eljárás alkalmazásával párhuzamos szerkezeteket eredményezne: az igei szerkezetekhez megkapnánk másik nyelvű fordításaikat is.

A modell kiterjesztéseit az 5.2 és az 5.3 fejezetben tárgyalom, a módszernek a párhuzamos igei szerkezetek kinyerésére való alkalmazásáról a 100. oldalon kezdődő 5.4. részben számolok be, az alábbi tézis összegzi ezt az ígéretes irányt:

7. tézis.
Megmutattam, hogy egy párhuzamos tagmondat (azaz két különböző nyelvű, egymásnak megfelelő tagmondat) közös reprezentációja kialakítható az eredeti modell szerinti formában: a központi elem a két (különnyelvű) igéből alkotott pár lesz, a bővítményeket pedig egy összesített halmazként rendelem e központi elem mellé. Ezzel előáll a párhuzamos korpuszok olyan reprezentációja, mely formailag megegyezik az egynyelvű korpuszok eredeti modell szerinti reprezentációjával. Az igei szerkezeteket kinyerő eljárást ezen a reprezentáción közvetlenül futtatva kétnyelvű, párhuzamos igei szerkezeteket, azaz szerkezeteket és a másik nyelvű megfelelőiket tudtam kinyerni. A módszer képes arra, hogy párba állítson olyan szerkezeteket is, melyek aszimmetrikusak, azaz a két nyelven teljesen eltérő felépítésűek.

A tézishez kapcsolódó publikáció:
(Sass, 2010d)

DOI:10.15774/PPKE.ITK.2011.004

A párhuzamos szerkezetekre vonatkozó vizsgálatokat egy holland–francia korpuszon végeztem. Az eredményben megkaptam például a holland '*nemen deel aan*' és a francia '*participer à*' alkotta aszimmetrikus párt (jelentésük: '*részt vesz vmiben*'). Látjuk, hogy amit a holland összetett igével fejez ki, azt a francia itt egy szóval, egy egyszerű igével.

A módszer segítségével a jövőben olyan nyelvtanulást segítő kétnyelvű szótárak állíthatók elő, melyek a használatból nyert egymásnak megfeleltetett igei szerkezetek révén elősegítik a jobb nyelvhasználatot, az anyanyelvi beszélők számára is természetes nyelvi produkciót. A kétnyelvű szótárak ilyen előállításának kidolgozása a jövő feladata, dolgozatom egy fontos lépés ebben az irányban.

DOI:10.15774/PPKE.ITK.2011.004

DOI:10.15774/PPKE.ITK.2011.004

Köszönetnyilvánítás

Köszönöm feleségemnek, *Dórinak*, az állandó támogatást és biztatást. Köszönöm, hogy a dolgozatírás sűrű időszaka alatt lényegében minden otthoni feladat alól mentesített. És a finom ebédeket. Köszönöm a gyerekeknek, *Micinek*, *Csöpinek*, *Lencsinek* és *Jáninak*, hogy megértették, hogy amikor a gépnél ülök nem szabad zavarni. És a hülyéskedéseket. Köszönöm *szüleimnek, húgomnak* és a tágabb családnak is a támogatást és biztatást. Köszönöm, hogy elolvasták és megbírálták az irományaimat, sőt volt, hogy az annotálásban is részt vettek.

Köszönöm témavezetőmnek, *Prószéky Gábornak*, a támogatást és biztatást, a baráti hangnemet, a szemesi konzultációkat. Köszönöm főnökömnek, *Váradi Tamásnak*, hogy az MTA Nyelvtudományi Intézetben lehetőséget adott arra, hogy a módszeremet a gyakorlatban is kipróbálhassam, és együtt elkészíthessük a *Magyar igei szerkezetek* szótárt. Köszönöm a konferencia-részvételek nagyvonalú támogatását, és azt, hogy közvetlen hozzáférést kaphattam a Magyar Nemzeti Szövegtárhoz. Köszönöm legközelebbi munkatársamnak, *Oravecz Csabának*, a folyamatos nyelvészeti és programozásbeli tanácsokat, és az angol nyelvű cikkek kijavítását. Köszönöm kollégnőmnek, *Pajzs Julinak*, hogy bevezetett a lexikográfiába és a közösen írt cikkeket. Köszönet a doktori iskola vezetőinek, *Roska Tamásnak* és *Szolgay Péternek*, hogy elfogadták, hogy munka mellett (egy kicsit más ütemben) végzem a doktori feladatokat.

Köszönet *Vajda Petinek*, aki talán 2009-ben azt mondta: „Hát akkor neked a Mazsola lesz a PhD-d." Köszönet *Vajda Feri* barátomnak, aki azt tanácsolta, hogy már a doktori tanulmányok elején kezdjem el a téziseket írni, bár nem fogadtam szót neki. Köszönet *Bottyán Gergőnek*, aki szerint „az a fontos, hogy amiket csinálunk, azt fel tudjuk fűzni egy szép gondolatmenetre."

Köszönet *Tihanyi Lacinak*, akinek hatására született meg a Mazsola, *Merényi Csabának* a Mazsola név ötletéért, *Héja Enikőnek* a szakmai beszélgetésekért és a holland korpusz részletes elemzéséért, *Kiss Margitnak* kritikus megjegyzéseiért és a szótári példaválasztás szempontjainak kidolgozásáért, *Gábor Katának* és *Varasdi Károlynak*, hogy rendelkezésemre bocsátották kéziratait.

Köszönet *Bankó Évának*, *Bérci Norbinak*, *Budinszky Andrásnak*, *Kis Balázsnak*, *Kuti Juditnak*, *Laki Lacinak*, *Miháltz Marcinak*, *Nagy Viktornak*, *Orosz Gyurinak*, *Ott Ferinek*, *Papp Gyulának*, *Pintér Tibinek*, *Pohl Gábornak*, *Simon Eszternek*, *Takács Dávidnak*, *Vincze Verának* és mindenkinek, akik támogattak, biztattak és segítségemre voltak a doktori évek és a dolgozatírás ideje alatt.

Köszönet azoknak, akik imádkoztak értem, és annak, aki ezeket az imákat meghallgatta.

DOI:10.15774/PPKE.ITK.2011.004

DOI:10.15774/PPKE.ITK.2011.004

A szerző publikációi

Könyv

Sass Bálint – Váradi Tamás – Pajzs Júlia – Kiss Margit 2010a. *Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára*. Tinta Könyvkiadó, Budapest.

Folyóiratcikk

Sass Bálint – Pajzs Júlia 2010b. Igei szerkezetek gyakorisági szótára – félautomatikus szótárkészítés nyelvtechnológiai eszközök segítségével. *Alkalmazott Nyelvtudomány*, 2010(1–2):5–32.

Könyvfejezet

Sass Bálint 2006a. Extracting idiomatic Hungarian verb frames. In Salakoski, Tapio – Ginter, Filip – Pyysalo, Sampo – Pahikkala, Tapio (eds.): *Advances in Natural Language Processing*, 303–309. Springer, Berlin Heidelberg New York. Lecture Notes in Computer Science, Vol. 4139.

Sass Bálint 2008. The Verb Argument Browser. In Sojka, Petr – Horák, Aleš – Kopecek, Ivan – Pala, Karel (eds.): *Text, Speech and Dialogue*, 187–192. Springer, Berlin Heidelberg New York. Lecture Notes in Computer Science, Vol. 5246.

Sass Bálint 2009a. Korpusznyelvészeti eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Sinkovics Balázs (szerk.): *LingDok 8. – Nyelvész-doktoranduszok dolgozatai*, 143–155. JATEPress, Szeged.

Sass Bálint 2009b. „Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Váradi Tamás (szerk.): *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból*, 117–129, MTA Nyelvtudományi Intézet, Budapest.

Sass Bálint – Pajzs Júlia 2010c. FDVC – creating a corpus-driven frequency dictionary of verb phrase constructions. In Granger, Sylviane – Paquot, Magali (eds.): *eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Cahiers du CENTAL 7. Presses universitaires de Louvain*, 263–272, Louvain-la-Neuve, Belgium.

DOI:10.15774/PPKE.ITK.2011.004

A szerző publikációi

Külföldi konferenciakötet

Pajzs Júlia – Sass Bálint 2010. Towards semi-automatic dictionary making. In *Proceedings of the XIV. EURALEX International Congress*, 453–462.

Sass Bálint 2007. First attempt to automatically generate Hungarian semantic verb classes. In *Proceedings of the 4th Corpus Linguistics conference*, Birmingham.

Sass Bálint 2009c. A unified method for extracting simple and multiword verbs with valence information and application for Hungarian. In *Proceedings of RANLP 2009*, 399–403, Borovets, Bulgária.

Sass Bálint 2009d. Verb Argument Browser for Danish. In *Proceedings of the 17th Nordic Conference of Computational Linguistics, NoDaLiDa 2009*, 263–266, Odense, Dánia.

Hazai konferenciakötet

Sass Bálint 2005. Vonzatkeretek a Magyar Nemzeti Szövegtárban. In Alexin Zoltán – Csendes Dóra (szerk.): *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2005)*, 257–264, Szeged.

Sass Bálint 2006b. Igei vonzatkeretek az MNSZ tagmondataiban. In Alexin Zoltán – Csendes Dóra (szerk.): *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2006)*, 15–21, Szeged.

Sass Bálint 2010d. Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2010)*, 102–110, SZTE, Szeged.

DOI:10.15774/PPKE.ITK.2011.004

Irodalomjegyzék

Abney, Steven 1996. Partial parsing via finite-state cascades. In *Proceedings of the 8th European Summer School in Logic, Language and Information (ESSLLI96) Robust Parsing Workshop*, 8–15, Prága, Csehország.

Artstein, Ron – Poesio, Massimo 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Atkins, B. T. Sue – Rundell, Michael 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Baldwin, Timothy – Villavicencio, Aline 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan.

Baldwin, Timothy 2005. The deep lexical acquisition of english verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4): 398–414.

Bárdosi Vilmos 2003. *Magyar szólástár. Szólások, helyzetmondatok, közmondások értelmező és fogalomköri szótára*. Budapest: Tinta Könyvkiadó.

Bárdosi Vilmos 2009. *Magyar szólások, közmondások értelmező és fogalomköri szótára*. Budapest: Tinta Könyvkiadó.

Bojar, Ondřej – Hajič, Jan 2005. Extracting translations verb frames. In *Proceedings of the Modern Approaches in Translation Technologies Workshop*, 2–6, Borovets, Bulgária.

Bouma, Gosse – van Noord, Gertjan – Malouf, Robert 2001. Alpino: Wide coverage computational analysis of dutch. In *Computational Linguistics in the Netherlands, CLIN 2000*. Rodopi.

Briscoe, Ted – Carroll, John 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, Washington, DC.

Burger, Harald 2003. *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Erich Schmidt Verlag, Berlin.

Cheng, Winnie – Greaves, Chris – Warren, Martin 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11(4):411–433.

DOI:10.15774/PPKE.ITK.2011.004

Irodalomjegyzék

Debusmann, Ralph 2004. Multiword expressions as dependency subgraphs. In *Proceedings of Workshop on MWEs*, 56–63, Barcelona, Spanyolország, ACL.

Dura, Elzbieta 2006. CULLER – a user-friendly conrpus query system. In *Proceedings of the Fourth International Workshop on Dictionary Writing Systems*, 47–52, Torino, Olaszország.

É. Kiss Katalin – Siptár Péter – Kiefer Ferenc 2003. *Új magyar nyelvtan*. Osiris Kiadó.

Evert, Stefan 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Evert, Stefan – Krenn, Brigitte 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics*, 188–195, Toulouse, Franciaország.

Fazly, Afsaneh – Stevenson, Suzanne 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the EACL*, 337–344, Trento, Olaszország.

Firth, John Rupert 1957. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, 1–32.

Forgács Tamás 2003. *Magyar szólások és közmondások tára*. Budapest: Tinta Könyvkiadó.

Forgács Tamás 2007. *Bevezetés a frazeológiába*. Budapest: Tinta Könyvkiadó.

Gábor Kata – Héja Enikő 2007. Clustering Hungarian verbs on the basis of complementation patterns. In *Proceedings of the ACL-SRW'07 conference*, Prága.

Gábor Kata 2005. Tagmondathatár-kijelölő rendszer. Kézirat. MTA, Nyelvtudományi Intézet.

Gábor Kata – Héja Enikő – Mészáros Ágnes 2003. Kötőszók korpusz-alapú vizsgálata. In Alexin Zoltán – Csendes Dóra (szerk.): *I. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2003)*, 305–306, Szeged, SZTE.

Goldberg, Adele E. 2006. *Constructions at Work*. Oxford University Press.

Grefenstette, Gregory 1998. The future of linguistics and lexicographers: Will there be lexicographers in the year 3000? In *Proceedings of EURALEX 1998*, 25–41, Liège.

Hanks, Patrick 2001. The probable and the possible: Lexicography in the age of the internet. In *Proceedings of AsiaLex 2001*, 1–15, Yonsei University, Szöul, Korea.

Hanks, Patrick 2005. Metaphors and meanings: a lexicographical approach to corpus analysis. In Kiefer Ferenc – Kiss Gábor – Pajzs Júlia (eds.): *Papers in Computational Lexicography, COMPLEX 2005*, 81–106. Budapest: MTA Nyelvtudományi Intézet.

Hanks, Patrick 2008. The lexicographical legacy of John Sinclair. *International Journal of Lexicography*, 21(3):219–229.

DOI:10.15774/PPKE.ITK.2011.004

Héja Enikő – Sass Bálint 2010. Többszavas kifejezések kezelése a párhuzamos korpuszokra épülő szótárkészítési módszertanban. In Vincze Veronika – Tanács Attila (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2010)*, 80–90, SZTE, Szeged.

Jackendoff, Ray 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Janssen, Maarten 2008. Meaningless dictionaries. In *Proceedings of the XIII. EURALEX International Congress*, 409–420, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

Kaalep, Heiki-Jaan – Muischnek, Kadri 2008. Multi-word verbs of Estonian: a database and a corpus. In *Proceedings of the LREC2008 workshop: Towards a Shared Task for Multiword Expressions*, 23–26, Marrakech, Marokkó.

Kálmán László 2006. Miért nem vonzanak a régensek? In Kálmán László (szerk.): *KB 120: A titkos kötet.*, 229–246.

Kilgarriff, Adam 1997. "I dont't believe in word senses". *Computers and the Humanities*, 31(2):91–113.

Kilgarriff, Adam – Tugwell, David 2001. Word Sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics, workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*, 32–38, Toulouse.

Kilgarriff, Adam – Rychly, Pavel – Smrz, Pavel – Tugwell, David 2004. The Sketch Engine. In *Proceedings of EURALEX 2004*, 105–116, Lorient, Franciaország.

Kilgarriff, Adam – Husák, Miloš – McAdam, Katy – Rundell, Michael – Rychly, Pavel 2008. GDEX: Automatically finding good dictionary examples. In *Proceedings of the XIII. EURALEX International Congress*, 425–432, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

Kim, Chang-Hyun – Hong, Munpyo 2006. A korean syntactic parser customized for korean-english patent mt system. In Salakoski, Tapio – Ginter, Filip – Pyysalo, Sampo – Pahikkala, Tapio (eds.): *Advances in Natural Language Processing*, 44–55. Springer, Berlin Heidelberg New York. Lecture Notes in Computer Science, Vol. 4139.

Kis Balázs – Villada Moirón, Begoña – Bouma, Gosse – Ugray Gábor – Bíró Tamás – Pohl Gábor – Nerbonne, John 2004. A new approach to the corpus-based statistical investigation of hungarian multi-word lexemes. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, volume V, 1677–1681, Lisszabon, Portugália.

Komlósy András 1992. Régensek és vonzatok. In Kiefer Ferenc (szerk.): *Strukturális magyar nyelvtan. I. Mondattan*, 299–527. Akadémiai Kiadó, Budapest.

DOI:10.15774/PPKE.ITK.2011.004

Irodalomjegyzék

Koutny Ilona – Wacha Balázs 1991. Magyar nyelvtan függőségi alapon. *Magyar Nyelv*, 87(4):393–404.

Kuti Judit – Varasdi Károly – Gyarmati Ágnes – Vajda Péter 2007. Hungarian WordNet and representation of verbal event structure. *Acta Cybernetica*, 18(2):315–328.

Levin, Beth 1993. *English Verb Classes and Alternations*. The University of Chicago Press.

Macken, Lieve – Trushkina, Julia – Paulussen, Hans – Rura, Lidia – Desmet, Piet – Vandeweghe, Willy 2007. Dutch Parallel Corpus. A multilingual annotated corpus. In *Proceedings of Corpus Linguistics 2007*, Birmingham, Nagy-Britannia.

Manning, Christopher D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Meeting of the Association for Computational Linguistics*, 235–242, Columbus, Ohio.

Martens, Scott – Vandeghinste, Vincent 2010. An efficient, generic approach to extracting multi-word expressions from dependency trees. In *Proceedings of the Workshop on MWEs*, 84–87, Beijing, China, ACL.

McCarthy, Diana – Keller, Bill – Carroll, John 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 73–80, Sapporo, Japán.

O. Nagy Gábor 1966. *Magyar szólások és közmondások.* Akadémiai Kiadó, Budapest.

Oravecz Csaba – Dienes Péter 2002. Large scale morphosyntactic annotation of the Hungarian National Corpus. In Hollósi Béla – Kiss-Gulyás Judit (szerk.): *Studies in Linguistics, Volume VI.*, 277–298, Debrecen.

Oravecz Csaba – Varasdi Károly – Nagy Viktor 2004. Többszavas kifejezések számítógépes kezelése. In Alexin Zoltán – Csendes Dóra (szerk.): *II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2004)*, 141–154, Szeged.

Oravecz Csaba – Nagy Viktor – Varasdi Károly 2005. Lexical idiosyncrasy in MWE extraction. In *Proceedings of the 3rd Corpus Linguistics conference*, Birmingham.

Pajzs Júlia 2000. Frazeológiai egységek a nagyszótárban. In Gecső Tamás (szerk.): *Lexikális jelentés, aktuális jelentés – Segédkönyvek a nyelvészet tanulmányozásához IV.*, 217–226. Tinta Könyvkiadó, Budapest.

Pajzs Júlia 2002. A corpus based investigation of collocations in Hungarian. In *Proceedings of EURALEX 2002*, 831–840, University of Copenhagen.

Pecina, Pavel 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC2008 workshop: Towards a Shared Task for Multiword Expressions*, 54–57, Marrakech, Marokkó.

Prószéky Gábor – Tihanyi László 1996. Humor - a morphological system for corpus analysis. In *In Proceedings of the first TELRI Seminar*, 149–158, Budapest.

DOI:10.15774/PPKE.ITK.2011.004

Prószéky Gábor – Koutny Ilona – Wacha Balázs 1989. Dependency syntax of Hungarian. In Maxwell, Dan – Schubert, Klaus (eds.): *Metataxis in Practice (Dependency Syntax for Multilingual Machine Translation)*, 151–181. Foris, Dordrecht, The Netherlands.

Pusztai Ferenc (szerk.) 2003. *Magyar Értelmező Kéziszótár*. Akadémiai Kiadó.

Ramisch, Carlos – Schreiner, Paulo – Idiart, Marco – Villavicencio, Aline 2008. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC2008 workshop: Towards a Shared Task for Multiword Expressions*, 50–53, Marrakech, Marokkó.

Recski Gábor 2010. Főnévi csoportok azonosítása szabályalapú és hibrid módszerekkel. In Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2010)*, 333–341, SZTE, Szeged.

Riedl Frigyes 1882. Simonyi kis nyelvtana. *Egyetemes Philológiai Közlöny*, 573–590.

Rundell, Michael 1998. Recent trends in english pedagogical lexicography. *International Journal of Lexicography*, 11(4):315–342.

Rundell, Michael 2009. The road to automated lexicography: First banish the drudgery... then the drudges? Elhangzott: *eLexicography in the 21st Century Conference*, Louvain-la-Neuve, Belgium.

Sag, Ivan – Baldwin, Timothy – Bond, Francis – Copestake, Ann – Flickinger, Dan 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of 3rd CICLING*, 1–15, Mexico City, Mexikó.

Sampson, Geoffrey R. 2007. Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory*, 3(1):1–32.

Seretan, Violeta 2008. *Collocation extraction based on syntactic parsing*. PhD thesis, University of Geneva.

Siepmann, Dirk 2005. Collocation, colligation and encoding dictionaries. Part I: Lexicological aspects. *International Journal of Lexicography*, 18(4):409–444.

Sinclair, John McH. 1987. *Collins Cobuild English Language Dictionary*. London: Harper-Collins publishers.

Sinclair, John McH. 1998. The lexical item. In Weigand, Edda (ed.): *Contrastive Lexical Semantics*, 1–24. Amsterdam Philadelphia: John Benjamins.

Stefanowitsch, Anatol 2006. Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory*, 2(1):61–77.

T. Litovkina Anna 2005. *Magyar közmondástár. Közmondások értelmező szótára példákkal szemléltetve*. Budapest: Tinta Könyvkiadó.

DOI:10.15774/PPKE.ITK.2011.004

Irodalomjegyzék

Tapanainen, Pasi – Piitulainen, Jussi – Järvinen, Timo 1998. Idiomatic object usage and support verbs. In *Proceedings of the 17th COLING – 36th ACL*, 1289–1293, Montreal, Canada.

Teubert, Wolfgang 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1):1–13.

Tognini-Bonelli, Elena 2001. *Corpus Linguistics at Work*. John Benjamins.

Trautner Kromann, Mathias 2003. The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Svédország.

Váradi Tamás 2002. The Hungarian National Corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, 385–389, Las Palmas, Spanyolország.

Váradi Tamás 2003. Főnévi csoport annotálása a CLaRK rendszerben. In Alexin Zoltán – Csendes Dóra (szerk.): *I. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2003)*, 65–71, Szeged, SZTE.

Váradi Tamás – Gábor Kata 2004. A magyar INTEX fejlesztésről. In Alexin Zoltán – Csendes Dóra (szerk.): *II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2004)*, 3–10, Szeged, SZTE.

Varasdi Károly 2005. Coordination. Kézirat. MTA, Nyelvtudományi Intézet.

Yarowsky, David 1993. One sense per collocation. In *Proceedings of the workshop on Human Language Technology*, 266–271, Princeton, New Jersey.

Zarrieß, Sina – Kuhn, Jonas 2009. Exploiting translational correspondences for pattern-independent MWE identification. In *Proceedings of the Workshop on MWEs*, 23–30, Singapore, ACL.

Zeman, Daniel – Sarkar, Anoop 2000. Learning verb subcategorization from corpora: Counting frame subsets. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, Athén, Görögország.

DOI:10.15774/PPKE.ITK.2011.004

Tárgymutató

DOI:10.15774/PPKE.ITK.2011.004