Alkalmazott Nyelvtudomány
X. évfolyam 1-2. szám 2010.

SASS BÁLINT ÉS PAJZS JÚLIA

MTA Nyelvtudományi Intézet és PPKE ITK IMT Doktori Iskola

sass.balint@nytud.hu, pajzs@nytud.hu

Igei szerkezetek gyakorisági szótára – félautomatikus szótárkészítés nyelvtechnológiai eszközök segítségével

One of the central issues in today's computational lexicography is the question of how much traditional manual work can be taken over by computers, how far lexicography can get with purely automated tools. In the present contribution we describe the details of a semi-automatic dictionary creation project. First we present the natural language processing toolchain which adds detailed linguistic analysis to its raw input text, turning it into what can be regarded as a raw dictionary. We go on to present the subsequent steps of manual lexicographic work the result of which is the final dictionary. It was possible to reduce the amount of manual lexicographic work significantly by using automated language technology tools. The resulting „Frequency Dictionary of Verb Phrase Constructions" has the following characteristics: (1) It is a *corpus-driven frequency* dictionary. (2) Its basic units are *verb phrase constructions*, not words, as was the case with the classical dictionary entries. (3) It is a *meaningless dictionary*, although it contains example sentences illustrating the meaning of the particular constructions. (4) The dictionary has now been created for Hungarian, but the language processing tools used are to some extent *language independent*, and can be applied to other languages, accordingly. (5) The dictionary is created largely by *automated tools*, thus, it can be produced for a relatively small budget. (6) It can be used in a number of areas, e. g. in linguistic research, in language teaching or in human language technology.

1. Szótárírás ma: automatizálás és frázisok

Már tíz évvel ezelőtt felmerült (Grefenstette, 1998), hogy meddig lesz szükség lexikográfusokra, manapság pedig az is kérdés, hogy meddig lesznek egyáltalán szótárak – és itt általában a hagyományos papíralapú szótárakra gondolnak – az online világban. Az biztosnak tűnik, hogy örök az idegen nyelvek megismerésének vágya, 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.

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. 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 használata a modern lexikográfiában elengedhetetlen követelménnyé vált. 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 (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 néznek 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 *több szóból álló lexikai egységek* – kollokációk, idiomatikus kifejezések, állandósult szókapcsolatok – jelentőségének felismerése és a korábbinál sokkal hangsúlyozottabb megjelenítése az új szótárakban. Sinclair (1998) úgy látja, hogy a nyelv valójában részben előre megkonstruált frázisokból épül fel, nem pedig egyes szavakból. A korpuszvezérelt szótárírás tapasztalatait így foglalta össze (a szerzők fordítása):

> A lexikográfia számos régóta elfogadott hagyománya megkérdőjeleződött: például az, hogy egy szónak inherensen van egy vagy több jelentése. Az eredeti 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 (1998: 2).

Sinclair (1998) 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 elhelyezésének 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 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 általában a sajátos értelműnek tekinthető állandósult szókapcsolatok gyűjtésére és értelmezésére, illetve példákkal való illusztrálására helyezik a hangsúlyt, azaz a hagyományosabb korpuszalapú megközelítéssel dolgoznak.

Jelen tanulmányban egy új, korpuszvezérelt szótárkészítési módszert mutatunk be, illetve annak alkalmazását egy konkrét szótár esetében. 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. Az 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.

A tanulmány további felépítése a következő. A 2. részben a szótárat meghatározó alapelveket fektetjük le. A 3. részben bemutatjuk, hogyan jutunk el a puszta szövegtől a nyers szótárig tisztán automatikus úton; a 4. rész az ezt követő manuális lexikográfiai munka részleteit ismerteti. Ezt követi egy szemelvény a szótárból, majd a tanulmány végén a módszer nyelvfüggetlenségéről, és a szótár lehetséges alkalmazásairól szólunk.

2. Megközelítés, elvek, a szótár tulajdonságai

Ebben a fejezetben az „Igei szerkezetek gyakorisági szótára” (ISZGYSZ) alábbi, részben már említett alapvető tulajdonságait fejtjük ki: korpuszvezérelt módon készül; alapegységei igei szerkezetek; a szótár nem tartalmaz definíciókat, a szerkezetek jelentését példamondatok világítják meg.

2.1. A korpuszvezérelt megközelítés

A szótár készítése során a sinclair-i szigorúan korpuszvezérelt megközelítést követjük (Tognini-Bonelli, 2001). 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 számba, és ezen elemek mellett gyakorisági mérőszámot is feltüntetünk.

2.2. Igei szerkezetek mint alapegységek

A szótár *központi igéből és annak névszói csoport bővítményeiből* álló szerkezeteket tartalmaz. Ezek a szótár alapegységei, lexémái, ezeket nevezzük a továbbiakban igei szerkezeteknek. Fontos megjegyezni, hogy az összes ilyen formájú kifejezést ideértjük a vonzatkeretektől (pl. *ad vkinek vmit*) a komplex igéken (pl. *hasznot húz vmiből*) át egészen a szólásokig (pl. *más malmára hajtja a vizet*). Ezt az általunk kontinuumnak tekintett teljes spektrumot egy szótárban fedjük le, nem csak egyes részeire figyelünk, ahogy az a speciális szótárakban (vonzatszótárak, kollokációs szótárak, szólástárak stb.) szokásos. Az ige a tagmondat „pillére”, ezért ezek a kifejezések a megnyilatkozások túlnyomó részét lefedik, azaz általuk az egész nyelv lexikonjáról átfogó képet kaphatunk.

1. táblázat: Példák igei szerkezetekre. LSzB: lexikálisan szabad bővítmény; LKB: lexikálisan kötött bővítmény, a pontos meghatározásokat lásd a szövegben. Az LKB-kben megjelenő ragos szavak szótő-rag alakban szerepelnek, az -A jel a birtokos személyragot jelenti.

példa	**szerkezet**	**LSzB**	**LKB**
(1) Egyszerű jogsértés történt csupán,	*x* történik	–	–
(2) Foglalkozik politikával?	*x* foglalkozik *y*-vAl	*-vAl*	–
(3) A szaxofonos vállat vont.	*x* váll-t von	–	*-t*
(4) A hivatal pontatlanságából azért hasznot is húztam.	*x* haszon-t húz *y*-bÓl	*-bÓl*	*-t*
(5) Az ő malmára hajtja a vizet,	*x* (vki) malom-A-rA hajt víz-t	–	*-t, -rA*

Az 1. táblázatban igei szerkezetekre láthatunk néhány példát. A névszói csoportokat a csoport esetragjával (illetve névutójával) és a csoport fejével reprezentáljuk. A névszói csoportok kétfélék lehetnek. *Lexikálisan szabad bővítmény*ek (továbbiakban *LSzB*) az ige vonzatainak (az ige melletti vonzati helyeknek) vagy bővítményeinek felelnek meg, itt az adott bővítményi pozíciót betöltő szó (ti. a csoport feje) széles szóosztályból választható. A *lexikálisan kötött bővítmények*nek (továbbiakban *LKB*) megfelelő bővítményi pozícióban egy konkrét szó szerepel. Az alapige és az LKB-k együtt *komplex igét* alkotnak, amennyiben (egy vagy több) névszói csoport szemantikailag szervesen hozzátartozik az igéhez, és az együttes jelentésük valamilyen mértékben nem kompozicionális, másképp fogalmazva: ha megváltoztatjuk az LKB konkrét szavát, akkor elvész a komplex ige együttes jelentése.

Kiemelendők azok a szerkezetek, melyek LKB-t és LSzB-t is tartalmaznak, ezek a vonzatos komplex igék, mint például a *részt vesz vmiben*, *egyet ért vkivel*, *kétségbe von vmit*. Az ilyen típusú szerkezetek *egyszerre* vonzatkeretek és a többszavas kifejezések: a kollokációs szótárakból vonzatuk miatt, a vonzatszótárakból pedig a jelen lévő LKB miatt „lógnak ki”[1]. A komplex igék sokkal gyakoribbak, mint azt az általános nyelvi intuíciónk sugallja, de gyakoriságukon kívül elméleti érvek is szólnak amellett, hogy általános célú szótárban nem kezelhetők marginális esetként, hanem akár önálló lexémaként – mint igék! – is megállják a helyüket. Három érvet sorakoztatunk fel, hogy miért sorolhatjuk a komplex igéket önálló elemekként az igék kategóriájába: egyrészt disztribúciós alapon látjuk, hogy igék helyén jelenhetnek meg a mondatban (vö: *megvizsgál vmit* ↔ *górcső alá vesz vmit*); másrészt az alapigétől eltérő új jelentéssel bírnak; valamint az alapigétől független új vonzatkerettel rendelkezhetnek: a *részt vesz* mellett megjelenő *-bAn* vonzat vagy az *egyet ért* melletti *-vAl* az alapige (*vesz,* illetve *ért*) mellett nem szerepelt. Ez az egységes felfogás azzal a jelentős előnnyel jár, hogy közvetlenül összevethetjük az egyszerű és a komplex igék tulajdonságait, például vonzatkereteiket vagy gyakorisági viszonyaikat.

Az ISZGYSZ-ben a lexémák valójában az egyes igei szerkezetek, a szótár mikrostrukturája nemcsak, hogy tartalmazza a többszavas kifejezéseket (frazeológiát), hanem kifejezetten frazeológia-központú, tekintve, hogy az alapelemek frázisok. A megszokotthoz közeli prezentáció érdekében azonban egy második lépésben az igei kifejezéseket az igék köré rendezzük, és így végül alapige-központú (*vesz, ért* stb.) hagyományos szócikkeket kapunk.

2.3. Definíció nélküli szótár korpuszpéldákkal

Említettük, hogy módszerünk a szótári anyaggyűjtés feladatát automatizálja. A definícióírás mai eszközeinkkel nem automatizálható, és egyik legidőigényesebb része a lexikográfiai munkának. Az ISZGYSZ nem tartalmaz definíciókat, *definíció nélküli* szótárnak (meaningless dictionary) nevezhető. Bár a definíció, a jelentés megadása a szótárak egyik legfontosabb jellemzője, van haszna az effajta szótáraknak is.

> 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: 412 ; a szerzők fordítása)

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.

Az ISZGYSZ-ben nincs definíció, de jelentéses, sőt lehetőleg egyjelentésű elemeket szándékozik felsorolni. 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. Azzal, hogy a kollokációt tesszük meg a szótár alapegységének, a poliszémia jelentős részétől automatikusan megszabadulunk. Valójában az igei szerkezetek nagy része valódi konstrukció, „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. Az ISZGYSZ koncepciója az, 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).

3. A szövegtől a szótárig

A teljes szótárkészítési folyamat az 1. ábrán tekinthető át. A nyers szövegtől (fent) a kész szótárig (lent) haladunk. Az első („automatikus”) szakaszban tisztán nyelvtechnológiai 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 morfológiai elemzés eredményeként ismertté válnak az egyes szavakhoz tartozó szótövek (pl. *vet*) és elemzések (pl. `V.Pe2`, azaz ige, felszólító mód, egyes szám második személy). A tagmondatra bontást követő szintaktikai elemzés azonosítja a tagmondat igéjét, és a különféle névszói csoportok esetét/névutóját és fejét. A jellegzetes igei szerkezeteket gyűjtő algoritmus lényegi tulajdonsága, hogy „kitalálja”, hogy melyik lexikálisan kötött fej releváns az adott igei szerkezet esetében: az ábrán látható *pillantást vet vmire* szerkezetben például a tárgy lexikálisan kötött, a *-rA* ragos bővítmény lexikálisan szabad. Az algoritmus az egyes szerkezetekhez gyakorisági mérőszámot is rendel a korpusz alapján. A szerkezetek rendszerezését követi a második szakasz manuális lexikográfiai munkája. Ennek során az automatikus szakasz hibáit javítjuk, az esetleges hibás igei szerkezeteket elhagyjuk, alkalmas példamondatokat választunk az egyes szerkezetekhez, illetve megjelöljük az idiomatikus kifejezéseket, így készül el a végleges szótár. Az alábbiakban ismertetjük a felhasznált nyelvtechnológiai eszközöket, a manuális lexikográfiai munkára a 4. részben térünk vissza.

3.1. Morfológiai elemzés és egyértelműsítés

A szótár alapjául szolgáló korpusz 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 automatikus morfológiai elemzés arra szolgál, hogy megállapítsuk az egyes szavak összes elvben lehetséges szótövét, szófaját és morfológiai elemzését; az ezt követő egyértelműsítési lépés pedig, hogy kiválasszuk ezek közül a legvalószínűbbet. Utóbbi a szó környezetében lévő jegyek alapján egymást kiegészítve statisztikai és szabályalapú módon történik. Például az „Idén nem terem sok alma.” mondatban a *terem* szónak a szótöve *terem* (és nem *tér*!), a szófaja ige (és nem főnév), az elemzése pedig `V.e3` (azaz ige, egyes szám, harmadik személy). A rendszer 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 tehát

az MNSZ-ben minden egyes szó mellett fel van tüntetve a szótő, a szófaj és a szó morfológiai elemzése.

3.2. Tagmondatra bontás

Az MNSZ-re épülő első feldolgozó lépés a tagmondatra bontás. Ennek célja, hogy olyan egységeket kapjunk, melyek egy igét és annak névszói csoport bővítményeit tartalmazzák. A morfológiailag elemzett szöveget szabályalapú megközelítéssel bontjuk tagmondatokra. A szabályok a szövegszavak és írásjelek sorozata fölött megfogalmazott reguláris kifejezések, azon alapulnak, hogy milyen a szövegben a központozás és a kötőszavak elhelyezkedése. Az egyik szabály például tagmondathatárt helyez el vessző után, amennyiben a vesszőt (esetleges kötőszó vagy határozószó közbeszúrásával) vonatkozó névmás követi. A szabályok néhány heurisztikával egészülnek ki: két finit ige közötti egyetlen írásjel/kötőszó mindenképpen tagmondathatár lesz, két finit ige közötti több írásjel/kötőszó közül pedig a leginkább jobbra esőt választjuk, csökkentve az esélyét annak, hogy hibásan, felsorolás közepére helyezzünk el tagmondathatárt. A tagmondatra bontó modul pontossága 83,6%, lefedése pedig 86,5%, azaz a bejelölt tagmondathatárok nagyjából 83,6%-a pontos, és az elvben helyes tagmondathatárok 86,5%-át találja meg a program (Sass, 2006). Ez a teljesítmény a további feldolgozáshoz elegendő, sok esetben csak olyan hibáról van szó, melyek a bővítmények meghatározására nincsen kihatással.

3.3. Szintaktikai elemzés

A részleges szintaktikai elemzés során nem törekszünk a tagmondatok teljes szintaktikai fájának felépítésére, a cél csupán az ige és a névszói csoportok azonosítása. A tagmondatra bontáshoz hasonlóan itt is szabályalapú megközelítéssel dolgozunk. A szabályok szintén a szövegszavak és írásjelek sorozata fölött megfogalmazott reguláris kifejezések, de ezek a szabályok többszintű reguláris nyelvtant alkotnak: egymásra épülnek, azaz a felismert csoportokból további szabályokkal, rekurzívan újabb, nagyobb kiterjedésű csoportok képezhetők (Sass, 2005).

Az ige(tő) azonosítása során az igetőhöz kapcsoljuk az esetleges elváló igekötőt, elhagyjuk a *-hAt* képzőt, mert 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.

A névszói csoportok két számunkra legfontosabb tulajdonsága az esetrag és a fej pozícióban megjelenő konkrét szó. A névutókat az esetragokkal azonos módon kezeljük, a bővítmények tehát esetragos vagy névutós névszói csoportok. (A továbbiakban, ha az esetragokról esik szó, a névutókat is beleértjük.) Alapesetben a csoportok a bennük szereplő utolsó szó tulajdonságait öröklik, ennek köszönhető, hogy a névszó esetragja a névszói csoport eset jegyébe kerül. A névutók (főként a személyragos névutók) természetesen ettől eltérő speciális kezelést igényelnek. Ha egy tagmondatban több azonos esetragú csoport szerepel, akkor egyszerűsítésként közülük csak az utolsót vesszük tekintetbe.

A fenti eljárással előállított szintaktikai elemzés részleges függőségi elemzésnek tekinthető, ami alapján nyilvánvalóan előállítható a tagmondatok olyan reprezentációja, ahol a tagmondat igéből és névszói csoport bővítményekből áll, a bővítmények reprezentációja pedig az esetrag, illetve a fej pozícióban megjelenő konkrét szó. „A szaxofonos vállat vont.” mondatnak (1. táblázat, (3) példa) reprezentációja tehát a következő (az alanyt `-0` jelöli):

`ige`=von `-0`=szaxofonos `-t`=váll

3.4. Jellegzetes igei szerkezetek gyűjtése

Az automatikus feldolgozás legfontosabb lépése az az eljárás, amely a szintaktikai elemzés eredményeként előálló korpuszreprezentáció alapján összegyűjti a jellegzetes igei szerkezeteket. A kifejlesztett algoritmus lényege, hogy automatikusan felismeri, hogy egyrészt mely bővítmények tartoznak szorosan az igei szerkezethez; másrészt hogy mikor lényegi elem a konkrét fej, és mikor csak az eset. Azaz például a *húz haszon-t -bÓl* esetében felfedezi, hogy az ige mellett egy lexikálisan kötött tárgy és egy szabad *-bÓl* esetragos bővítmény alkotja a szerkezetet. A gyakoriságra épülő algoritmus összesíti az adott igét tartalmazó mondatokat, és meghatározza az igéhez tartozó igei szerkezeteket. Alapötlete a következő: induljunk ki a teljes korpuszreprezentációból, és hagyjuk el azokat a bővítményeket, aelyek nem részei a szerkezetnek, illetve (az esetet megtartva) azokat a fejeket, aelyek szintén nem részei a szerkezetnek. Az algoritmus a következő lépésekből áll:

1. Vesszük a korpuszból az összes tagmondat-reprezentációt. Váltakozva töröljük belőlük a fejeket, így az 3.3. rész végén látható példából az alábbi három további szerkezet keletkezik:
   `ige`=von `-0 -t`
   `ige`=von `-0`=szaxofonos `-t`
   `ige`=von `-0 -t`=váll
   Erre az átalakításra van szükség ahhoz, hogy a végső listában LSzB-t tartalmazó szerkezetek is megjelenhessenek, mivel a korpuszból vett eredeti tagmondatokban természetesen minden bővítményi pozíción szerepel valamilyen konkrét szó.
2. Hossz szerint csökkenő sorba rendezzük az igei szerkezetek 1. lépés

szerint kiegészített teljes listáját. Egy szerkezet hosszát a benne található esetek és fejek összesített száma adja.

3. A leghosszabbtól kezdve elhagyjuk azokat, melyeknek a gyakorisága 5-nél kisebb. Az elhagyott szerkezetek gyakoriságát olyan eggyel (ha nincs ilyen, akkor kettővel, ha ilyen sincs, akkor hárommal stb.) rövidebb kerethez *adjuk hozzá*, mely illeszkedik az eredeti keretre. A rövidebb keret akkor illeszkedik, ha bővítményi pozícióinak halmaza az eredeti keretének részhalmaza, és ahol az eredeti keret LKB-t tartalmaz, ott a rövidebb keretben nincs eltérő konkrét szó. Az 'ige=von -0 -t=váll' 3 hosszúságú keret például illeszkedik az 'ige=von -0=szaxofonos -t=váll' 4 hosszúságú keretre.
4. Végül a megmaradó szerkezeteknek a (3. lépésben leírt módon számított) kumulatív gyakoriság szerint rendezett listája adja az összegyűjtött igei szerkezeteket.

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 tárgyi fej a *váll* szó, az alanyi fejek viszont sokkal variábilisabbak az ilyen 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 'ige=von -0 -t' pedig azért nem „nyelheti el” az összes ilyen mondatot, mert *két* egységgel rövidebb az „A szaxofonos vállat vont.” típusú mondatoknál, így azoktól közvetlenül nem tud gyakoriságot örökölni. Az algoritmus pontossága megfelelő, a lexikálisan kötött bővítményt is tartalmazó szerkezetek esetében 80% fölötti arányban eredményez idiomatikus értelmű igei szerkezeteket (Sass, 2009a). Eredményként előállnak az egyes igei szerkezetek a gyakorisági mérőszámukkal együtt. A *vet* igéhez tartozók a 2. ábrán láthatók.

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]

2. ábra: A vet igéhez tartozó szerkezetek. Az LSzB-ket a rag jeleníti meg, az LKB-kben a tartalmas szó félkövérrel, a rag kötőjellel hozzákapcsolva jelenik meg, -A jel a birtokos személyragot jelenti. Szögletes zárójelben a szerkezethez tartozó gyakorisági mérőszám.

Fontos kiemelni, hogy a korábbi fejezetekben ismertetett klasszikus nyelvelemző modulokkal ellentétben az igei szerkezeteket gyűjtő algoritmus már egy valódi specifikus lexikográfusi részfeladatot – az anyaggyűjtést – vált 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 automatikus anyaggyűjtés kiküszöböli ezeket a problémákat: a lexikográfus keze alá dolgozva összegzi a korpuszban található információt.

3.5. A szerkezetek automatikus rendszerezése

A jellegzetes igei szerkezetek között számos olyan van, amely egy másik szerkezet specializációjának tekinthető. 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 érezhetjük, 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. A kérdés az, hogy hogyan jelenítsük meg a szótárban a bonyolult specializációs viszonyokat, miközben az eredeti gyakorisági szempontunkra 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”) szerkezetben *pontosan egy* olyan bővítményi pozíció van lekötve, amely az általánosabb szerkezetben lexikálisan szabad (LSzB), és a specializált szerkezet gyakorisági mérőszáma kisebb. 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. Azonban ez nem mindig van így, ilyenkor a lexikográfus felülbírálhatja az automatikus rendszer döntését, mégpedig azáltal, hogy idiómának (önálló jelentéssel bíró új szerkezetnek) jelöl meg egy szerkezetet, és ezzel az alárendelt szerkezetet áthelyezi a felső szintre, a gyakoriság szerinti megfelelő helyére.

Az alárendelt (csak gyakori szót példázó) szerkezetek megjelenítésére két lehetőség van. Egyrészt megjeleníthetjük önálló egységként, valahogyan (pl. beljebb szedéssel) jelezve, hogy az általánosabb szerkezet alá tartozik:

alkalmaz -t [3209]
  alkalmaz **módszer**-t [278]

A másik lehetőség, hogy csak felsoroljuk a jellemző szavakat az általánosabb szerkezetnél. Ez a megoldás talán jobban mutatja, hogy az alárendelt szerkezetek csupán gyakori, jellemző változatai az általánosabb szerkezetnek:

alkalmaz -t [3209] / **módszer**-t [278]

Mindkét esetben érvényes marad, hogy minden szerkezet a saját jogán rendelkezik gyakorisági mérőszámmal, azaz jelen esetben a 278 a 3209-en felül értendő.

3.6. Példagyűjtés

Az automatikus szakasz utolsó lépéseként példákat gyűjtünk az egyes szerkezetekhez. Minden szerkezethez olyan példamondatokat (valójában a reprezentációból következően példa-tagmondatokat) rendelünk, amelyek illeszkednek a szerkezethez: az ige mellett megfelelő LSzB-k és LKB-k vannak jelen benne. Ilyen példákat nyilván egyszerűen találhatunk a korpuszunkban, amelyből 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 leggyakoribb 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 következőkben bemutatandó korpuszlekérdező eszköz automatikus használatával valósul meg.

3.7. A korpuszlekérdező eszköz

A szintaktikailag elemzett korpuszhoz elkészült egy, a korpuszreprezentációnak megfelelő speciális korpuszlekérdező eszköz: a *Mazsola* (Sass, 2007). Regisztráció után szabadon hozzáférhető a http://corpus.nytud.hu/mazsola címen. Alapvető funkciója, hogy bemutassa a keresett ige leggyakoribb bővítményeit, bővítménykeretét, az ige mellett adott toldalékkal előforduló legjellegzetesebb kollokátumokat (3. ábra). A kollokátumokat – az ún. *salience* (Kilgarriff & Tugwell, 2001) mértékkel mért – jellegzetességük szerint sorba rendezve prezentálja. Amellett tehát, hogy a Mazsola egy önálló nyelvészeti kutatóeszköz igék és bővítmények, illetve igei szerkezetek korpuszvezérelt tanulmányozására,

jelen munkálat manuális szakaszában is nagy hasznát vettük. Segítségével bármikor ellenőrizhettük az igei szerkezeteket, a hozzájuk tartozó összes korpuszpéldával együtt, illetve további példákat kereshettünk segítségével a korpuszban, mikor a rendszer által felkínált példamondatok közül egyik sem volt megfelelő. Az alábbiakban konkrét példán mutatjuk be az eszköz működését.

A 3. ábrán látható a Mazsola felülete. A képernyő bal felső mezőjében választhatjuk ki a lekérdezni kívánt korpuszt, ez alatt tüntethetjük fel a vizsgálni kívánt igetövet. Alatta, három sorban a kívánt bővítmény(ek)et specifikálhatjuk csak esetrag/névutó vagy esetrag/névutó és szótő megadásával. (A 'Nem' jelölőnégyzet megjelölésével a találati halmazból kizárni kívánt elemeket választhatunk ki.) A legalul látható 'Szó' mezőben szabadszavas kereséssel szűkíthető a vizsgálat. Ha a 'Teljes mondatlefedés'-t jelöljük meg, csak azokat a 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' gomb segítségével azt állíthatjuk be, hogy melyik bővítményi pozícióban megjelenő jellegzetes szótövek listáját kérjük. A lekérdezőfelület alatt látjuk az eredményt: a kérdezett bővítményként megjelenő tipikus szavakat jellegzetesség szerint csökkenő sorrendben. A szavakra kattintva a megfelelő konkrét korpuszbeli példamondatokhoz jutunk.

A program eredményeinek elemzésével megvizsgálhatjuk az igék legjellemzőbb bővítményeit. Példánkban a *köt vmit vmihez* szerkezet jellegzetes tárgyragos (3. ábra) és jellegzetes *-hOz* ragos (4. 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. *megköszörüli a torkát*, *mosolyt fakaszt*, *a gyanú árnyéka sem vetődik rá*, *üsse kő*, *hoz a konyhára* stb.) – 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, történetileg korábbról adatolható jelentést. Erről a kérdésről részletesebben lásd Hanks (2005).

Az idiómák és szólások azonosítása után megvizsgálva az eredményeket, és az egyes bővítményi helyeken megjelenő szavakból szemantikai csoportokat képezve feltérképezhetjük a különféle igei szerkezeteket, illetve a szerkezetek jelentéselemeit. A *köt vmit vmihez* szerkezet esetében a szó szerinti jelentés (*kutyát fához*) gyakoriságát jóval meghaladja azaz 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.

3.8. Megbízhatóság

Amint láttuk, az automatikus feldolgozó lépések egyike sem tökéletes, a nyelvtechnológiában 100%-os pontosságot elérni lényegében lehetetlen. Bár emiatt 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. A statisztikai alapú általános állítások igazságára az alkalmazott eljárásban előforduló ritka hibák nincsenek számottevő hatással (Teubert, 2005; Kilgarriff, *et al.*, 2004).

4. A manuális lexikográfiai munka

A ma szokásos szótáríró rendszerekben a korpuszkezelés és szócikkek szerkesztése két elkülönülő alrendszert alkot, a szótáríró „viszi át” a manuálisan kiválasztott nyelvi adatokat a korpuszlekérdezőből a szerkesztőprogramba. A mi módszerünk egy ennél fejlettebb megközelítést képvisel: fontos kiemelni, hogy az automatikus szakasz végén (ld. az 1. ábra felső részét) maguk a nyers szócikkek (igék köré rendezett igei szerkezetek) állnak elő. Ezek 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 (pl. az XMetal-lal) végezheti a manuális lexikográfiai munkát[2]. 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.: „Törlendő”) alapján.

Láttuk, hogy a jellegzetes igei szerkezeteket gyűjtő algoritmus (3.4. rész) az egyes igei szerkezetekhez kumulatív gyakorisági mérőszámot is rendel. A szótár készítése során alapvető a korpuszvezérelt gyakorisági szempont, csak azok a szerkezetek kerülnek be, melyek a korpuszban kellő gyakorisággal előfordulnak. A cél az volt, hogy hozzávetőlegesen 2000 ige szerepeljen a szótárban, ennek érdekében az egységes gyakorisági küszöböt 250-nek választottuk: azaz pontosan azokat a szerkezeteket vettük be, melyek a korpuszban legalább 250-szer megtalálhatók. Így 2347 ige 6854 szerkezete alkotja a nyers szótárat az automatikus szakasz végén. Ezek típus szerinti megoszlása a 2. táblázatban látható.

2. táblázat: A nyers szótár igei szerkezeteinek megoszlása.

típus	**példa**	**db**	**%**
1 LSzB	*foglalkozik -vAl*	2808	41%
2 LSzB	*ad -t -nAk*	1166	17%
1 LKB	*von váll-t*	1138	17%
1 LKB + 1 LSzB	*húz haszon-t -bÓl*	923	13%
puszta ige	*történik*	631	9%
egyéb	*hajt malom-A-rA víz-t*	188	3%
		6854	100%

A lexikográfus feladata, hogy eldöntse, hogy az adott szerkezet valóban létezik-e, vagy csak valamilyen automatikus lépés hibás működése folytán jelenik meg; hogy alkalmas példamondatot válasszon; valamint hogy eldöntse a szerkezetekről, hogy önálló jelentéssel bíró, idiomatikus szerkezetek-e. Ezekről a feladatokról lesz szó az alábbiakban. A lexikográfiai munkát Pajzs Júlia és Kiss Margit végezte.

4.1. Létezik-e a szerkezet?

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. Fontos, érdemi munka annak eldöntése, hogy a program által felkínált szerkezetek valóban léteznek-e.

Gyakori hibaforrás a morfológiai elemzést követő egyértelműsítés (ld. 3.1. rész) pontatlan volta: mivel sok esetben (például az ikes igék egyes szám első személyében) a határozott és határozatlan tárgyas ragozás egybeesik, a szintaktikai elemzés pedig nem csupán az explicit tárggyal rendelkezőket tekinti tárgyggyal rendelkező igének, hanem azokat is, amelynek igeragja határozott tárgyas alakú, gyakori hiba, hogy tárgyatlan igéknél is tárgyas szerkezetet feltételez a program.

Sokszor nehezen dönthető el, létezőnek tekintsük-e a program által felkínált szerkezetet. Az egyik legtöbb fejtörést okozó a puszta igés szerkezet (pl. *bejön vmi*): elvben ilyenkor az igén kívül csak alanyt tartalmaz a mondat. Ilyenkor, ha számos példamondat alapos áttanulmányozását követően meggyőződünk róla, hogy az állítmánynak mindig van valamilyen határozói bővítménye (mód, idő, hely stb.), csak formailag ezek nem annyira egységesek, hogy a határozókon megjelenő toldalék (*-n, -kor, -bAn*) kellő gyakoriságban külön szerkezetbe so-

rolja őket, külön megjelöltük a szerkezetet. Ezeket nem számítottuk az elfogadott, helyes szerkezetek közé, de valószínűleg helyet fognak kapni a végső szótárban.

Az, hogy a lexikográfusok az igei szerkezetek mekkora hányadát találják elfogadhatónak, a teljes automatikus szakasz minőségéről ad fontos információt. A munkálat kiértékelésekor a következő eredményt kaptuk: a lexikográfusok a 6854 igei szerkezet közül 6243-at fogadtak el jónak, 369 igei szerkezetet hibásnak ítéltek, illetve 121 esetben egy igéhez tartozó valamennyi szerkezetet (összesen 242-t) hibásnak ítéltek. Utóbbi esetben általában az igetőazonosítás volt rossz. Az automatikus szakasz pontossága tehát 6243 / 6854=91,1%. Elmondhatjuk tehát, hogy az automatikusan előállított nyers szócikkek jó minőségűek, a lexikográfusok viszonylag ritkán találkoznak hibás szerkezettel.

4.2. Példaválasztás

A nyers szócikkek megfelelő számú, formailag illeszkedő példamondatot (példa-tagmondatot) tartalmaznak (ld. 3.6. rész). A lexikográfus feladata, hogy ezek közül a legjobbnak tűnőt kiválassza. Amennyiben egyik felkínált példamondat sem tűnik igazán alkalmasnak, lehetőség van arra, hogy a Mazsola korpuszlekérdező (3.7. rész) manuális használtatával jobbat keressen, és azt illessze be a példák közé.

Az alkalmas példamondatok kiválasztásához (Kilgarriff*, et al.*, 2008) nyomán az alábbi szempontrendszert alakítottuk ki[3]:

- az ige ne igenévi vagy továbbképzett (pl. *-hAt* képzős) formában szerepeljen a példamondatban;
- ha tárgyas szerkezetet szemléltetünk, lehetőleg valóban legyen explicit tárgy is a mondatban, ne csupán az igei személyrag fejezze ki, hogy van tárgya az igének;
- lehetőleg explicit alanya is legyen az igének;
- az igének az adott mondatban/tagmondatban lehetőleg ne legyen más bővítménye, mint az a szerkezet, amelynek illusztrálására kiválasztottuk;
- az illusztrálandó toldalékok lehetőleg tartalmas szavakon szerepeljenek, ne pusztán névmásokon;
- az LKB-t tartalmazó szerkezetekben kiemelten szereplő kollokátum lehetőleg ne szerepeljen a kollokátumot nem tartalmazó szerkezet-variáns példamondatában;
- lehetőleg válasszunk olyan példamondatot is, amelyben a szó eredeti, konkrét jelentésében szerepel (*vetem a magot; a tó vizébe vetette magát*);
- a kiválasztott példamondatokban lehetőleg kerüljük az aktuális politikára utaló közismert tulajdonneveket (pártok, ismert politikusok nevei) és általában igyekezzünk elkerülni az olyan példamondatok kiválasztását, amely a társadalom valamelyik csoportja/rétege számára sértőként is értelmezhető;

- bár a használt korpusz jelentős része a sajtónyelvből származik, törekszünk arra, hogy minél változatosabb nyelvhasználatot illusztráljunk.

Ezek a szempontok a szócikkek készítése és kölcsönös ellenőrzése közben folyamatosan alakultak ki. Azt tartottuk szem előtt, hogy az általunk feltételezett felhasználók: szótárírók, nyelvészek, nyelvtanárok, nyelvtanulók jó eséllyel találhassanak olyan példamondatokat szótárunkban, amelyeket maguk is felhasználhatnak, illetve kedvet kapjanak a Mazsola korpuszlekérdező önálló használatához. A szempontok egy része alapján automatikusan is szűkíthető, pontosítható lenne a felkínált példamondatok halmaza: például *-hAt* toldalékúak és főnévi igenévi alakok mellőzése a példamondatok közül, csak explicit tárggyal rendelkező példamondatok felkínálása stb. Érdemes lehet az atomatizálható szempontokat a későbbiekben beépíteni a példagyűjtő eljárásba (ld. 3.6. rész), és ezzel a manuális munka újabb részletét automatizálni.

4.3. Idiómák megjelölése

A 3.5. részben említettük, hogy a lexikográfus felülbírálhatja a szerkezetek automatikusan kialakított alárendeltségi viszonyait azáltal, hogy idiómának jelöl meg egy szerkezetet. A kritérium a következő: ha egy szerkezet egy LSzB-jének lekötésével önálló jelentéssel bíró új szerkezet jön létre, akkor a létrejött szerkezet idióma; ha az LSzB-ben csak egy ott tipikus, gyakori szó jelenik meg új jelentés nélkül, akkor az nem idióma. A *kijön* puszta igéhez képest a *kijön a lépés* idióma, a *vesz rész-t -n* szerkezethez képest a *vesz rész-t tárgyalás-n* viszont nem. Az idiómák tehát mindig komplex igék. A munkálatok végén azt találtuk, hogy a szerkezetek 15%-a (923 darab) volt idióma. A szótárban az idiómák valószínűleg a legfelső szinten fognak megjelenni, és feltüntetjük mellettük azt az információt, hogy idiómák.

4.4. Költségigény

Megközelítésünk lényegi pontja, hogy az automatikus szakaszban (1. ábra) 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 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. További előny, hogy a szótári munkálatok a bármikor folytathatók, a szótár kiegészíthető anélkül, hogy a már elvégzett munka kárba veszne. Egyszerűen csökkentjük a gyakorisági küszöböt, majd feldolgozzuk és beillesztjük az ezáltal bekerülő szerkezeteket.

5. A szótár végső formája

A kész szótár szócikkeinek szerkezetét a 5. ábrán látható példa mutatja be. A hagyományos szótári megjelenítéshez hasonlóan az igei szerkezeteket igék köré csoportosítva prezentáljuk.

vet [15728]
vet -nAk **vég**-t [ID] [1463] vessen véget az erőszaknak,
vet **szem**-**A**-rA -t [ID] [805] Hasonló diszkriminációkat vetnek az albán hatóságok szemére,
vet -rA **pillantás**-t [ID] [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 [ID] [377] vesse papírra az új problémákat.
vet **szám**-t -vAl [ID] [297] Vessünk számot eddigi politikánkkal,
vet -rA **fény**-t [ID] [267] ez rossz fényt vet az edzők nevelőmunkájára.
vet -bA -t [252] a tó vizébe vetette magát.

5. ábra: Példaszócikk a kész szótárból.

Az alapigét követi a Mazsola lekérdező által szolgáltatott előfordulási száma. Ezután a gyakoriság csökkenő sorrendjében tüntetjük fel a tipikus szerkezeteket, amelyben az alapige előfordult; az esetleges idióma [ID] megjelöléssel együtt. Szögletes zárójelben a szerkezet gyakorisági mérőszámát láthatjuk, majd a példamondat következik. Látjuk, hogy a szócikkbe csak az említett 250-es küszöbértéknél gyakoribb szerkezetek kerültek be. A példa jól illusztrálja, hogy a komplex igék milyen változatos formában jelennek meg.

Az új jelentést nem hordozó kollokátumokat tartalmazó szerkezeteket – ha gyakoriságuk az általánosabb szerkezetnél kisebb – az általánosabb szerkezet „alá” rendelve tüntetjük fel, a 3.5. rész végén leírt egyik módon. A példaszócikkben azonban ilyen nincs, mert bár a '*vet* **papír**-rA -t [377]' és a '*vet* -rA **fény**-t [267]' kisebb gyakoriságuk miatt a '*vet* -rA -t [380]' alá tartoznának, mégis a felső szintre kerülnek, mivel önálló jelentéssel bíró idiomatikus szerkezetek (ld. 4.3. rész). Az általános keretüknél gyakoribb szerkezeteket (pl. '*vet* -rA **pillantás**-t [708]') mindenképp kiemeljük, ezeknek gyakoriságuknál fogva tulajdonítunk az általános keretüknél nagyobb jelentőséget, még akkor is, ha esetleg nem idiomatikusak.

Az hagyományos módon, alapige szerint betűrendbe rendezett prezentáció mellett a szótár számos mutatót is tartalmaz, ezek az XML alakból emberi beavatkozás nélkül automatikusan generálhatók. A *gyakoriság szerinti mutató* szigorúan véve maga a tanulmány címében szereplő gyakorisági szótár. Igétől függetlenül egyetlen gyakorisági listába rendezve mutatja be a szerkezeteket. Az *általános bővítménykeretek szerinti mutató* az igék mellett megjelenő bővítményi kombinációkat (pl.: *-tÓl -ig* vagy *-t -vAl*) listázza, felsorolva a hozzájuk tartozó igéket. Segítségével, szándékunk szerint, izgalmas grammatikai és szemantikai elemzések készíthetők majd. Áttekinthetjük például, melyek azok az igék, amelyek egy vagy több általános bővítménykerettel gyakran fordulnak elő, csoportokat képezhetünk olyan igékből, amelyek több szerkezetben is jellemzően előfordulnak. A *kötött szavak szerinti mutató* azokat a névszói fejeket listázza, melyek igei szerkezetek lexikálisan kötött bővítményeiben előfordulnak. Ez a mutató a szótár „kifordításának" tekinthető, mivel az alapfunkcióval szemben – ti. hogy adott ige milyen névszókkal kollokál igei szerkezeteket alkotva –, itt arra kereshetjük a választ, hogy adott névszók milyen igékkel állnak szoros kapcsolatban. Segítségével vizsgálhatjuk a hasonló névszók mint bővítmények mellett előforduló igéket, így képet kaphatunk arról, mely igék tűnnek a gépi elemzés számára leghasonlóbbnak szintaktikai és/vagy szemantikai szempontból. A fentiek az igekötők automatikus különválasztása után kiegészíthetők az *igekötőmentes alapigék szerinti mutató*val (itt a *szembeállít -t -vAl* szerkezet az *állít* igéhez kerülne), valamint az igekötővel kiegészített általános kereteket (pl.: *fel -rA, ki -bÓl*) tartalmazó *igekötős keretek szerinti mutató*val.

6. Nyelvfüggetlenség

A szótár bemutatása után kitérőt teszünk, ebben a fejezetben az alkalmazott automatikus eszközök (ld. az 1. ábra felső részét) nyelvfüggetlenségét vizsgáljuk meg. Az automatikus eljárásoknak általában is külön jelentőséget ad, ha valamilyen mértékben nyelvfüggetlenek tudnak lenni. Ilyenkor kis munkabefektetéssel egyéb nyelvekre az eredetihez hasonló eredményeket lehet segítségükkel elérni.

Az alkalmazott automatikus eszközök két részre oszthatók. A klasszikus nyelvelemző eszközök – a morfológiai elemző és egyértelműsítő (3.1.), a tagmondatra bontó (3.2.), és a szintaktikai elemző (3.3.) – nyilvánvalóan nyelvfüggők. Ezek azonban sok nyelvre már elkészültek, illetve a szükséges minőségben nem túl jelentős ráfordítással létrehozhatók. Várható, hogy az alapvető nyelvtechnológiai eszközkészlet (Basic Language Resource Kit, http://www.blark.org) részeként néhány éven belül számos nyelvre rendelkezésre fognak állni.

A szintaktikailag elemzett korpuszra épülő további automatikus eszközökről pedig – kiemelendő a jellegzetes igei szerkezeteket gyűjtő algoritmus (3.4.) és a példagyűjtésben is használt korpuszlekérdező eszköz (3.7.) – az alábbiakban

megmutatjuk, hogy minden bizonnyal nyelvfüggetlenek. A korpuszreprezentáció (ld. a 3.3. rész végét) lényegében mindössze annyit feltételez a kezelendő/feldolgozandó nyelvről, hogy van benne predikátum-argumentum struktúra; más szóval, hogy tagmondatai igék köré szerveződnek, és ezeknek az igéknek a dependensei bizonyos névszói csoportok, amiknek az igéhez való viszonyát valahogyan (magyarban pl. felszíni szinten az esetragokkal) meg lehet ragadni. A jellegzetes igei szerkezeteket gyűjtő algoritmus működésének egyetlen feltétele, hogy ilyen reprezentációjú bemenő korpuszt kapjon.

Teoretikus alapon tehát azt várjuk, hogy a módszer tetszőleges nyelvre működtethető. Ezt támasztják alá a dán nyelv tekintetében elvégzett alábbi vizsgálatok. A 300000 szavas dán függőségileg elemzett korpuszon (Trautner Kromann, 2003) megmutattuk, hogy reprezentációnk egyszerű szabályokkal létrehozható (Sass, 2009b). A szerkezeteket gyűjtő algoritmus lefuttatásakor a korpusz kis mérete miatt 5 helyett 2-es küszöböt alkalmaztunk (ld. 3.4. rész, 3. lépés). Az eredményben azt tapasztaljuk, hogy komplex igék (pl. *få lov til* 'engedélyt kap', *have brug for* 'szüksége van vmire') a kis korpuszméret miatt sajnos nem jönnek ki, de az eredmények bíztatók, amint azt a 6. ábrán látható két nyers szócikk mutatja.

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

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

Látható tehát, hogy az automatikus szakasz második felében található eljárások nyelvfüggetlenek. Úgy is fogalmazhatunk, hogy módszerünk alkalmazásának feltétele tagmondatokra bontott, szintaktikailag megfelelően elemzett korpusz, *vagy* az ennek előállításához szükséges morfológiai elemző, tagmondatra bontó és szintaktikai elemző modul megléte. Mivel utóbbiak sok nyelvre már most rendelkezésre állnak, úgy véljük, hogy a módszer sikerrel adaptálható más nyelvekre is.

7. Alkalmazások

A létrejövő „Igei szerkezetek gyakorisági szótára” az alapvető definíció nélküli szótári funkciókon (ld. 2.3. rész) kívül az alábbi célokra használható fel. Segédeszköz lehet más *lexikográfiai* munkák készítésekor, melyek a magyar nyelvvel (is) foglalkoznak. Korpuszból nyert adatokat foglal össze, manuálisan ellenőrizve és javítva, alkalmas korpuszpéldákkal; így kiegészítve a hozzá tartozó korpuszlekérdező eszközzel (3.7. rész) gazdag korpusznyelvészeti *kutatóeszköz*. Nyelvi adatokat felhasználó *kísérletekben* (például a pszicholingvisztikában) fontos szempont a gyakoriság, ehhez adatokat szolgáltat a szótár. Értékes lexikális erőforrás, amit a *nyelvtechnológia* számos területe hasznosíthat az információ visszakereséstől a gépi fordításig. Hasznos lehet olyan nyelvtanároknak, kutatóknak, akik nyelvtanítási célú *tananyagot* készítenek. Itt szintén nyilvánvalóan fontos szempont a gyakoriság, illetve a különféle igei szerkezetek.

A fentieken túl az ISZGYSZ egyfajta haladó nyelvtanulóknak szóló *tanulói szótár* (learner's dictionary), mely lehetővé teszi, hogy a nyelvtanulók „idiomatikusan írhassanak és beszélhessenek” (Hanks, 2008), azaz valóban úgy fogalmazzák meg a mondanivalójukat, ahogy azt egy magyar tenné. Könnyen meghatározhatjuk belőle azokat a névszókat, amelyek adott igével komplex igét vagy kollokációt alkotnak (pl. *állít -nAk emlék-t* vagy *állít bíróság·elé -t* stb.), de az 5. részben említett kötött szavak szerinti mutató segítségével a fordított irányú kollokációs kapcsolatokat is számba vehetjük, így kiderül, hogy a magyarban az adott névszó melyik igé(ke)t „szereti” és ezekkel az igékkel milyen kifejezéseket alkot, amint ez a 3. táblázatban látszik.

3. táblázat: Névszók igei kollokációi.

követelmény	→	megfelel követelmény-nAk
ajándék	→	kap ajándék-bA -t
bizonyíték	→	van bizonyíték -rA
búcsú	→	vesz búcsú-t -tÓl

Annak a döntésünknek, hogy nem csak idiomatikus szerkezeteket, hanem kompozicionális kollokációkat is közlünk, nagy előnye, hogy képet kapunk a névszók kollokációs viselkedéséről is, azaz pl. hogy a *szerződés*sel „mit lehet tenni”, a *szerződés* minek a bővítménye szokott lenni. Itt jegyezzük meg, hogy bár a Mazsola program elsődlegesen az igék tipikus bővítményeinek vizsgálatára készült, kereshetjük vele adott névszókhoz mint bővítményekhez tartozó igéket is. Ha a felületen (ld. a 3. ábrát) szótőként a *szerződés* szót adjuk meg, és az 'Eloszlás' gombot az (üresen hagyott) igető mező mellé állítjuk, az eredmény-

ben legelöl a *köt*, *megköt*, *aláír* igék szerepelnek, de mindjárt ezután következik a *felmond*, *felbont*, *lejár*, majd kicsit hátrébb a *bont* és a *felrúg* is.

8. Befejezés

Tanulmányunkban bemutattuk az általunk kifejlesztett korpuszvezérelt, félautomatikus szótárkészítési módszert, illetve az ezzel a módszerrel készülő új szótárat. A különféle nyelvelemző, lexikai kinyerő és rendszerező programok használatával a lexikográfiai munkaigényt alacsonyan tudtuk tartani. Az automatikus eszközök konkrét lexikográfusi részfeladatokat is képesek voltak átvenni. A gyakori, jellemző igei szerkezetek teljes spektrumát lefedő „Igei szerkezetek gyakorisági szótára" több lexikográfiai újdonságot is tartalmaz: alapegységei nem szavak, hanem – igét és bővítményeit tartalmazó – szószerkezetek; az anyaggyűjtés automatikusan történik, a nyers szócikkek a lexikográfus nyelvi intuíciójától függetlenül automatikusan állnak elő; autentikus korpuszpéldák világítják meg a szerkezetek jelentését; valamint a Magyar Nemzeti Szövegtárból származó gyakorisági mérőszámokat is tartalmaz. Egyebek mellett fontos segédeszköz lehet a nyelvoktatásban: igék és névszók kollokációs viszonyainak ismerete elengedhetetlen a haladó nyelvtanuló számára, mivel ezek teszik a nyelvhasználatot folyékonnyá és az anyanyelvi beszélőpartner számára is természetessé.

Az automatikus szintaktikai elemzés során a bővítményeket pusztán formai jegyek (ti. az esetragok) alapján kapcsoltuk az igékhez, ez a felszíni jegyeken alapuló megközelítés természetesen kihatással van a szótárra is. Ezért van, hogy 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 fedhet le, pl. *nyer -vAl*: *pontozással*, *lelkesedéssel*, *kisgazdákkal*; vagy *nyílik -rA*: *résnyire*, *körútra*. Az esetragok szemantikai kérdéseitől tehát eltekintettünk, illetve nem foglalkoztunk a bővítményként megjelenő szavak szemantikai osztályozásával sem. Jelentős előrelépést jelenthetne, ha a jövőben ugyanezt a módszert egy szemantikai annotációval ellátott korpuszon próbálhatnánk ki. Ide tartozna például a hely-, idő- és módhatározók automatikus felismerése, és bővítményi kategóriaként való kezelése, valamint a különböző szemantikai kategóriák, és ezáltal szemantikus alapú szerkezetek (pl. *vág* ÉLŐ-*hOz* ÉLETTELEN-*t*) azonosítása.

A tanulmány célja az volt, hogy megmutassuk, hogy módszerünk értékes szótár létrehozására alkalmas. Nyitva áll az út a további erre épülő fejlesztések előtt: a szemantikus annotációval bíró korpusz használata mellett, nyilván valamennyi automatikus eszköz teljesítményén lehet javítani (ld. például a morfológiai egyértelműsítés hibájából adódó problémát a 4.1. részben, vagy példaválasztás kiterjedtebb automatizálására vonatkozó javaslatokat a 4.2. részben), a legígéretesebb irány mégis a módszer más nyelvekre, esetleg szaknyelvekre való adaptálása lehet.

Jegyzetek

1. Megjegyzendő, hogy a modern kollokációs szótárak figyelmet fordítanak a vonzatok gondos feltüntetésére (Forgács, 2003; Bárdosi 2009).
2. Köszönet Mártonfi Attilának az XMetal használata során nyújtott segítségéért.
3. A szempontrendszer kidolgozása jelentős mértékben Kiss Margit érdeme.

Irodalom

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

Bárdosi V. (2009) Magyar szólások, közmondások értelmező és fogalomköri szótára. Budapest: Tinta Könyvkiadó.

COBUILD. (1987) Collins Cobuild English Language Dictionary. London: Harper Collins Publishers.

Forgács T. (2003) *Magyar szólások és közmondások tára*. Budapest: Tinta Könyvkiadó.

Forgács T. (2007) *Bevezetés a frazeológiába*. Budapest: Tinta Könyvkiadó.

Goldberg, A. E. (2006) *Constructions at Work*. Oxford: Oxford University Press.

Grefenstette, G. (1998) The future of linguistics and lexicographers: Will there be lexicographers in the year 3000? *Proceedings of EURALEX 1998*, Liège. pp. 25-41.

Hanks, P. (2001) The probable and the possible: Lexicography in the age of the internet. *Proceedings of AsiaLex 2001*, Seoul: Yonsei University. pp. 1-15.

Hanks, P. (2005) Metaphors and meanings: a lexicographical approach to corpus analysis. In: Kiefer, F., Kiss, G. & Pajzs, J. (eds.) *Papers in Computational Lexicography, COMPLEX 2005*. Budapest: Linguistics Institute HAS. 81-106.

Hanks, P. (2008) The lexicographical legacy of John Sinclair. *International Journal of Lexicography*, 21/3. pp. 219-229.

Janssen, M. (2008) Meaningless dictionaries. *Proceedings of the XIII EURALEX International Congress*, Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. pp. 409-420.

Kilgarriff, A. (1997) „I dont't believe in word senses". *Computers and the Humanities*, 31/2. pp. 91-113.

Kilgarriff, A. & Tugwell, D. (2001) Word Sketch: Extraction and display of significant collocations for lexicography. *Proceedings of the 39th Meeting of the Association for Computational Linguistics, workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*. Toulouse. pp. 32-38.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004) The Sketch Engine. *Proceedings of EURALEX 2004*. Lorient. pp. 105-116.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, P. (2008) GDEX: Automatically finding good dictionary examples. In: *Proceedings of the XIII EURALEX International Congress*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. pp. 425-432.

O. Nagy G. (1966) *Magyar szólások és közmondások.* Budapest: Akadémiai Kiadó.

Oravecz, Cs. & Dienes, P. (2002) Large scale morphosyntactic annotation of the Hungarian National Corpus. In: Hollósi, B. & Kiss-Gulyás, J. (eds.) *Studies in Linguistics, Volume VI*. Debrecen. 277-298.

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

Pajzs, J. (2002) A corpus based investigation of collocations in Hungarian. P*roceedings of EURALEX 2002*. Copenhagen: University of Copenhagen. pp. 831-840.

Rundell, M. (1998) Recent trends in english pedagogical lexicography. *International Journal of Lexicography*, 11/4. pp. 315-342.

Rundell, M. (2009) The road to automated lexicography: First banish the drudgery... then the drudges? Meghívott előadás az „eLexicography in the 21st Century" konferencián, Louvain-la-Neuve.

Sass B. (2005) Vonzatkeretek a Magyar Nemzeti Szövegtárban. In: Alexin Z. és Csendes D. (szerk.) *III Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2005)*. Szeged. 257-264.

Sass B. (2006) Igei vonzatkeretek az MNSZ tagmondataiban. In: Alexin Z. és Csendes D. (szerk.) *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2006)*. Szeged. 15-21.

Sass B. (2007) „Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. *I. Alkalmazott Nyelvészeti Doktorandusz Konferencia*. Budapest: MTA Nyelvtudományi Intézet. 137-149.

Sass, B. (2009a) A unified method for extracting simple and multiword verbs with valence information and application for Hungarian. *Proceedings of RANLP 2009*. pp. 399-403.

Sass, B. (2009b) Verb Argument Browser for Danish. In: Jokinen, K. & Bick, E. (eds.) *Proceedings of the 17th Nordic Conference of Computational Linguistics, NoDaLiDa 2009*, NEALT. 263-266.

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

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

Teubert, W. (2005) My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10/1. pp. 1-13.

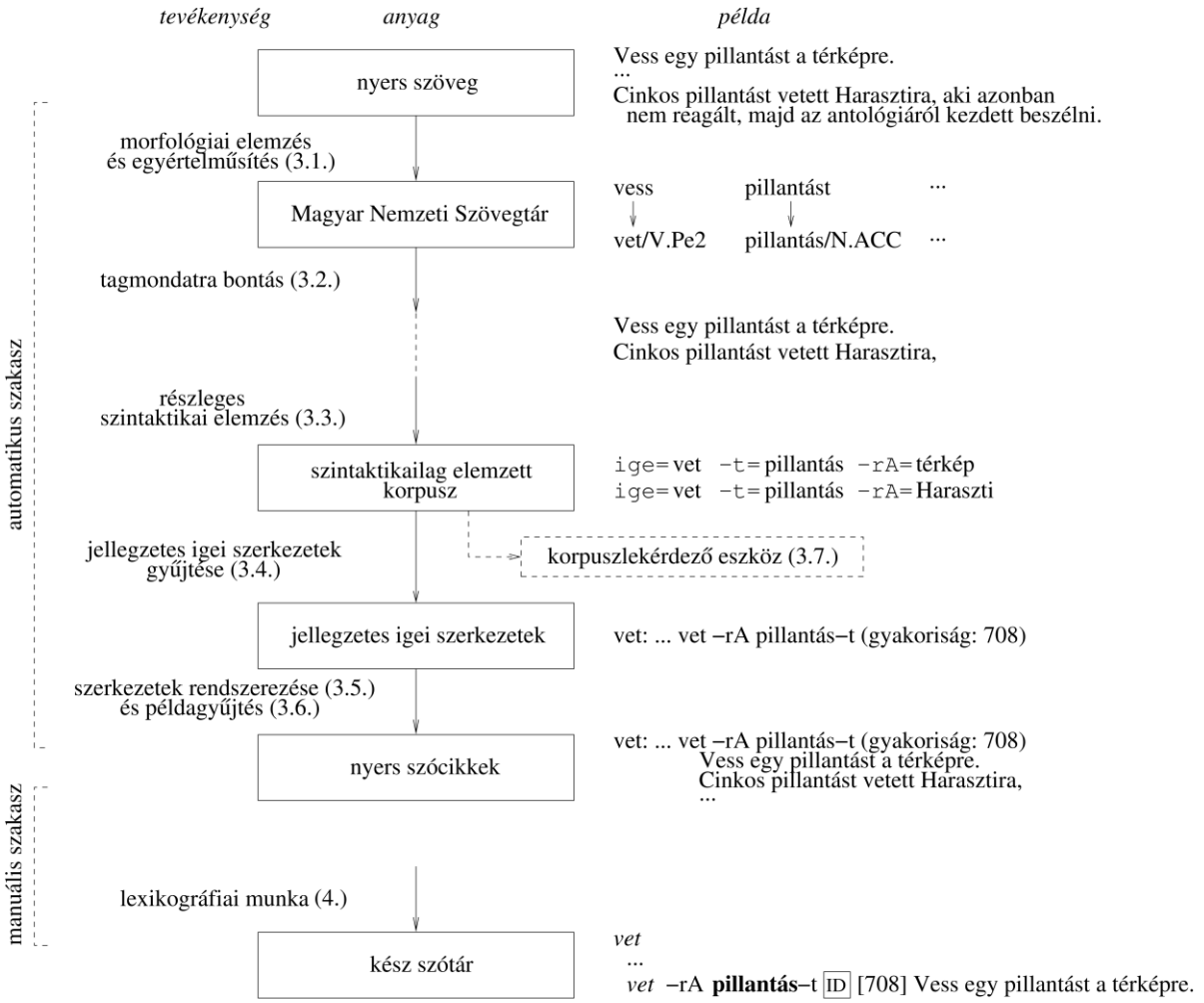
Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

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

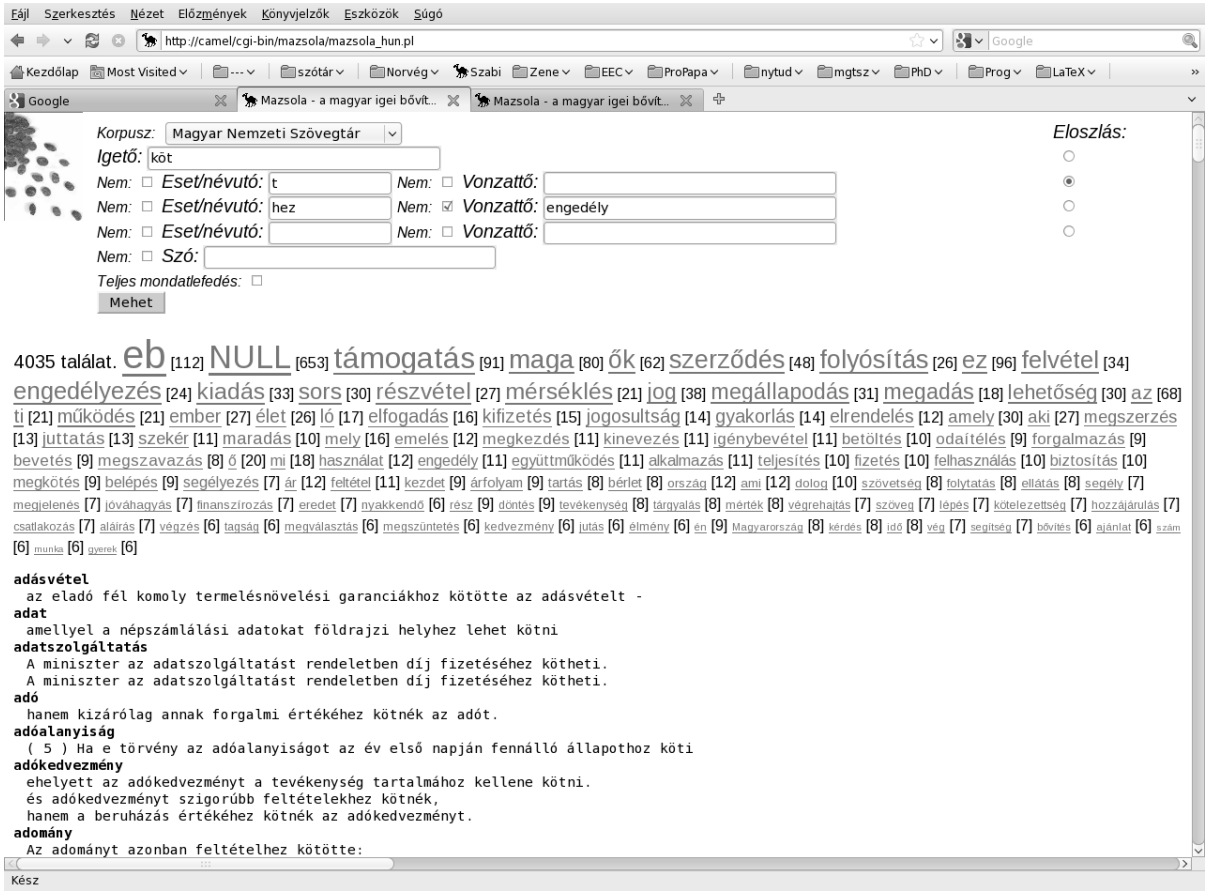
Váradi, T. (2002) The Hungarian National Corpus. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*. pp. 385-389.

Yarowsky, D. (1993) One sense per collocation. *Proceedings of the workshop on Human Language Technology*. pp. 266-271.

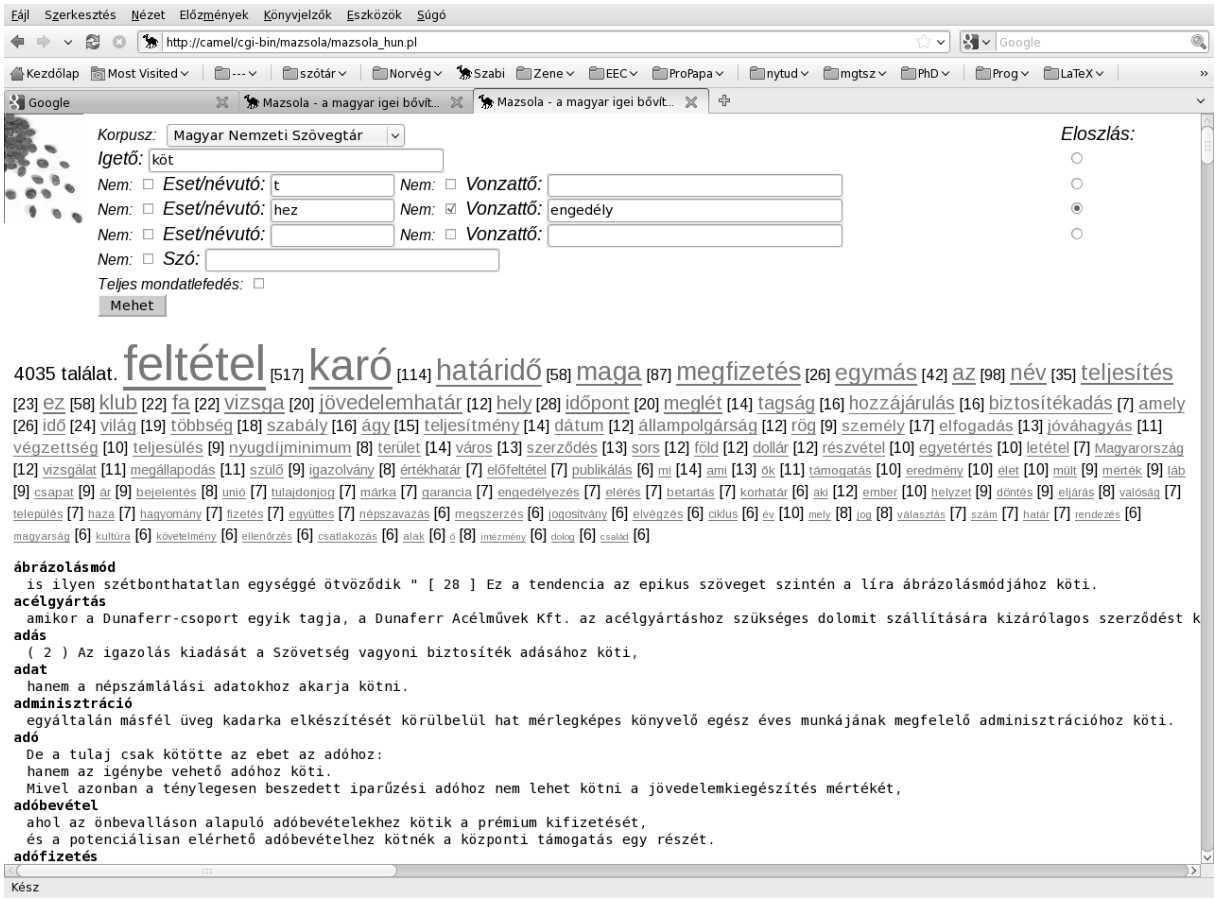
Függelék



1. ábra: A szótárkészítési folyamat. A zárójelben feltüntetett számok a tanulmány megfelelő fejezeteire utalnak.



3. ábra: A Mazsola felülete: a köt vmit vmihez szerkezet, a benne előforduló jellegzetes tárgyragos szavak listájával. (A NULL a tárgyas ragozású igével rendelkező, de explicit tárggyal nem rendelkező példamondatokat jelöli.)



4. ábra: A Mazsola felülete: a köt vmit vmihez 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.)