Nyelvtechnológia a lexikográfia szolgálatában

Pajzs Júlia

Napjaink szótáraink elkészítése és publikálása számos területen összefonódik a nyelvtechnológia eredményeivel. A tanulmányban e szerteágazó kérdéskörnek csupán néhány részletét emelem ki: korpuszépítés és elemzés, többelemű szókapcsolatok automatikus kinyerése, nyers szótári szócikkek generálása, a szótári adatbázisok szerepe a mai szótárak előállításában. Magyar egynyelvű szótárak előkészítésekor szerzett tapasztalataimat kívánom megosztani.

1. Korpuszépítés

Bár az 1980-as években a Magyar Történeti Szövegtár összeállításakor még az látszott egyedüli biztonságos megoldásnak, ha kézzel rögzítjük az előre gondosan kiválogatott szövegrészleteket, ma mind az egynyelvű, mind a többnyelvű korpuszok gyűjtésekor már számítógépes formában lévő szövegekből érdemes összeválogatni a felhasználásra szántakat. Az elsősorban az ezredforduló szóanyagát tartalmazó Magyar Nemzeti Szövegtár (VÁRADI 2002) anyaga már így állt össze. Speciális esetben − mint Mikes Kelemen írói szótárának jelenleg is folyamatban lévő készítésekor (KISS 2012) – az a célszerű, ha a feldolgozandó anyagot professzionális módon beszkenneljük, majd ellenőrizzük és javítjuk. Magyar egynyelvű szótárak készítésekor, a fent említett, már lekérdezhető állapotú Magyar Történeti Szövegtár és az újabb és bőségesebb szóanyagot tartalmazó Magyar Nemzeti Szövegtár használatán kívül, kellő körültekintéssel gyűjtött különféle internetes forrásból származó anyagokból felépíthetjük saját korpuszunkat. Tudnunk kell azonban, hogy ilyenkor feltétlenül tisztázandó a szöveg felhasználása a szerzői jog birtokosával.

2. Előfeldolgozás, szövegszavak felismerése

A különféle forrásból származó szövegek egységesítéshez kiszűrendők vagy szabványos (XML) formába átalakítandók a további felhasználás szempontjából irreleváns (pl. tipográfiai) információk. Szükséges az egységes kódkészletre való konverzió. Célszerű az

írásjelek leválasztása a szövegszavakról betűközök közbeiktatásával. A bekezdéshatárok szintén egységesen jelölendők. Ha lehetséges és szükséges az oldalhatárok jelzése, ez is az előfeldolgozás során iktatandó be.

Ha már egységesítettük a különféle forrásokból származó szövegeket, viszonylag könnyű a szövegszavak automatikus felismertetése: szövegszó jelölt lehet minden, ami nem írásjel, segédjel vagy puszta numerikus karaktersorozat. Az egységesített szövegek egyszerű feldolgozására, lekérdezésére, szövegszólista készítésére, konkordancia készítésére számos viszonylag könnyen beszerezhető és felhasználóbarát programcsomag készült. Például a magyar szövegfeldolgozásra is alkalmassá tett NooJ program, amely letölthető a nooj4nlp.net címről, vagy a többnyelvű párhuzamos korpuszok feldolgozására is alkalmas ParaConc, amely az athel.com/parallel_corpora.html címen érhető el.

3. Morfológiai elemzés, lemmatizálás, szófaji besorolás

Míg az angol szövegeknél a tő- és toldalékmorféma felismerése egyszerű művelet, mindössze néhány tucatnyi rendhagyó alakot kell önállóan kezelni, a magyar toldalékok és szótövek helyes felismertetése önmagában is komoly kihívást jelentő feladat volt. Az utóbbi évtizedekben többször, mind jobb minőségben és mind hatékonyabb változatokban elkészült Humor program (legelső változat: PRÓSZÉKY et al. 1979, későbbiekben egyebek közt: PRÓSZÉKY et al. 1992, 1994, PRÓSZÉKY 2000, NOVÁK 2003) képes felismerni a komplex, sok toldalékból – akár képzők és ragok egymásutánjából álló – magyar szövegszavakat, sőt a változó tövű szavak aktuálisan előforduló tövét és szótári szóalakját is ismeri. Első változatának működési elvéről már írtam (PAJZS 1990) bővebben.

SZÖVEGSZÓ		LEXÉMA	SZÓFAJ	TOLDALÉKOK KÓDJA
1.	ábrázolás	ábrázol	N	gerund+nom+sg
2.	ábrázolása	ábrázol	N	gerund+3+sg+pssg+ps+nom
3.	ábrázolásával	ábrázol	N	gerund+3+sg+pssg+ps+ins
4.	ábrázolással	ábrázol	N	gerund+ins+sg
5.	ábrázolt	ábrázol	A	perfpart+nom+sg
6.	ábrázolt	ábrázol	V	indef+3+past+sg
7.	ábrázoltak	ábrázol	A	perfpart+pl+nom

8.	ábrázoltak	ábrázol	V	indef+3+past+pl

1. ábra a Humor morfológiai elemzővel elemzett szövegszavak

Mint az *1. ábra* táblázatának 5-6. és 7-8. soraiban látjuk, a szövegszavak jelentős részének egynél több lehetséges korrekt elemzése van. A leggyakoribb eset, hogy már maga a toldalék többértelmű, például mert a múlt idő jelei gyakran egybeesnek a befejezett melléknévi igenév alakjaival. Az sem ritka, hogy maga a lexéma is homonim (pl. *vár*), bár ennek egyes toldalékolt alakjairól egyértelműen eldönthető, hogy főnévi vagy igei előfordulása-e a szónak – például: *várak* N+pl, *várok* V+1+sg –, de számos alak toldalékolt formában is homonim (*várnak, várat*). Folyó szövegekben igen gyakoriak az ún. homográfok, amikor különböző szótövek toldalékolt formája egybeesik (pl. *nézet*), néhány − egyébként folyó szövegekben kifejezetten gyakran előforduló − szövegszó pedig akár egyidejűleg példázhatja a homonímia és homográfia jelenségét (*volt* V+3+past+sg, A+perfpart+sg, N+sg). Az itt látható rövidítések feloldását NOVÁK (2003) tartalmazza. A jelenségkört részletesen NOVÁK – PINTÉR (2006) járja körül.

A szövegszavak felismerésének és lemmatizálásának egy igen speciális esetével kellett megbirkóznunk a *Nagyszótár* (ITTZÉS 2006, 2011) számára gyűjtött *Magyar Történeti Szövegtár* régi szövegeinek feldolgozásakor. Mivel ez a korpusz több mint két évszázad szóanyagát öleli fel, és az egységes magyar helyesírás kialakulása csupán az 1930-as évekre tehető, ugyanaz a szóalak a régi szövegekben számtalan változatban előfordul. A XVIII. század végéről, esetenként XIX. század elejéről származó szövegekben olyan régi karakterek is előfordulnak, amelyek ma már egyáltalán nem használatosak. Amikor 1985-ben elkezdtük rögzíteni a szövegeket, az látszott a legésszerűbb megoldásnak, – amit azután az azóta eltelt idő többszörösen is igazolt, – ha mind a speciális karaktereket, mind a ma is használatos magyar ékezetes karaktereket az angol ábécé betűivel és számok kombinációjával jelöljük. Az ötlet gazdájáról, Prószéky Gáborról elnevezett kódolás (PRÓSZÉKY 1985) segítségével betűhíven tudtuk rögzíteni a legkülönfélébb forrásokból származó változatos hardver/szoftver környezetben felvitt szövegeinket. Az így keletkező szövegek lemmatizálását úgy igyekeztünk megoldani, hogy kísérletet tettünk a régi karakterek és szövegszavak mai alakhoz hasonlóvá alakítására, viszonylag egyszerű szabályok alkalmazásával. Például az ’o23’-mal jelzett karakter mai formában lehet ’ö’ vagy ’ő’; a ’ts’ jelölheti a mai ’cs’-t, a ’tz’ a mai ’c’-t stb. A szabályok nem pusztán a régi karakterek átalakítására szorítkoztak; igyekeztünk figyelembe venni szabályos toldalékvariánsokat, az igekötők helyesírásának régi variánsait is.

Erről a kísérletről részletesebben egyebek közt (KISS et al. 2004)-ben számoltunk be. Röviden összefoglalva az általunk tesztelt megoldás:

– A Humor program elemzi a szövegszavakat.

– Ha egy szót nem sikerült elemezni, egy heurisztikus program átalakítja, majd újra elemezni próbálja azt.

– Ha az átalakított változat elemzése sikeres, a felismert vagy felismerni vélt alakot őrizzük meg.

2. ábra A *fekszik* ige elemzett lekérdezésének eredménye 2002-ben

Az elemzett lekérdezés előnyeit láthatjuk a *2. ábrá*n: a *fekszik* igének nemcsak a *feküd* tövű változatait találja meg a program, hanem történeti szövegben előforduló, a mai helyesírásnak már nem megfelelő *feküdgy* felszólító alakot valamint az s43-al jelölt alakot is.

Időközben É. Kiss Katalin vezetésével megindult a Magyar Generatív Történeti Szintaxis munkálata, amely, minta a nevéből is kitűnik, elsődlegesen nem szótári célra gyűjt és elemez régi magyar nyelvemlékeket, de akár a már most hozzáférhető anyag, akár ennek mintájára készülő hasonló történeti korpusz szótári felhasználása is érdekes lehet. Az e munkálathoz kötődő Régi Magyar Konkordancia program Sass Bálint munkája (corpus.nytud.hu/rmk/).

4. Egyértelműsítés

Amint az *1. ábrá*n láthattuk, a sikeres elemzés végeredménye gyakran többértelmű. Ez a probléma lényegében minden nyelvnél felmerül: míg az angol esetében főként a több szófajú szavak nagy száma és a leggyakoribb *-s* toldalék többértelműsége okozza ezt a jelenséget, a magyarban inkább a toldalékolt alakok többféleképpen elemezhető volta, nem utolsósorban a már fentebb is említett múlt idő jelének és a befejezett melléknévi igenév képzőjének azonos alakúsága, valamint számos egyéb jelenség (homonímia, homográfia) okozza ezt. A Magyar Történeti Szövegtár sikeres morfológiai elemzése után azt találtuk, hogy a szövegszavak több mint 30%-ának van egynél több elemzése. A többféleképpen elemezhető alakok egyértelműsítésére több megoldás kínálkozik:

a) kézi egyértelműsítés,

b) az ún. lokális szabályokon alapuló egyértelműsítés,

c) statisztikai módszeren alapuló egyértelműsítés.

A Történeti Korpusz feldolgozásakor rövid ideig kísérleteztünk az a) megoldással, de néhány hónap után be kellett látnunk, hogy ennek szakember-, költség- és időigénye olyan magas lenne, amely már messzemenően nem térülne meg a tényleges szótári munka szempontjából. Ezután próbálkoztunk a b) megoldással: e kísérlet eredményéről részletesebben több publikációban (PAJZS 1997a, PAIS – PAJZS 1998, PAJZS 1999) számoltunk be. Itt csak egészen röviden ismertetem az eljárás lényegét: PERL-utasítások segítségével igyekeztem néhány alapvető szabályt megfogalmazni, hogy adott többértelműségek környezetében előforduló egyéb szavak függvényében az elemző által felkínált megoldások közül melyiket válassza az egyértelműsítő eljárás. Például ilyen szabályokat használtam (a szabályok alkalmazásának sorrendje kötött volt!):

4.1. ha egy mássalhangzóval kezdődő ige és főnév közül kell választani, és közvetlenül előtte az *a* névelő található, a főnevet válassza;

4.2. ha az *az* szót magánhangzóval kezdődődő főnév vagy melléknév követi, tekintse névelőnek;

4.3. ha az *az* után egy ige következik, tekintse névmásnak;

4.4. ha ige és más szófaj közül kell választani és nincs másik ige a tagmondatban (a következő vesszőig vagy mondatzáró írásjelig terjedő részben), válassza az igét stb.

Mint látjuk, ezekben a szabályokban tulajdonképpen keverednek a grammatikai és statisztikai szempontok: általában azt lehet rájuk mondani, valószínűbb, hogy a javasolt megoldás a helyes. A szabályokat első lépésben a MULTEXT-EAST projekt keretében készített szövegkorpuszon, Orwell *1984* című regényének anyagán teszteltem, fejlesztettem. Ezt a szöveget a tesztfuttatást követően kézzel ellenőriztük, így képet kaptunk arról, az egyes

szabályok milyen gyakran kerültek alkalmazásra, és melyik milyen gyakran bizonyult helyesnek. A kísérlet eredményének részleteiről a fent említett publikációkban számoltunk be. A tesztelés és továbbfejlesztés után a teljes elemzett Történeti Korpusznak is elkészült egy olyan változata, amelyet a fenti módon egyértelműsítettünk. Az eredmény alapján készült egy nyers címszójegyzék, amely tartalmazta az adott címszó korpuszbeli előfordulásának számát, valamint első és utolsó előfordulásának évét. Egyebek közt ez a nyers címszólista segítette hozzá a Nagyszótár lexikográfusait ahhoz a felismeréshez, hogy pusztán a Történeti Szövegtárból kiindulva nem képzelhető el a Nagyszótár kellő színvonalú elkészítése, óhatatlanul szükséges az archíválisés cédulák és egyéb, időközben elektronikus formában hozzáférhetővé vált források használata is. A Történeti Szövegtár az akkori (1997) állapotában inkább az Értelmező kéziszótár korpuszalapú felújítására, bővítésére lett volna csupán alkalmas (PAJZS 1997b).

Ebben az időszakban a szótári munkálatokkal párhuzamosan megalakult az MTA Nyelvtudományi Intézetének Korpusznyelvészeti osztálya, Váradi Tamás vezetésével, mely fő céljául a korszerű korpuszgyűjtési feldolgozási módszerek alkalmazását és fejlesztését tűzte ki. Itt hamarosan össze is állt az ezredforduló elektronikus formában hozzáférhető szövegeiből a Magyar Nemzeti Szövegtár anyaga, amelynek morfológiai elemzését szintén a Humor elemző akkori változatával végezték el, a szövegek egyértelműsítésére statisztikai módszert használtak, a végeredmény pontossága megközelítette a 98%-ot (ORAVECZ 2002).

5. (Részleges) szintaktikai elemzés

Számos projekt keretében többen kísérleteztek és kísérleteznek magyar szintaktikai elemző program kidolgozásával (pl.: KIS et al. 2003, GÁBOR 2007, PRÓSZÉKY et al. 2004).

Ha a lekérdező programot legalább részleges szintaktikai elemzésen átesett szövegekhez illesztik, a szótárak forrásanyagául szolgáló lekérdezések is árnyaltabbak lehetnek. Mintául szolgálhat az angol nyelvre készített és ezt követően néhány más nyelvre is kifejlesztett WordSketches (sketchengine.co.uk) program (KILGARIFF – TUGWELL 2001), valamint az ennek tapasztalatai alapján készült DANTE-projekt (webdante.com). Ennek keretében egy olyan angol lexikai adatbázis készült, amely nagy korpusz (1,7 GigaByte) elemzett feldolgozásán alapult, elsődleges célja ugyan az angol-ír szótár angol felének előkészítése volt, de azzal a szándékkal, hogy az újonnan készülő bármilyen kétnyelvű angol szótárak erre a friss adatbázisra épülhessenek. Ehhez az elgondoláshoz részben hasonló a Patrick Hanks

által a 2012-es Budapesten rendezett EFNIL konferencián felvetett „pattern dictionary” gondolata is (minta: deb.fi.muni.cz/pdev).

Magyar vonatkozásban Sass Bálint fejlesztett a Sketchengine-hez valamelyest hasonló korpuszelemző és -lekérdező programot (SASS 2011). Első lépésként tagmondatokra bontotta a korpuszt, majd részleges szintaktikai elemzést hajtott rajtuk végre. Így foglalta össze ezt a műveletet: „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 jelenlévő határozószókat például figyelmen kívül hagyjuk. Ezek alapján a reprezentáció már kialakítható.” (SASS 2011, 37). Sass a tagmondatra bontáskor szabályalapú megközelítéssel dolgozik, a szabályokat reguláris kifejezések segítségével írja le. A névutókat és az esetragokat hasonló módon kezeli, mivel funkciójuk is hasonló. Amennyiben nem jelenik meg explicit tárgy a tagmondatban, de az igeragozás tárgy jelenlétére utal (azaz ún. határozott tárgyas az ige ragozása), a program ugyanúgy kezeli, mintha explicit tárgy lenne a tagmondatban: egy speciális NULL tárgyat feltételez. A részleges szintaktikai elemzés következtében a program eredménye megmutatja, hogy adott ige bővítményeként milyen toldalékú névszók jelennek meg gyakran, vagy akár kötelezően – ez utóbbiak a vonzatok. Ha egy adott toldalékú bővítmény konkrét szavakkal van gyakran kitöltve, ezeket a szavakat is felsorolja a gyakoriság csökkenő sorrendjében. A Magyar igei szerkezetek szótára (Sass et al. 2010) összegzi ezeknek az adatoknak leggyakrabban előforduló részét (a legalább 250 kumulált gyakorisági értékű elemeket), míg a Mazsola (corpus.nytud.hu/mazsola) programmal részletesen egyenként kérdezhetjük le az igéket (függetlenül attól, hogy a fenti szótárban megtaláltuk-e őket) és bővítményeiket. A Mazsola programot részletesen SASS (2009) ismerteti. A program által nyújtott lehetőségeket akkor tudjuk igazán értékelni, ha összevetjük más, a magyar nyelvre is használt lekérdező programokkal.

6. Lekérdező programok

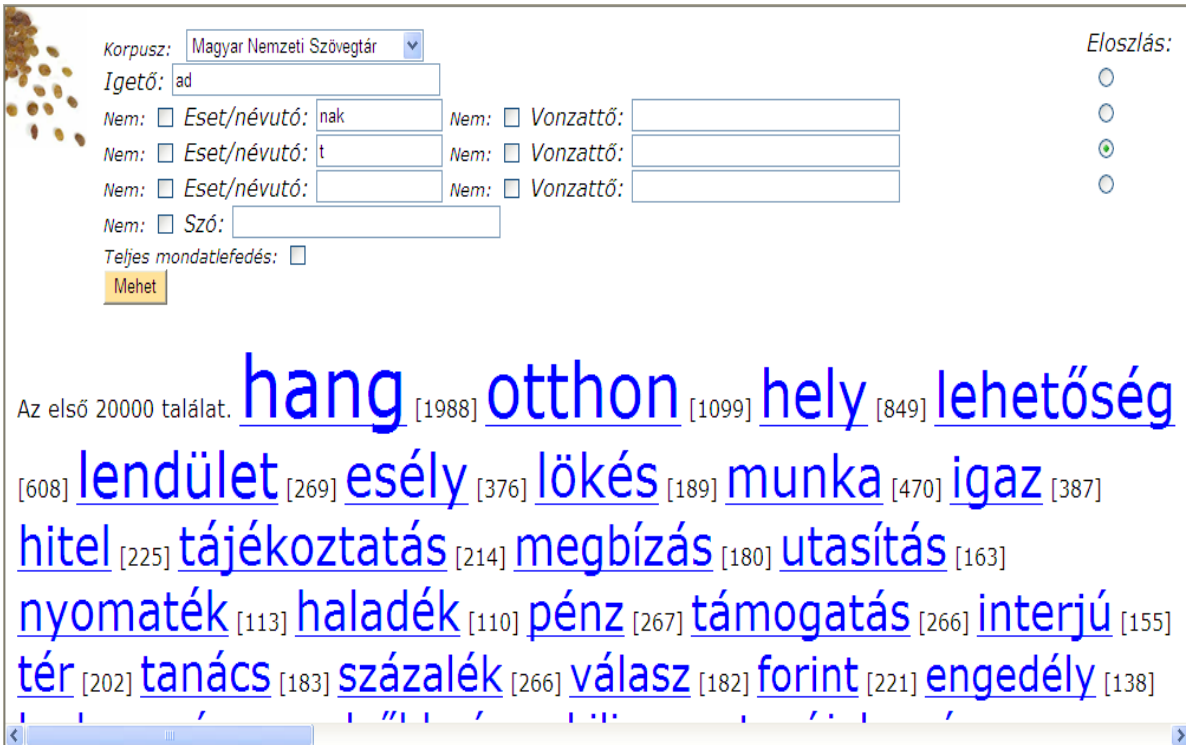
A nem elemzett korpuszokra használt lekérdező programok közül említésre érdemes a Nagyszótár készítéséhez beszerzett és adaptált − az 1980-as évek végén készült Open Text (korábban PAT) lekérdező program, amelyhez Váradi Tamás fejlesztett magyar felületet (VÁRADI – PAJZS 1997). Egyik változata ma is elérhető, ezzel kutathat a nagyközönség a

Magyar Történeti Korpuszban (nytud.hu/hhc). Ennek segítségével kereshetünk egyedi szövegszavakat, vagy kereshetünk bizonyos karakterekkel kezdődő szavakat. Kereshetjük két szó együttes előfordulását vagy kötött formában előforduló szóegyüttest. Bármelyik lekérdezést végezhetjük a teljes, 25 millió szövegszóból álló korpuszon vagy annak bizonyos részein. A keletkezés éve, a szerző személye vagy a szöveg műfaja szerint szűkíthetjük a keresést. A találatokat egyszerű konkordancialistában, vagy rövid bibliográfiával tekinthetjük át, bármely kiválasztott találatra kattintva megnézhetjük nagyobb (néhány mondatnyi) szövegkörnyezetét és pontos bibliográfiai adatait.

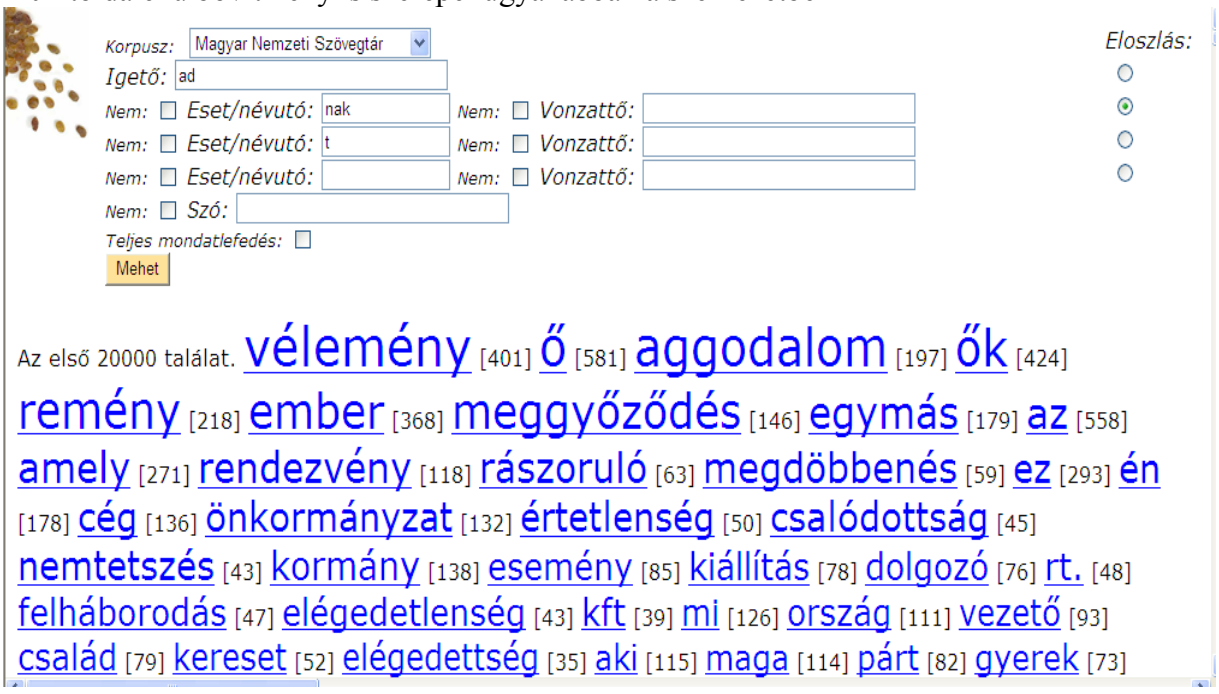
A Magyar Nemzeti Szövegtárhoz használt lekérdezőprogram motorja az IMS Corpus Workbench program, amelyhez sok tekintetben a Történeti Korpuszhoz hasonló magyar nyelvű lekérdező felületet illesztettek (VÁRADI 2002). A korpusz méretén (2012-ben 187 millió szövegszó) kívül a legfontosabb különbség, hogy itt az elemzett és egyértelműsített korpuszban kereshetünk. Így nem pusztán – esetleg egymás közelében előforduló – szövegszavakat tudunk keresni, hanem lexémákat is, megadott toldalékokkal, vagy a környezetükben előforduló szavak toldalékainak megadásával. Így nem csak a Mazsola programmal kereshetjük ki, mondjuk az *ad* ige környezetében előforduló *–t* ragos névszókat, hanem az MNSz lekérdező felületével is. A különbséget a *3.* és a *4–5. ábra* szemlélteti.

2.	még valaki át kell hogy	**adja**	*V.Te3*	a szovjet követnek a jegyzéket
3.	hogy neked, még nem	**adtam**	*V.TMe1*	a világ szájára; "
4.	csinálni belõle, fölösleges hangulatkeltésre	**ad**	*V.e3*	alkalmat a kormánnyal szemben,
5.	a színháznak, el kell	**adnia**	*V.INRe3*	áruját, az elõadást,
6.	vontatta Madeleine-Bastille omnibusz egy halottaskocsinak	**adta**	*V.TMe3*	át a helyét. Új
7.	személynek a másik szülõ nem	**adta**	*V.TMe3*	át. Vagyis mindkét szülõ
8.	tagállamok képviselõi, mintegy lökést	**adva**	*V.HIN*	az amerikai közvéleménynek is ahhoz
9.	egy-egy eseményt, és ez	**adja**	*V.Te3*	az igazi bonyodalmat. Lényeges
10.	egy iszákost), jelt	**ad**	*V.e3*	az indulásra. Néhány nap
11.	, aki szintén igenlõ választ	**adott**	*V.Me3*	az invitációra. Az ünneplés
12.	Attilától eredeztették, és ez	**adta**	*V.TMe3*	az uralom alapját. A
13.	, tehát egy általános fogalmat	**adjon**	*V.Pe3*	-e, vagy pedig csak
14.	Jasper Avenue-n a legkülönfélébb nációk	**adnak**	*V.t3*	egymásnak randevút, noha ez
15.	jöhet a következõ. Ne	**add**	*V.TPe2*	fel, hogy megéld az
16.	mûvészetre szakosodott múzeum többé nem	**ad**	*V.e3*	hivatalos attesztet, nõ a
17.	Ha az enzimet a sertéstáphoz	**adják**	*V.Tt3*	hozzá, csak 56 százalékkal
18.	A német hatóságok 72 szakembernek	**adták**	*V.TMt3*	ki a munkavállalási engedélyt,
19.	kell fellapoznia, melyet 1973-ban	**adott**	*V.Me3*	ki Marosi Ildikó. Ebben
20.	tágyakat és állóalapokat. f.	**adjon**	*V.Pe3*	ki saját szakfolyóiratot, jelentessen
21.	irányító posztjaiért, õk hiába	**adtak**	*V.Mt3*	ki utasításokat, a kisebb
22.	könyvében. Mûkincsrablások Európában címmel	**adta**	*V.TMe3*	közre a Száz eltûnt mûtárgy
23.	RJSZ-tõl játékengedélyt, ha nem	**adják**	*V.Tt3*	le a vb-n használt sportszereket
24.	mellett mûködõ állandó választott bírósághoz	**ad**	*V.e3*	majd be a vállalat.
25.	Visszahívom a jogászt. "	**Adja**	*V.TPe3*	meg a saját számát,
26.	csak az utolsó betûig végigolvasva	**adják**	*V.Tt3*	meg magukat. Ez a
27.	vannal azok, akik nem	**adnak**	*V.t3*	semmit a zenei kulturájukra.
28.	a " nyereségesség " elérésére	**ad**	*V.e3*	tanácsokat. És ki gondolná
29.	volt, hogy ha nem	**adták**	*V.TMt3*	volna el mindenüket, akkor
30.	összegeket. Amint korábban hírül	**adtuk**	*V.TMt1*	, a Mazsihisz szervezésében két
31.	egyes nemzeti bankokat a kezébe	**adja**	*V.Te3*	, amelyeket felhasznál, amelyek
32.	gyors és pontos diagnózist képesek	**adni**	*V.INF*	- csak éppen rendkívül drágák

3. ábra MNSz-lekérdezés: az *ad* közelében *–t* ragos szó



4. ábra Az *ad* ige bővítményeként előforduló leggyakoribb *–t* toldalékos névszók, amikor egy *–nak* toldalékú bővítmény is szerepel ugyanabban a szerkezetben



5. ábra Az *ad* ige bővítményeként előforduló leggyakoribb *–nak* toldalékos névszók, amikor egy *–t* toldalékú bővítmény is szerepel ugyanabban a szerkezetben

Míg az MNSz lekérdezési eredményeiben – különösen gyakori szavak esetén – igen nehéz áttekinteni, mely szavak fordulhatnak elő leggyakrabban például az *ad* ige tárgyragos

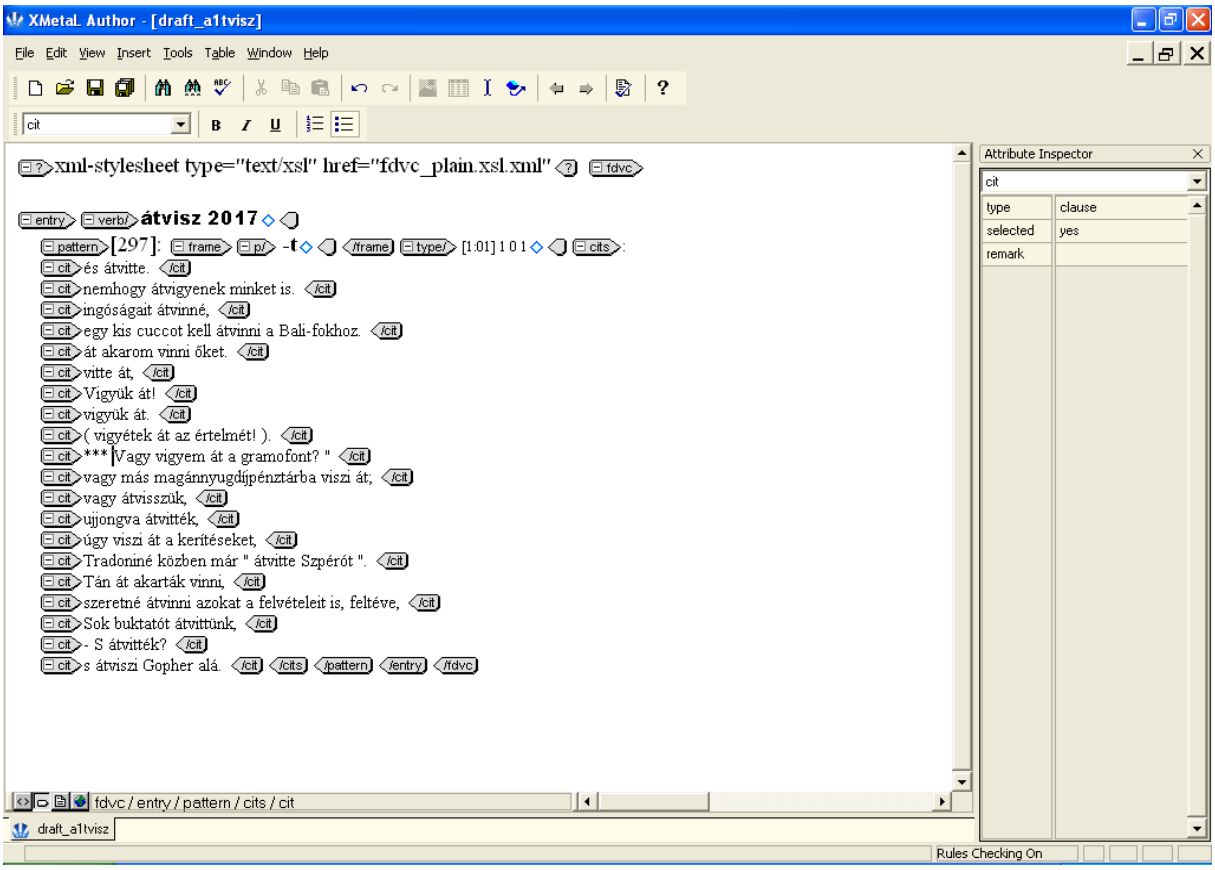
bővítményeként, a Mazsola program használatakor kimondottan szembeszökő az eredmény: úgy tűnik, várakozásainkkal ellentétben nem annyira konkrét tárgyak adásáról írunk igazán gyakran, hanem inkább sajátos jelentésű szókapcsolatokban használjuk: *hang*ot *ad*unk, *otthon*t, *hely*t, *lehetőség*et. (Valójában természetesen számos különféle konkrét tárgy *adás*áról is történik említés a korpuszban, csak ugyanarról a konkrét tárgyról nincs 250-nél több ismétlődő adat a korpuszban. Kivételt képez ez alól a *pénz*.) A Mazsola programban egy egyszerű kapcsolóval (Eloszlás) beállíthatjuk, melyik toldalékolt szó gyakorisága szerint rendezze az eredményt, így azt is áttekinthetjük, kinek/minek adunk valamit leggyakrabban. Az előző eredményhez hasonlóan itt is az átvitt értelmű szókapcsolatok aránya szembeszökő. A kétféle rendezés együttes eredménye: leggyakrabban *vélemény*ünknek és *aggodalm*unknak *ad*unk *hang*ot.

A Mazsola program elsősorban arra használható, hogy áttekintsük az igék legtipikusabb bővítményeit, az adott bővítményként előforduló szavakat és az igét tartalmazó több szóból álló kifejezéseket. Az MNSz-korpusz azon példamondatait is megtekinthetjük vele, amelyekből ezek az eredmények adódtak. Itt azonban rendszerint csak tagmondatok szerepelnek, azaz csak az egyszerű mondatokat láthatjuk teljes egészében, és semmilyen információnk nincs a mondatok lelőhelyéről. Ha bármilyen okból szükségünk van az adott mondat nagyobb szövegkörnyezetére és/vagy (hozzávetőleges) lelőhelyére, kikereshetjük ugyanazt a mondatot az MNSz lekérdezője segítségével. Ha itt a szokásosnál nagyobbra állítjuk a találat méretét, és bejelöljük, hogy mutassa meg a bibliográfiai adatokat is, alaposabban megvizsgálhatjuk a kiválasztott idézeteket. A Magyar igei szerkezetek szótárának készítésekor nem egyszer e két lekérdező program kombinált használatával döntöttük el egy-egy a program által javasolt szerkezet érvényességét, vagy így kerestünk megfelelő szótári példamondatot.

7. Nyers szócikkek automatikus generálása

A Magyar igei szerkezetek munkálatánál azt a megoldást találtuk leghatékonyabbnak, ha Sass Bálintnak az a programja, amely összesítette a legalább 250-szer előforduló igei szerkezeteket, és az igék szerint csoportosította az adatokat, az összesített eredményt közvetlenül olyan XML formátumban készíti el, amely már behívható bármely XML-szerkesztőbe. Az adott munkálathoz az XMetal szerkesztőt használtuk, mivel ezt már

korábban a Nagyszótár számára beszereztük, és jó tapasztalatunk volt használata során. A 6. ábrában egy rövid példaszócikkből látunk részletet.



6. ábra Munka közben a *Magyar igei szerkezetek*: az *átvisz* szócikke szerkesztése

A Sass Bálint által előkészített minden egyes nyers szócikket behívtuk az XMetal szövegszerkesztőbe. Szerkesztőként (Pajzs Júlia és Kiss Margit) eldöntöttük, hogy a javasolt címszó és a javasolt szerkezet valóban létezik-e. Ha igen, átolvastuk a program által felkínált, rendszerint 10, esetenként 20 példamondatot. Ha valamelyik megfelelőnek tűnt, egyetlen teendőnk volt, hogy ráálljunk a kurzorral a választott példamondatra, majd a képernyő jobb felső sarkában található argumentum szerkesztőben a példamondat „selected” argumentumát egy kattintással „Yes”-re állítsuk. Az esetek egy részében (kb. 20-25 százalékában) a felkínált példamondatok egyikét sem éreztük elég jónak, szótári példamondatként igazán alkalmasnak. Ilyenkor a Mazsola programmal és/vagy az MNSz. lekérdező programjával kerestünk megfelelő példamondatot, azt másoltuk a szócikkbe, és ezután állítottuk a bemásolt példamondat „selected” argumentumát „Yes” értékre. Ez a megoldás sokkal megbízhatóbb és gyorsabb volt, mintha akár minden szerkezethez magunknak kellett volna példát keresni és ide másolni, akár nekünk kellett volna ténylegesen kitörölni a feleslegesnek ítélt mondatokat.

A szerkesztés úgy folyt, hogy minden szócikket először az egyik szerkesztő nézte át, és választotta ki a megfelelő példamondatokat, esetlegesen megjegyzésekkel látta el valamelyik mezőt (pl. egy-egy szerkezet esetén: „mindig jelzővel!”), ezután átadta a másik szerkesztőnek, aki szintén látta az összes példamondatot, és módja volt arra, hogy másikat javasoljon kiválasztásra. Ilyenkor megjegyzésként arról is szót váltottunk, milyen szempontból találjuk megfelelőbbnek saját választásunkat. Az így végzett munka során igen hamar kialakult egy közös szempontrendszer a mondatok kiválasztására (ezeket részletesebben lásd: SASS–PAJZS 2010: 20–21) és a felmerült egyéb problémák egységes megoldására. Amelyik lexémát vagy szerkezetet nem találtunk jónak, azt is csupán a megfelelő elem „remark” argumentumában jelöltük törlésre, szükség esetén szöveges magyarázattal kiegészítve. Így ezeket a kérdéseket is mindketten mérlegeltük, majd a végső változat elkészítése előtt Sass Bálintnak módjában állt dönteni, elfogadja-e javaslatainkat. Úgy gondolom, hasonló megoldás talán más szótári munkálat hatékonyságának fokozására is alkalmas lehet.

A szócikkekből a többszöri átnézés után már újabb programok sorozatával állt elő a kész szótár: a mutatók mindegyikét programok generálták, a nyomtatott változat előkészítése is programok segítségével történt.

8. A többelemű szókapcsolatokról

A Magyar igei szerkezetek elkészítése és az ezt megelőző Mazsola program nagy lépés abba az irányba, hogy a rendelkezésünkre álló korpuszokból szakszerűen kigyűjtsük azokat a többelemű szókapcsolatokat, amelyek akár egy-, akár többnyelvű korszerű magyar szótárak készítésekor nélkülözhetetlenek. A modern, korpuszvezérelt lexikográfia egyik legfontosabb felismerése, hogy nem az egyes szavak jelentését érdemes boncolgatni, sokkal inkább az ismétlődő szósorozatok, többé-kevésbé állandósult szókapcsolatok elemzése alapján célszerű összeállítani a szócikkek egységeit és értelmezni az adott környezetbeli jelentést. A Mazsola program és a Magyar igei szerkezetek a mondatban központi szerepet betöltő igékre és bővítményeire koncentrál. Szükséges lenne hasonló lekérdezőprogram – és esetleg szótárszerű feldolgozás – a névszói csoportokra is: jó lenne, ha ugyanilyen könnyen megkérdezhetnénk, melyek egy adott főnév legtipikusabb jelzői, illetve fordítva: egyes melléknevek mely főnevek jelzőjeként fordulnak elő tipikusan. Akár készíthetnénk hasonló eljárással olyan korpuszvezérelt komplex kollokációs szótárt, amely mind a névszói, mind az

igei szerkezetekről együttesen ad információt. Olyan megoldást is lehetségesnek tartanék, ahol a nyomtatott változatban csak a szótári rész lényege szerepelne, minimális mennyiségű példával, a mutatók és a további bőséges példamondatok pedig egy CD-mellékletben lennének megtalálhatók.

Ezen túlmenően megfontolásra érdemesnek tartanám egy az angol DANTE-projekthez hasonló magyar korpuszvezérelt szótári adatbázis elkészítését is, amely mind korszerű magyar egynyelvű, mind kétnyelvű szótárak alapjául szolgálhatna.

9. Összegzés

Ahogy a tanulmányból reményeim szerint kitűnik, a ma készülő szótárak elkészítésének minden fázisában igen jól hasznosíthatók a nyelvtechnológia egyes eredményei. Elsődlegesen a szótár típusától függ, milyen mértékben hagyatkozhatunk az automatikus eljárásokra, milyen az aránya az aprólékos, lexikográfiai szakértelmet igénylő, csakis kézzel végezhető feladatoknak. A nyelvtechnológia eredményeinek hasznosítása nem csupán a szótár előállítását teheti hatékonyabbá, számos lehetőséget nyújt arra is, hogy a készülő szótárak minőségileg is eltérjenek a hagyományostól. Ilyen a korpusz alapján gyűjtött többelemű szókapcsolatok szerepének megváltozása, a valódi korpuszbeli példák sokaságának elhelyezése kisebb szótárakban is, az észlelt gyakoriság figyelembevétele az egyes szócikkek kidolgozásánál.

A szótárírás már maga is egyfajta szolgálat, hiszen nem az egyéni ötletek izgalmas kifejtésére ad lehetőséget, hanem rendszerint nagyobb munkacsoportban való fegyelmezett – sokszor rutinszerű, ezért nem ritkán unalmas – és kitartó munkavégzést követel. A nyelvtechnológia úgy szolgálhatja a „szolgákat”, ha a rabszolgamunka, a rutinfeladatok mind nagyobb részét tudja átvállalni, és egyszersmind meg tud felelni az adott munkálat által megkövetelt megbízhatóság kritériumának is.

Irodalom

GÁBOR Kata (2007) Syntactic Parsing and Named Entity Recognition for Hungarian with Intex. In: S. Koeva, D. Maurel, M. Silberztein (szerk.): Formaliser les langues avec

l'ordinateur: De Intex à Nooj, Presses Universitaires de Franche-Comté, Besançon, 353–366.

ITTZÉS Nóra (szerk.) (2006) A magyar nyelv nagyszótára I-II. MTA Nyelvtudományi Intézet, Budapest.

ITTZÉS Nóra (szerk.) (2011) A magyar nyelv nagyszótára III-IV. MTA Nyelvtudományi Intézet, Budapest.

KILGARIFF, 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, Toulouse: Association for Computational Linguistics. 32–38.

KIS Balázs − NASZÓDI Mátyás − PRÓSZÉKY Gábor (2003) Komplex (magyar) szintaktikai elemző rendszer mint beágyazott rendszer. In: Alexin Zoltán; Csendes Dóra (szerk.) Az 1. Magyar Számítógépes Nyelvészeti Konferencia előadásai (MSZNY), SZTE, Szeged, 145–150.

KISS Gabriella − KISS Margit − PAJZS Júlia (2004) A Nagyszótár történeti korpuszának elemzéséről. Magyar Nyelv C. évf. 2. sz., 185–191.

KISS Margit (2012) A digitális Mikes-szótár. Magyar Tudomány (megjelenés alatt)

NOVÁK Attila (2003) Milyen a jó Humor? In: Alexin Zoltán; Csendes Dóra (szerk.) Az 1. Magyar Számítógépes Nyelvészeti Konferencia előadásai. SZTE, Szeged, 138–145.

NOVÁK Attila M. − PINTÉR Tibor (2006) Milyen a még jobb Humor. In: Alexin Zoltán; Csendes Dóra (szerk.) A 4. Magyar Számítógépes Nyelvészeti Konferencia előadásai, Szeged: SZTE, 60–69.

ORAVECZ Csaba (2002) Large scale morphosyntactic annotation of the Hungarian National Corpus. Hollósi B.; Kiss-Gulyás J. (ed.). Studies in Linguistics, Volume VI. Debrecen: Institute of English and American Studies, University of Debrecen, 277–298.

PAJZS Júlia (1990) Számítógép és lexikográfia. MTA Nyelvtudományi Intézet, Budapest.

PAIS Judit − PAJZS Júlia (1998) Using local rules for disambiguation of Homographs in Hungarian corpora. Proceedings of the EURALEX ’98 Conference. University of Liege, Liege, 239–248.

PAJZS Júlia (1997a) Synthesis of results about analysis of corpora in Hungarian. Linguisticae Investigationes XXI-2. John Benjamins, Amsterdam, 349−365

PAJZS Júlia (1997b) Milyen szótár készíthető a nagyszótári korpuszból. in: Kiss G. − Zaicz G. (szerk.) Szavak – nevek – szótárak. Írások Kiss Lajos 75. születésnapjára. Budapest, MTA

NYTI, 289–297.

PAJZS Júlia (1999) Homonimák és homográfok egyértelműsítése a nagyszótári korpuszban. Gecső Tamás (szerk.) Poliszémia, homonimia Segédkönyvek a nyelvtudományok tanulmányozásához. Tinta Könyvkiadó, Budapest, 245–248.

PRÓSZÉKY Gábor − KISS Zoltán − TÓTH Lajos (1979) Magyar nyelvű szövegek számítógépes morfológiai vizsgálata. SOFTTECH D41. SZÁMKI, Budapest.

PRÓSZÉKY Gábor (1985) Automatizált morfológiai elemzés a nagyszótári munkálatokban. Kézirat, MTA Nyelvtudományi Intézet.

PRÓSZÉKY Gábor – TIHANYI László (1992) A Fast Morphological Analyzer for Lemmatizing Agglutinative Languages. In: Kiefer, Ferenc; Gábor Kiss; Júlia Pajzs (eds) Papers in Computational Lexicography (COMPLEX), Linguistics Institute of the HAS, Budapest, 265–278.

PRÓSZÉKY Gábor – PÁL Miklós – TIHANYI László (1994) Humor-based Applications. Proceedings of the 15th International Conference on Computational Linguistics (COLING), University of Kyoto, Kyoto, Japan, 1241–1244.

PRÓSZÉKY Gábor (2000) Számítógépes morfológia. In: Kiefer Ferenc (szerk.): Morfológia (Strukturális magyar nyelvtan III.), Akadémiai, Budapest, 1021–1064.

PRÓSZÉKY, Gábor – TIHANYI, László – UGRAY, Gábor (2004) Moose: a robust high-performance parser and generator. Proceedings of the 9th Workshop of the European Association for Machine Translation, Foundation for International Studies, La Valletta, Malta, 138–142.

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

SASS Bálint (2011) Igei szerkezetek gyakorisági szótára – egy automatikus lexikai kinyerő eljárás és alkalmazása. PhD-dolgozat, PPKE ITK, Budapest.

SASS Bálint – PAJZS Júlia (2010) Igei szerkezetek gyakorisági szótára – félautomatikus szótárkészítés nyelvtechnológiai eszközök segítségével Alkalmazott Nyelvtudomány X. évf. 1–2. szám MTA Nyelvtudományi Bizottság, Alkalmazott Nyelvészeti Munkabizottság, Veszprém, 5–32.

SASS Bálint – VÁRADI Tamás – PAJZS Júlia – KISS Margit (2010) Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára. Tinta Könyvkiadó, Budapest.

VÁRADI Tamás – PAJZS Júlia (1997) A magyar irodalmi és köznyelv nagyszótárának korpusza a HUNGARNET közösség számára. Proceedings of the Networkshop ’97 conference. Budapest, NIIF, CD.

Váradi Tamás (2002) ”The Hungarian National Corpus”. Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), Paris: ELRA, 385–389.