A készülő Akadémiai nagyszótár számítógépes vonatkozásai

Pajzs Júlia

MTA Nyelvtudományi Intézet
Lexikográfiai és Lexikológiai Osztály
1068 Budapest Benczúr u. 33.
pajzs@nytud.hu

Abstract. The project for the Academic Dictionary of Hungarian is presented from computational point of view. The major steps are the following: collection of the 25 million running word Historical Corpus of Hungarian, lemmatization, disambiguation, user friendly retrieval interface (www.nytud.hu/hhc), frequency database of the entries, on-line compilation of the dictionary entries with the XML module of the Corel Office 2000 WordPerfect 9 program. Presentation of the TEI based DTD of the dictionary.

Bevezetés

Az Akadémiai nagyszótár munkálatai 1985-ben indultak újra. Mintául a francia *Trésor de la langue française* projektum szolgált, amelyben számítógépes korpuszra és hagyományos módon gyűjtött cédulákra támaszkodva készítették el a francia nyelv 1789 utáni történeti szótárát.

1. A korpusz összeállítása és rögzítése

A magyar szótár forrásanyagául szolgáló történeti korpuszba rögzítendő anyagot irodalomtörténészek és különféle területek szakértői válogatták ki számunkra. Az anyag kijelölésével párhuzamosan megindult a próbaszövegek bevitele (kezdetben Commodore 64 típusú számítógépeken). Először saját kódrendszert alakítottunk ki a rögzítésre, az ékezetes és történeti karaktereket az angol ábécé betűiből és számok kombinációjából álló kódrendszerrel jelöltük (ún. Prószéky kód: á=a1, ö=o2 stb.). Amint megjelent a TEI SGML szabvány, áttértünk a korpusz TEI alapú kódolására, az ékezetes és történeti karaktereket azonban – a biztonságos hordozhatóság és egyértelmű konvertálhatóság érdekében – továbbra is Prószéky kódban tároljuk.

2. Morfológiai elemzés

A munkálat elindításakor elhatároztuk, hogy egy morfológiai elemző program segítségével könnyebben kezelhetővé, lekérdezhetővé tesszük az anyagot. Az elemző program terveit Prószéky Gábor készítette. A megvalósításhoz felhasználtuk Elekfi

László: *Szókincsünk nyelvtani alakrendszere* című művét. Az elemző program első változatát magam készítettem [18]. Később Tihanyi László csatlakozott a munkacsoporthoz, és ő folytatta a program írását, amelynek továbbfejlesztett változata HUMOR morfológiai elemző program néven vált ismertté, és a HELYES-E? alapjává.

Az elemző használata lehetővé tette, hogy ne pusztán szövegszavakat, szóalakokat keressünk a korpuszban, hanem lexémákat is. Így a lexikográfusoknak nem kell a szavak minden lehetséges toldalékolt alakját egyesével kikeresgélni, egyszerre lekérdezhetik az *alszik*, a *bokor* vagy a *hó* szó összes toldalékolt alakját.

Míg a mai szövegekre a HUMOR megfelelő hatékonysággal alkalmazható, a régiekre természetszerűleg lényegesen kevésbé. Ezért egy olyan heurisztikus algoritmusokkal dolgozó programot fejlesztettünk [12], [13], [14] amely az esetek jelentős részében lehetővé teszi a régies alakok helyes elemzését is.
Az eljárás lényege a következő:

A HUMOR program elemzi a szövegszavakat.

- Ha egy szót nem sikerült elemezni, a heurisztikus program átalakítja, majd újra elemezni próbálja azt.
- Ha az átalakított változat elemzése sikeres, a felismert, vagy felismerni vélt alakot őrizzük meg.

Az átalakításokat a történeti szövegekben megfigyelhető szabályszerűségek alapján végezzük el. A program első változata csaknem 30 százalékkal növelte a lemmatizálható szövegszavak mennyiségét.

3. Egyértelműsítés

Az 1990-es évek közepén fejlesztettem és teszteltem egy szabályalapú egyértelműsítő programot [18],[19]. Bár a történeti szövegek sajátosságai miatt az eredmények távolról sem voltak olyan jók, mint amilyenekről akár a nemzetközi, akár a hazai korpusznyelvészeti és nyelvtechnológiai kutatások számot adnak, a korpusz használhatóságát, lekérdezhetőségét jelentősen javította az eljárás alkalmazása.

4. Címszólista

Az elemzett és egyértelműsített korpusz alapján készítettem egy olyan címszógyakorisági listát, amely a gyakorisági adatokon túl a szó első és utolsó előfordulásának idejét is tartalmazta. Ebből nem csak a korpusz címszóállományának gazdaságáról kaphattunk képet, hanem a szavak időbeli eloszlásáról is (Kiss G. – Pajzs 2000, Pajzs 1997). A Morphologic Kft.-ben Tihanyi László készítette el a cimszójegyzék Mobidic felületű változatát.

5. Lekérdezés

A korpusz lekérdező programját, amely külső felhasználók részére is szabadon hozzáférhetővé teszi a korpuszt (www.nytud.hu/hhc) Váradi Tamás készítette [17]. A

lekérdezőnek korábban volt egy `telnet` alatt működő változata is, ez azonban ma már csak a belső felhasználók számára használható, biztonságtechnikai okokból. Lexikográfiai felhasználásra a lekérdezőnek kissé módoított változatát fejlesztettük ki, a `telnet` alattt működő változatot – Váradi Tamás programjának felhasználásával – fejlesztettem, ez az elemzett és egyértelműsített korpuszon működik, és a PAT (Open Text) lekérdező programot használja motorként. Egy másik, szintén a lexikográfusok speciális igényeit kiszolgáló programot Nagy Viktor készített számunkra, ennek lekérdező modulja a stuttgarti egyetemen fejlesztett Corpus Workbench program. Egyelőre ez is csak a belső felhasználók számára érhető el.

6. A szótári adatbázis készítése

A szócikkeket közvetlenül XML formában készítjük, a WordPerfect 9. XML szerkesztő moduljának felhasználásával. A szótár DTD-jét, és az ehhez tartozó XML applikációt – a TEI standard javaslatainak figyelembevételével – magam készítettem, az utóbbi időben Mártonfi Attila munkatársam tartja karban.

6.1 A korpusz Dokumentum típus definiciója

```
<!DOCTYPE dic
[
<!ELEMENT dic       (fileName, (entry | entryxr)+)>
<!ELEMENT entry  ((remark?, head, (sense | sengr)*,
xref*), compby?)>
<!ELEMENT head   (lemma, usg*, gramgrp?, (usgvar?,
variant)*, usg*, xref*, freq?)>
<!ELEMENT lemma  (#PCDATA | hom)*>
<!ELEMENT hom    (#PCDATA)>
<!ELEMENT variant (#PCDATA)>
<!ELEMENT gramgrp (subc*, pos*, gov?, lbl*)>
<!ELEMENT pos    (#PCDATA)>
<!ELEMENT subc   (#PCDATA)>
<!ELEMENT lbl    (#PCDATA | mention)*>
<!ELEMENT mention (#PCDATA |hint | hom)*>
<!ELEMENT usg    (#PCDATA)>
<!ELEMENT sense  (sennu?, (mainsens | sumsens)?,
subsen*)>
<!ELEMENT sennu  (#PCDATA)>
<!ELEMENT mainsens (usg*, gramgrp?, reflex?, usg*,
(def|defrep)*, dom*,hideg | eg)*, (re* | coll*))>
<!ELEMENT sumsens (usg*, defsum, coll*)>
<!ELEMENT sengr  (sgrnu, gramgrp, xref*, sense* )>
<!ELEMENT sgrnu  (#PCDATA)>
<!ELEMENT def(#PCDATA | gloss | hint | abbr | tr |
        mention| syn| dom )*>
<!ELEMENT defrep  (#PCDATA | gloss | hint| abbr| tr |
        mention | syn| dom )*>
```

```
<!ELEMENT defsum  (#PCDATA | gloss | hint | abbr | tr |
mention | syn | dom)*>
<!ELEMENT hint    (#PCDATA | usgphr?)*>
<!ELEMENT hinteg  (#PCDATA | hide )*>
<!ELEMENT coll    (usgphr*, (ph | hint)*, ((subd*) |
usgphr*, (def|defrep|defsum), (eg | hideg)*)))>
<!ELEMENT subd      (sdnu,usg*, (def|defrep), (eg |
hideg)*)>
<!ELEMENT eg        (cit,bibl)>
<!ELEMENT hideg     (cit,bibl)>
<!ELEMENT hide      (#PCDATA | hinteg | ref | ref2 | ph |
abbr)*>
<!ELEMENT bibl     (wdate, (author | pubTitle), id, p)>
<!ELEMENT xref    (xrtype?,xr*)>
<!ELEMENT xr      (#PCDATA | hom)*>
<!ELEMENT xrtype  (#PCDATA)>
<!ELEMENT wdate   (#PCDATA)>
<!ELEMENT cit     (#PCDATA | hinteg | ref | ref2 | ph |
abbr | hide )*>
<!ELEMENT ref     (#PCDATA | abbr )*>
<!ELEMENT ref2    (#PCDATA)>
]>
```

6.2 Az egyes tagek jelentése

<abbr>	a példamondaton belül rövidített vagy tildével helyettesített szó kiegészített, feloldott része
<author>	szerző, magyar fordító
<bibl>	bibliográfiai adatok egysége
<cDate>	a szócikk írásának időpontja
<cit>	idézet
<cName>	a szócikkíró neve
<coll>	értelmezett szókapcsolatok egysége
<compby>	a szócikkre vonatkozó információk blokkja
<deduced>	szóadatok nélküli, szókapcsolatból kiemelt címszó, paradigmatikus alakból elvont alakváltozat szögletes zárójelben
<def>	értelmezés
<defrep>	helyettesítő értelmezés csúcsos zárójelben
<defsum>	összefoglaló értelmezés
<dictions>	szótári hivatkozás (példamondat helyett, ill. szócikk végi blokkon belül)
<dom>	fogalomköri besorolás (az értelmezés része is lehet)
<eg>	a példamondat egysége
<entry>	önálló szócikk, ill. szócikkfejes utaló szócikk
<gloss>	az értelmezés csúcsos zárójeles kiegészítő része
<gov>	vonzat

<gramgrp>	grammatikai információk blokkja
<head>	szócikkfej
<hide>	rejtés (a teljes idézet vagy annak egy része rejtve)
<hideg>	a teljes rejtett példák (idézet + bibliográfia hivatkozás is) egysége
<hint>	az értelmezés kerek zárójeles vagylagos része, ill. az értelmezett szókapcsolat kimaradható része
<hinteg>	a példamondaton belüli szöveges kiegészítések és kihagyások szögletes zárójelben
<hom>	homonima indexszáma
<id>	a forrás kódszáma
<lbl>	nyelvtani kiegészítés a szófaj után kerek zárójelben
<lemma>	címszó
<mainsens>	adatolt főjelentés
<mention>	az <lbl>-en, ill. a <def>-en belüli kurzív rész (pl. `-t raggal hsz-szerűen; ill. 'abcűg kiáltással lehurrog vkit'`),
<nCorp>	korpuszbeli adatok száma
<nElse>	CD-adatok száma
<nSlip>	cédulák száma
<orauthor>	eredeti szerző (fordításoknál, a szócikkbe nem kerül be)
<p>	oldalszám
<ph>	az értelmezett szókapcsolat főváltozata adatként + a példamondat kiemelt részeként
<pos>	szófaj
<pubDate>	megjelenés éve (a szócikkbe nem kerül be)
<pubTitle>	a mű címe
<re>	bokrosított szócikk alcímszava
<ref>	a címszó előfordulása a példamondatban, kivéve értelmezett szókapcsolat részeként
<reflex>	igéből és *magát* tárgyból álló szerkezet (önálló jelentésben)
<remark>	a szócikkírás közbeni megjegyzések, figyelmeztetések, itt emeljük ki a filológiai ellenőrzendőket
<sdnu>	aljelentés **a)**, **b)** stb. betűjele
<sengr>	szófaji blokk
<sennu>	jelentésszám
<sense>	a jelentés blokkja
<sgrnu>	a szófajt jelölő római szám
<sources>	szócikk végi szótári hivatkozások blokkja
<status>	a szócikk állapota
<subc>	a szófajt megelőző szófaji kiegészítők (ts, tn)
<subd>	aljelentés egysége (ha ugyanaz a frazéma több jelentésben is használatos)
<subnu>	jelentésárnyalat száma
<subsen>	jelentésárnyalat
<sumsens>	adatok nélküli összefoglaló jelentés
<syn>	szinonima az értelmezésben
<tr>	az értelmezés kerek zárójeles kiegészítő részei (pl. latin növénynév)

<tra>	a ford. rövidítés egysége
<type>	a szótári hivatkozások **Vö.** rövidítése
<usg>	lexikai minősítés, ha a teljes szócikkre, ill. egy szófajra, jelentésre vagy frazéma aljelentésére vonatkozik
<usgphr>	az értelmezett szókapcsolatra, ill. a szólásra vonatkozó lexikai minősítés
<usgvar>	egy alakváltozatra vonatkozó lexikai minősítés
<variant>	alakváltozat a szócikkfejben
<vol>	kötetszám (a szócikkben nem jelenik meg)
<wdate>	keletkezés éve
<xr>	az a címszó, amelyre utalunk, ill. a szótári hivatkozásokban kiírandó címszó
<xref>*	összetételi és frazeológiai utalások blokkja (szócikk végén, belsejében, ill. csak ilyen szócikkfejes utalások szófaji minősítése után)
<xrtype>	az utalás típusának rövidítése

ágyszék fn

1. *(rég)* 'kanapé, pamlag': mind-a'-hárman az *Ágyszékre* le-ültenek (1790 Dugonics András C1468, 111) | {Ereszkedj-le édesem erre az *ágyszékre* (1793 Magyar játék-szín C2989, 22)} | Egy/zerre felugrott [Dorottya] az *ágy/zékröl* (1803 Csokonai Vitéz Mihály 1800069016, 48).

2. *(nyj)* 'lábazaton álló egyszerű deszkaágy': Az ágynak legősibb formája: az *ágyszék*, mely négyszegletű lábakból, deszkákból összerótt, kezdetleges tákolmány (1922 A magyar nép művészete CD07).

Vö. CzF.; ÚMTsz.

Szócikkíró: *Kéthely Anna*; első változat kelte: 2003. 01.

Szerkesztö: *Ittzés Nóra*; szerkesztés kelte: 2003. 03.

Cédulák száma: 30. Korpuszbeli adatok száma: 1.

A szócikk állapota: 3.

```
<entry><head><lemma>ágyszék</lemma>
<gramgrp><pos>fn</pos></gramgrp></head>
<sense><sennu/><mainsens>
<usg>rég</usg>
<def>kanapé, pamlag</def>
<eg><cit>mind-a'-hárman az <ref>ágy-székre</ref> le-u24ltenek </cit>
<bibl><wdate>1790</wdate><author>Dugonics András</author>
<id>C1468</id><p>111</p></bibl></eg>
<hideg><cit>Ereszkedj-le édesem erre az <ref>ágyszékre</ref></cit>
<bibl><wdate>1793</wdate><pubTitle> Magyar játék-szín </pubTitle>
<id>C2989</id><p>22</p></bibl></hideg>
```

Fig 1. Egy mintaszócikk kinyomtatott formában és XML változatának eleje

A <hideg> cimkével jelölt példamondatok csak a szótár elektronikus változatában fognak megjelenni. Más esetekben csak a példamondat egy részét rejtjük el a nyomtatáskor, ezeket a <hide> taggal jelöljük meg. Összességében a példamondatok bő egyharmada csak az adatbázis változatban lesz látható.

A készülő szótári adatbázis használata

Az XML változatban készülő szócikkekből nem csak a nyomtatott és az elektronikus változat állítható elő könnyedén, már munka közben is segíti, hogy több szempontból lekérdezhessük, ellenőrizhessük szócikkeinket. A szócikkek strukturális és mennyiségi ellenőrzését és lekérdezését pillanatnyilag Perl programok segítségével oldjuk meg, amelyek ACCESS adatbázisba exportálják a legfontosabb adatokat, ezekből azután a legkülönbözőbb lekérdezéseket generálhatjuk (pl. a példamondatok időbeli eloszlása, leggyakrabban idézett szerzők, melyik mondatokban szerepel Kádár János, vagy 1956 stb.) Ezen adatbázisok segítenek a munkatársak teljesítményének pontos, naprakész mérésében is.

Referenciák

[1] B. Lőrinczy É.-Gerstner K.: Lehet-e végre a magyar nyelvnek nagyszótára *Magyar Tudomány* 105 (1998): 261–71.

[2] Csengery K.-Ittzés N.(eds): Mutatványok az Akadémiai nagyszótárból MTA Nyelvtudományi Intézet, Budapest, 2002.

[3] Elekfi L.: Nagyszótári tervek és lehetőségek I–II *Magyar Nyelv* 93 (1997): 183–99, 296–311.

[4] Elekfi L.: Melléklet a Nagyszótári tervek és lehetőségek c. közleményhez: *út Magyar Nyelv* 94 (1998): 235–53.

[5] Elekfi L.: Nagyszótári tervek és lehetőségek III. *Magyar Nyelv* 94 (1998): 374–8.

[6] Elekfi L.: Mit tartalmaz a Szókincsünk nyelvtani alakrendszere c gyűjtemény? *Magyar Nyelv* 93 (1997): 63–8.

[7] Elekfi L.: A Magyar ragozási szótár és „Szókincsünk nyelvtani alakrendszere (1997): 213–21.

[8] Elekfi L.: Eltérő toldalékokban mutatkozó jelentéskülönbségek *Magyar Nyelvőr* 122 (1998): 305–17.

[9] Elekfi L.: Semantic differences of sufficial alternates in Hungarian. *Acta Linguistica Hungarica* 47 (2000): 145–77.

[10] Elekfi L.: Homonimák felismerése toldalékos alakok alapján. *Magyar Nyelvőr* 124 [2000]: 146–63.

[11] Gerstner K.: Cédulák és fájlok – A Magyar akadémiai nagyszótár alapjairól. In: Kiefer F.–Gósy M. (eds): *Helyzetkép a magyar nyelvtudományról* MTA Nyelvtudományi Inézet, Budapest, 2000, 35–43.

[12] Kiss Gabriella .– Pajzs J.: An attempt to develop a lemmatiser for the Historical Corpus of Hungarian. *Proceedings of CL 2001*. University of Lancaster, (2001),

[13] Kiss, Gabriella – Kiss, Margit – Pajzs, Júlia: Normalisation of Hungarian Archaic Texts Proceedings of COMPLEX 2001, University of Birmingham, (2001), pp. 83-94.

[14] Kiss L.-Pajzs J.: A magyar irodalmi és köznyelv nagyszótára (1533–1990) *Magyar Nyelv* 85 (1989): 129–36.

[15] Pajzs J.: Creating a Historical Dictionary of Hungarian with the Aid of Computer. In: *T. Magay–J. Zigány: BUDALEX '88 Proceedings.* Akadémiai Kiadó, Budapest, (1990), 559–63.

[16] Pajzs J.: Réalisation assistée par ordinateur de grands dictionnaires français et hongrois. *Cahiers d'études hongroises 3/91 Centre Interuniversitaire d'Études Hongroises Université Paris III.* Institut Hongrois de Paris, 47–54.

[17] Pajzs J.-Váradi T.: A magyar irodalmi és köznyelv nagyszótárának korpusza a HUNGARNET közösség számára. *A Networkshop '97 konferencia anyaga.* CD Budapest, NIIF, (1997).

[18] Pais J.–Pajzs J.: Using local rules for disambiguation of Homographs in Hungarian corpora. *Proceedings of the EURALEX '98 Conference.* University of Liège, 1998, 239–48.
[19] Pajzs J.: Synthesis of results about analysis of corpora in Hungarian. *Linguistiae Investigationes* XXI/2. John Benjamins, Amsterdam, (1997), 349–65.
[20] Pajzs, J.: Making Historical Dictionaries with the Computer. *Proceedings of EURALEX 2000,* University of Stuttgart, (2000), 249–59.