II. szám, 1986. június

KORNAI ANDRÁS

SZÓTÁRI ADATBÁZIS AZ AKADÉMIAI NAGYSZÁMÍTÓGÉPEN

Az MTA Nyelvtudományi Intézete
1250 Budapest, Szentháromság utca 2.

A cikk elsősorban az akadémiai IBM 3031-en működő szótári adatbázis jövendő felhasználóinak szól. Az első rész a rendszer keletkezésének történetét írja le - itt a forrásként felhasznált szótári anyagról kap tájékoztatást az olvasó. A második rész a rendszer jelenlegi állapotát és a legnagyobb (mintegy 80000 rekordból álló) file rekordjainak belső struktúráját ismerteti. Végül a harmadik rész rendszer továbbfejlesztésének várható menetéről szól.

0. Bevezetés

Ebben a munkában a Magyar Tudományos Akadémia IBM 3031-es számítógépén üzemelő szótári adatbázis (a továbbiakban: SZOTA1R) keletkezésének történetét (1. rész), jelenlegi állapotát (2. rész) és továbbfejlesztésének lehetőségeit (3. rész) írom le. Ez úton szeretnék köszönetet mondani mindazoknak, akik a SZOTA1R létrehozásához hozzájárultak: Bodó Éva (MTA SZTAKI), Détári György (MTA SZTAKI), Éltető László (Softinvest), Füredi Mihály (MTA Nyelvtudományi Intézet), Knuth Előd (MTA SZTAKI), Könyves Tóth Kálmán (Egyetemi Számítóközpont), Papp Ferenc (ELTE), Prószéky Gábor (OPKM), Tóth Péter (MTA SZTAKI), Vámos Tibor (MTA SZTAKI) - segítségük nélkül ez a rendszer nem jöhetet volna létre.

1. A múlt

1983 nyarán Könyves Tóth megemlítette Prószéky Gábornak, hogy a Papp Ferenc-féle a tergo szótár (a továbbiakban VégSz., ld. Papp 1969a) alapját képező ún. Debreceni Thesaurus lyukkártyái az MKKE egyik folyosóján tárva-nyitva álló szekrényben vannak: tartani lehet tőle, hogy az anyag elvész vagy megsemmisül. Prószéky vett egy lakatot és lezárta a szekrényt, majd szólt nekem. Ebben az időszakban a SZTAKI kiváló körülményeket biztosított a számítógépes nyelvészeti kutatásokhoz: az IBM 3031 a KGST legnagyobb számítógépe volt, és nemzetközi mércével mérve is jó középgépnek számított.

Bár addig csupán egy nem túl nagy toldaléktárat (Veenker 1968, magyarul ismerteti Papp 1969b) vittem fel a gépre, ebből is jól látszott, hogy a legnagyobb nehézséget az adatok hibátlan rögzítése jelenti. Az egyes címszavak illetve címszócsoportok kiválasztása ugyan jóval könnyebb volt, mintha az eredeti kiadványt kellett volna átlapozni, de ez a könnyebbség nem állt arányban a befektetett munkával. A számítógépes adatbázisokat a hagyományos "kézi" nyilvántartásoknál (pl. cédulakatalógus) mindenképpen előnyösebbé teszi nagyobb gyorsaságuk és hibamentes működésük: sajnos ezek az előnyök néhány ezer adat esetén még nem kárpótolják a felhasználót az adatbevitel nehézségeiért.

A régi lyukkártyák ebből a dilemmából kínáltak kiutat: úgy tűnt, hogy az Értelmező Szótár kishíján hatvanezer címszavát lényegében munka nélkül újra számítógépes környezetben lehet majd tanulmányozni. Éppen ezért május 17-én egy taxival az egész 35 doboznyi anyagot átszállítottuk a SZTAKI-ba és egy targoncán betoltuk a gépterembe. Néhány napot vártunk, hogy a kiszáradt kártyák a levegő nedvességét magukba szívják, majd megkezdtük a beolvasást. Az

első tíz-tizenöt doboz minden baj nélkül lement, de később egyre több kártyát kellett különvennünk: ezekről kézzel másolatot kellett csinálni, mert megrongálódtak. Egy ponton a beolvasott kártyákat tartalmazó file* túllépte a megengedett maximális méretet, és az egész olvasást újra kezdett kezdeni - az óriási file-méret még később is sok bajt okozott.

Néhány éjszakai megfeszített munka után végre az egész anyag együtt volt négy-öt file-ban - belenéztünk, és döbbenten láttuk, hogy munkánk eredménye betűk, számok és speciális karakterek áttekinthetetlenül zavaros halmaza. Hamarosan kiderítettük, hogy Papp Ferencék a Hollerith-szortgépes rendezés érdekében egy sajátos kódot vezettek be: szerencsére a szokásos karakterek visszaállítását egy erre a célra írt program segítségével viszonylag egyszerűen meg lehetett oldani, mert a karakterkódok megfejtését megtaláltuk az egyik lyukkártyán. A végeredményt a magyar ékezetes magánhangzókra kifejlesztett ún. Prószéky-kódban kaptuk meg: ebben az á megfelelője a1 (ugyanígy é helyett e1, és hasonlóan a hosszú í-re ó-ra és ú-ra); ö-nek o2 és ő-nek o3 felel meg (ü-re és ű-re hasonlóan). E kód fő előnye az, hogy belül marad azon a szűk karakterkészleten, ami minden számítógépen szabványos, de ennek ellenére gyakorlatilag korlátlanul bővíthető. (Ezt a rugalmasságot én elsősorban a szanszkrit szövegek latin transzliterációjában fellépő mellékjelek kódolásában használom ki, de a módszer elvben minden latin alapú ábécénél alkalmazható. A régebbi korok magyar grafémáinak kódolását ld. Prószéky 1985.)

Az immár jól olvasható anyagról kiderült, hogy a VégSz. anyagán kívül mást is tartalmaz: az etimológiai szótár anyagát a-tól gy-ig, számos orosz ige egyfajta kódolását, továbbá valamit, amiről azóta sem tudtuk kideríteni, hogy micsoda. Ezeket az anyagokat ma ETIM NAGY, OROSZ NAGY, ill. IBMDOBOZ NAGY nevű file-okban tároljuk - ez utóbbi név arra utal, hogy a file alapját képező kártyák egy IBM feliratú dobozból kerültek elő. A szótári file-okat egyesítettem egy olyan file-ba, ami már az IBM-en használatos ún. CMS file-formátumban volt: ez kapta a SZOTAR nevet.

A következő munkafázis az anyag rendezése volt: ezt nagyban megkönnyítette, hogy a kártyákon a kódok között sorszám is szerepelt. Ezt és a munkaszámot (ami minden kártyán ugyanaz volt) a későbbiekben eltávolítottam -

*A file (ejtsd fájl) szót a szabványban is rögzített magyar szakkifejezés ("adatállomány") mindeddig nem tudta kiszorítani a számítógépes szaknyelvből. Ennek oka valószínűleg nem csak az, hogy a magyar kifejezés négy szótaggal hoszabb, hanem az is, hogy a file szó értelmét (tkp. tetszőleges adatok egységbe foglalt halmazáról, dossziéről van szó) az adatállomány szó jelentése nem adja vissza, sőt még csak meg sem közelíti.

az egyes rekordok így 72 hosszúak lettek. Kb. másfélezer rekordot kellett kézzel kijavítani: az így nyert SZOTA1R DATA file lényegében megegyezett a VégSz. nyomtatott változatával (az utóbbi egyes nyilvánvaló sajtóhibái is javításra kerültek). A címszavak mellett feltüntetett ún. debreceni kódok részletes értelmezését a VégSz. előszavában megtalálja az olvasó: az eredeti anyag ezen túlmenően a szavak hosszára (nyomdai n-ben), stílusértékére és eredetére vonatkozó információt is tartalmaz. Mutatványképpen álljon itt a "szocializmus előtti szóhasználat" minősítésű szavak listája:

ADO1PE1NZTA1R	BE1LISTA1Z	DI1SZDOKTOR	GAZDATISZT
ADO1PRE1S	BE1RCSE1PLE1S	DI1SZMAGYAR	GRO1FNE1
ADO1TISZT	BE1RESGAZDA	DUGSEGE1LY	GRO1FNO3
ALJEGYZO3	BE1RESLEGE1NY	EGYKERENDSZER	GYA1MPE1NZ
ARATO1MUNKA1S	BI1RO1VISELT	EGYKE1Z	GYA1RIPAROS
ARATO1SZTRA1JK	BOLSEVISTA	ELCSEHESI1T	GYA1RTULAJDONOS
A1LLAMKO2LTSE1GES	BORDE1LY	ELEMISTA	GYERMEKMENHELY
A1LLAMSORSJA1TE1K	BORDE1LYHA1Z	ELMAGYAROSI1T	HABILITA1L
A1LLAMSORSJEGY	BORDE1LYOS	ELTOLONCOL	HADIMILLIOMOS
A1RU1VA1LTO1	BOTOSISPA1N	ENGEDE1LYES	HADSEREGSZA1LLI1TO1
A1RVAPE1NZ	CI1MZETES	EXCELLENCIA1S	HA1ZBIRTOK
A1RVAU2GY	CI1VIS	E1RDEKHA1ZASSA1G	HA1ZICSELE1D
BANKFIU1	CSELA1K	E1VJA1RADE1K	HA1ZISZOLGA
BANKHA1Z	CSELE1DHA1Z	FAJMAGYAR	HA1ZIU1R
BANKKO2LCSO2N	CSELE1DKO2NYV	FELA1R	HA1ZMESTERNE1
BANKTO3KE	CSELE1DLAKA1S	FEZO3R	HA1ZPARANCSNOK
BANKUZSORA	CSELE1DLA1NY	FE1LPROLETA1R	HELYSZERZO3
BANKU2ZLET	CSELE1DLE1PCSO3	FIATALU1R	HENTESINAS
BANKVEZE1R	CSELE1DNYU1ZA1S	FIZETE1SCSO2KKENTE1S	HENTESLEGE1NY
BANKZA1RLAT	CSELE1DSE1G	FIZETE1STELEN	HENTESSEGE1D
BA1LANYA	CSELE1DSOR	FOLYAMODVA1NY	HERCEGI
BA1RO1I	CSELE1DSZERZO3	FO2LDBIRTOKOS	HERCEGNO3
BA1RO1NO3	CSELE1DSZOBA	FO3HIVATALNOK	HERCEGSE1G
BA1RO1SA1G	CSENDO3RO3RS	FO3ME1LTO1SA1GU1	HIVATALSZOLGA
BECSÜLETBI1RÓSA1G	CSENDO3RSE1G	GABONABE1R	HU1SIPAROS
BENO3SU2L	CSENDO3RSORTU3Z	GARNISZA1LLO1	ILLETME1NYHIVATAL
BETEGBIZTOSI1TA1S	DIA1KVEZE1R	GAZDAIFJU1	IMPRESSZA1RIO1
BE1LISTA	DI1JBIRKO1ZO1	GAZDAKO2R	INASISKOLA
BE1LISTA1S	DI1JNOK	GAZDALEGE1NY	INGATLANIRODA

INGYENHELY
IPARISKOLA
IPARKAMARA
IPARMA1GNA1S
IPAROSINAS
IPAROSTANUL01
IPARRAJZISKOLA
IPARTESTU2LET
IRODAIGAZGAT01
I1NSE1GAD01
I1NSE1GMUNKA
I1NSE1GSEGE1LY
JAVI1T01NTE1ZET
JA1TE1KKLUB
JO1SZA1GIGAZGAT01
JO1TE1KONYKODIK
KABINETIRODA
KAMAT
KARDPA1RBAJ
KASZINO1TAG
KASZI1RNO3
KASZNA1R
KATONAISKOLA
KAUCIO1
KA1NTORTANI1T01
KEGYDI1J
KEGYDI1JAS
KERESKEDO3SEGE1D
KERTE1SZSEGE1D
KE1JNO3
KE1KHARISNYA
KE1NYSZERKO2LCSO2N
KE1NYSZERNYUGDI1JAZ
KICENZU1RA1Z
KIMENO3NAP
KISBE1RES
KISEMBER
KISTA1JGEROL
KISTO3KE1S
KOMORNA
KOMORNYIK
KONZORCIUM
KORMA1NYFO3TANA1CSOS
KORMA1NYLAP
KORMA1NYSAJTO1
KORMA1NYTANA1CSOS
KORONABIRTOK
KORTES
KORTESFOGA1S
KORTESHADJA1RAT
KORTESKEDE1S
KORTESKEDIK
KOSZTKAMAT
KO2NYVU2GYNO2K
KO2ZALAPI1TVA1NY
KO2ZE1PBIRTOK
KO2ZE1PBIRTOK
KO2ZE1PBIRTOKOS
KO2HIVATALNOK
KO2ZRENDO3R
KO2ZSE1GHA1ZA
KO2ZSE1GTANA1CS
KO2ZTISZTVISELO3
KULTUSZMINISZTER
KULTUSZMINISZTE1RIUM
KULTUSZTA1RCA
KU1RIAI
KU2LDVE1NY
KU2LTERU2LET
KVESZTOR
KVESZTU1RA
LEA1NYKERESKEDELEM
LEA1NYKERESKEDO3
LELENCHA1Z
LELENCU2GY
LEVENTEOKTATO1
LIBE1RIA1S
LO1KUPEC
LUDOVIKA1S
MAGA1NALKALMAZOTT
MAGA1NBANK
MAGA1NBANKA1R
MAGA1NDETEKTI1V
MAGA1NHIVATALNOK
MAGA1NISKOLA
MAGA1NNYOMOZO1
MAGA1NTISZTVISELO3
MAGA1NTITKA1R
MARHAKERESKEDO3
MEGYEHA1ZA
MENEDZSER
MENTO3EGYESU2LET
MINISZTERELNO2KSE1G
MOZIS
MUNKAADO1
MUNKANE1LKU2LI
MUNKANE1LKU2LISE1G
MUNKATA1BOR
MUNKA1SBIZTOSI1TA1S
MUNKA1SBIZTOSI1TO1
MUNKA1SEGYLET
MUNKA1SELLENES
MUNKA1SELO2ADA1S
MUNKA1SKE1RDE1S
MUNKA1SKIZA1RA1S
MUNKA1SKO2R
MUNKA1SLAP
MUNKA1SNYU1ZO1
MUNKA1SSZTRA1JK
NAGYBE1RLO3
NAGYBIRTOK

NAGYBIRTOKOS
NAGYIPAROS
NAGYKAPITALISTA
NAGYKERESKEDO3
NAGYME1LTO1SA1GU1
NAGYSA1GA
NAGYVILA1GI
NAPIDI1JAS
NE1PKONYHA
NE1PKO2R
NE1VHA1ZASSA1G
NO3EGYLET
NYOMORTANYA
OLA1H
OLA1HSA1G
ORSZA1GZA1SZLO1
O3FELSE1GE
O3FENSE1GE
O3ME1LTO1SA1GA
O3NAGYME1LTO1SA1GA
O3SKUTATA1S
O3PRO1BA
PANAMA
PANAMA1ZA1S
PANAMA1ZIK
PANAMISTA
PARASZTNYU1ZO1
PARKETT
PARVENU2
PA1RBAJ
PA1RBAJDU2H
PA1RBAJHO3S
PA1RBAJKE1PES
PA1RBAJKO1DEX
PA1RBAJOZIK
PA1RBAJSEGE1D
PA1RBAJVE1TSE1G
PA1RTHARC
PA1RTKASSZA
PA1RTVILLONGA1S
PE1NZU2GYIGAZGATO1
PE1NZU2GYIGAZGATO1SA1G
PLUTOKRATA
PLUTOKRA1CIA
POLGA1RISTA
POLGA1RO3RSE1G
PO1TADO1
PROSTITU1CIO1
PROTEZSA1L
RANGLISTA
RANGOSZTA1LY
RA1C
RA1DIO1TA1RSASA1G
REMUNERA1CIO1
RENDO3RBI1RO1
RENDO3RKAPITA1NYSA1G
RENDO3RTISZTVISELO3
RE1SZVE1NYES
SAJTO1FO3NO2K
SEGE1DLEVE1L
SEGE1DVIZSGA
SEGE1LYDI1JAS
SEGE1LYEGYESU2LET
SUSZTERINAS
SZAMA1RLE1TRA
SZEGE1NYADO1
SZEGE1NYHA1Z
SZEGE1NYNEGYED
SZEGE1NYSZAG
SZEGE1NYU2GY
SZEGO3DME1NY
SZEGO3DME1NYES
SZEGO3DTET
SZERETETADOMA1NY
SZERETETHA1Z
SZU2KSE1GMUNKA
TANA1CSJEGYZO3
TANA1CSNOK
TANA1RKE1PZO3
TANONCE1V
TANONCIDO3
TANONCISKOLA
TA1RSALKODO1NO3
TA1RSASA1GBELI
TA1VHA1ZASSA1G
TEKINTETES
TESTO3R
TESTO3RSE1G
TISZTISZOLGA
TISZTIU2GYE1SZ
TISZTU1JI1TO1
TOLONCHA1Z
TOLONCKOCSI
TOLONCOL
TO2MEGNYOMOR
TO2RVE1NYBI1RO1
TO2RVE1NYTELENI1T
TO3KEPE1NZ
TO3ZSDEJA1TE1K
TO3ZSDELOVAG
TO3ZSDETAG
TO3ZSDEU2GYNO2K
TO3ZSDE1ZIK
URADALMI
URASA1GI
UTCALA1NY
UZSORABE1R
U1JGAZDAG
U1RHO2LGY
U1RIAS
U1RIASSZONY

U1REMBER	VAGYONVA1LTSA1G	VOLONTO3R
U1RIHA1Z	VA1LTO1BE1LYEG	ZA1RDA NO2VENDE1K
U1RILA1NY	VA1RMEGYEHA1ZA	ZUGBANKA1R
U1RINO3	VERSENYTA1RGYALA1S	ZUGISKOLA
U1RISZOBA	VE1DLEVE1L	ZUGSAJTO1
U1RLOVAS	VE1DO3LEVE1L	ZUGSZA1LLI1TO1
U1RNO3	VE1DO3O3RIZET	ZSELLE1RHA1Z
U1RVEZETO3	VE1GELBA1NA1S	ZSELLE1RSOR
VAGYONADO1	VICEHA1ZMESTER	ZSI1RO1
VAGYONDE1ZSMA	VIRILISTA	ZSU1RFIU1

Az Eltető László által kifejlesztett adatbázis-kezelő rendszer (amelyet a SZOTA1R-on 1984 végén demonstráltunk) nem csupán a fentihez hasonló listák rendkívül gyors összeállítását teszi lehetővé, hanem azt is, hogy az anyagot bővítsük, javítsuk. Különösen hasznosnak bizonyult, hogy a rendszer (részletesen ismerteti Éltető 1985) lehetővé teszi a rekordstruktúra megváltoztatását, tehát új szempontok bevezetését is. Az első ilyen változtatás a szavak mássalhangzó-magánhangzó szerkezetét mutató ún. CV-csontváz (CV skeleton, ld. például Clements - Keyser 1983) bevezetése. Egy célprogram segítségével minden szóhoz (a példában 'ILLEMTANA1R') egy új, a CV-csontvázat tartalmazó mezőt rendeltünk (a példában 'VCCVCCVCVVC'). A program természetesen nem tudott minden digráfról, trigráfról ill. hangzókiesésről automatikusan dönteni, így a 'VI1Z-SUGA1R' stb. típusú szavak CV-csontvázát kézzel kellett kijavítani. (Az összes kétes esetet, tehát mintegy 15 ezer szót át kellett nézni, de szerencsére csak néhány százat kellett kijavítani.)

A második fontos változtatást az tette lehetővé, hogy Füredi Mihály végleges formába öntötte a Gyakorisági Szótár (a továbbiakban GyakSz.) számítógépes változatát. Mivel a két anyag ugyanazon a lemezen volt, nem volt nehéz "összefésülni" őket. Arról természetesen nem lehetett szó, hogy a GyakSz. összes adatát átvegyük, a SZOTA1R felhasználóinak azonban ilyen részletes tájékoztatásra nincs is szükségük. Éppen ezért az erre a célra kialakított F(rekvencia) mezőben csak egy egyszámjegyű kódot tüntettünk fel. Ez 0 akkor, ha a szó nem szerepel a GyakSz.-ban; 1 akkor, ha 1 gyakorisággal szerepel; 2 akkor, ha többször szerepel, de ugyanabban az anyagrészben; 3 akkor, ha kétszer szerepel, de különböző anyagrészekben; 4 akkor, ha a statisztikai eszközökkel kialakított ún. módosított gyakoriság (Fmod) 0 és 2 közé esik; 5 akkor, ha Fmod 2 és 4 közé esik; 6 akkor, ha Fmod 4 és 8 közé esik; 7 akkor,

ha Fmod 8 és 20 közé esik, végül 8 akkor, ha Fmod legalább 20.

A kódok jelentését a rendszer egy "FU" file-ban tárolja. Az F-kód esetén ez a file a következő:

```
0     NOT IN GYAKLEX
1     FABSZ = 1
2     FABSZ GE 2 AND FMOD = 0
3     FABSZ = 2 AND FMOD NE 0
4     FMOD LT 2 AND FMOD GT 0
5     FMOD GE 2 AND FMOD LT 4
6     FMOD GE 4 AND FMOD LT 8
7     FMOD GE 8 AND FMOD LT 20
8     FMOD GE 20
```

Azt, hogy ez a file milyen (hogy t.i. az első pozícióban szerepel a kód, és a hatodikon kezdődik a kód jelentésének a leírása), a rendszer egy újabb file-ban tárolja: ez az ún. rekordleírás. Esetünkben ez a következőképpen fest:

```
*** FILENEV:        FU01F
*** REKORDHOSSZ:     80
*** ENTRYK SZAMA:     4

     ENTRYNEV        TIPUS  SZAM   HELY  HOSSZ  OFFSET

1:   KOID              N      1      0     2
2:   JELENTE1S         M      1      5    55
3:   MEGJEGYZE1S       X      1     60    20
4:   KOID_JEL          O      1      0    60
```

A FU01F névből kiderül, hogy a 01-es file (t.i. a SZOTA1R) F nevű kódjáról van szó. Maga a kód csak számjegy lehet (numerikus, azaz N típus), de jelentése bármilyen alfanumerikus karaktert tartalmazhat (mixed, azaz M típus). A megjegyzés rovatban akár speciális karaktereket is használhatunk (extra, azaz X típus). Ha a kódra és jelentésére egyszerre akarunk rákérdezni, azaz a rekordokat az elsőtől a hatvanadik pozícióig tekintjük, akkor egy olyan mezőre van szükségünk, ami már másutt is említett pozíciókból épül fel (overlay, azaz O típus).

Az F-ben tárolt információ a gyakoriságról meglehetősen durva, de ezért cserébe igen megbízható tájékoztatást ad. Az adatok statisztikai természete miatt nagyobb pontossággal ("több tizedesre") csak a felső zónában lévő (F=8) szavak gyakoriságát lenne érdemes megadni, ezek az adatok azonban a GyakSz. kiadásával elérhetők lesznek. Az alsó zóna adatai sokban függnek a választott mintától, hiszen már egyetlen ívnyi adat hozzáadása számos szó abszolút gyakoriságát emelheti 0-ról 1-re vagy 1-ről 2-re - a nagyobb abszolút gyakoriságú szavak r e l a t í v gyakorisága természetesen nem fog lényegesen megváltozni.

Átvettünk a GyakSz.-ból néhány olyan kódot is (szófaj, homonímia-kód), ami az egyes homonímák azonosítását könnyíti meg: tekintve, hogy a homonímák beosztása a két anyagban nem ugyanolyan, ezek összefésülése csak kézi munkával, esetről esetre haladva lesz megvalósítható. Ezek a rekordok tehát valójában nem jelentenek új szócikkeket, a SZOTA1R kibővülése (jelenleg durván 80 ezer rekordból áll) tehát azoknak a szavaknak köszönhető, amelyek a VégSz.-ben nem szerepelnek, de a GyakSz.-ban igen.

A SZOTA1R file rekordleírása jelenleg a következő:

*** FILENEV: SZOTA1R
*** REKORDHOSSZ: 110
*** ENTRYK SZAMA: 51

	ENTRYNEV	TIPUS	SZAM	HELY	HOSSZ	OFFSET
1:	SZO1	M	1	0	31	
2:	O2SSZETETTSE1G	N	1	31	1	
3:	HOMONIMIA	N	1	32	1	
4:	FAJOK	M	1	33	3	
5:	JELENTE1SSZA1M	N	1	36	2	
6:	STI1LUS	N	1	38	2	
7:	TO3TI1OUS	N	1	40	2	
8:	TA1RGYRAG	N	1	42	2	
9:	TO2BBESSZA1M	N	1	44	2	
10:	BIRTOKOSRAG	N	1	46	2	
11:	EREDET	N	1	48	1	
12:	KE1PZO3	N	1	49	1	
13:	ZU3R	X	1	50	1	

14:	HOSSZ	N	1	51	2
15:	SORSZA1M	N	1	53	5
16:	HIA1NY	M	1	58	1
17:	CVZU3R	Z	1	59	1
18:	CSONTVA1Z	X	1	60	31
19:	SZO1-1-BETU3	0	1	0	1
20:	SZO1-2-BETU3	0	1	0	2
21:	SZO1_3_BETU3	0	1	0	3
22:	SZO1_4_BETU3	0	1	0	4
23:	SZO1_5_BETU3	0	1	0	5
24:	SZO1_10_BETU3	0	1	0	10
25:	SZO1_10_BETU3	0	1	0	12
26:	SZO1FAJ	2	3	33	1
27:	FO3SZO1FAJ	0	1	33	1
28:	ATERGO	0	1	0	31
29:	ATERGO_1	0	1	0	31
30:	ATERGO_2	0	1	0	31
31:	ATERGO_3	0	1	0	31
32:	ATERGO_4	0	1	0	31
33:	ATERGO_5	0	1	0	31
34:	NOMRAG	0	1	42	6
35:	C1TERGO	0	1	60	31
36:	C2TERGO	0	1	60	31
37:	C3TERGO	0	1	60	31
38:	C4TERGO	0	1	60	31
39:	C5TERGO	0	1	60	31
40:	CVTERGO	0	1	60	31
41:	H	N	1	91	1
42:	SZF	N	1	92	2
43:	F	N	1	94	1
44:	MARADE1K	X	1	95	15
45:	EGE1SZ	0	1	0	110
46:	C1BETU3	0	1	60	31
47:	C2BETU3	0	1	60	31
48:	C3BETU3	0	1	60	31
49:	C4BETU3	0	1	60	31
50:	C5BETU3	0	1	60	31
51:	C8BETU3	0	1	60	31

Az overlay mezőkre főként technikai okokból van szükség: ezek segítségével lehet a tergo, illetve rövidített (pl. első három betű) kereséseket végezni. Jelenleg kizárólag a szófajkód szerepel többször (egy szónak legfeljebb három szófaja lehet), de nem lenne nehéz más kódokat (pl. a tárgyragra vonatkozót) úgy átalakítani, hogy az ingadozásokat, az eredeti kódolásnak megfelelően, kettős kóddal jelöljük.

3. A jövő

A GyakSz.-szal való összefésülés azzal a következménnyel járt, hogy a CV-csontváz (és a hozzá tartozó atergo-mezők) kivételével egyik szempont szerint sem teljes a kódolás: azok mellől a szavak mellől, amelyek a GyakSz.-ból származnak, hiányzik a debreceni kód, és azok mellől, amelyek a GyakSz. félmillió szavas kiinduló anyagában nem szerepeltek, hiányzik (ill. 0) a gyakorisági kód. (Ez persze már önmagában elárul valamit az ilyen szavak gyakoriságáról.) Természetesen ezeket a hiányokat jó lenne megszüntetni, ez azonban meglehetősen összetett feladat. Tekintve, hogy a SZOTA1R kutatási célokra jelen állapotában is jól felhasználható, a teljességre törés önmagában nem indokolhatja a pótlólagos kódolással járó hatalmas munkát. Éppen ezért az alábbiakban a SZOTA1R jövőjét elsősorban a folyamatban levő nagyszótári munkával kapcsolatban vizsgálom.

A nagyszótári munkálatok a magyar lexikográfia, sőt, talán az egész magyar nyelvtudomány ezideig legnagyobb vállalkozását jelentik. Ezt nem csupán az erre a célra betervezett hatalmas pénzösszegek (melyek egy részéből a Nyelvtudományi Intézet kifejezetten erre a célra dedikált számítógép vásárlását tervezi), hanem az összeségében több mint 100 évet átfogó munkafolyamat is jól mutatja. E tanulmány keretei nem teszik lehetővé, hogy a vállalkozás eddigi sikereiről és jövőbeni terveiről részletesen írjak (nem is érzem magam hivatottnak erre) - megelégszem annak bemutatásával, kódról kódra, hogy a SZOTA1R egyes mezőinek kiegészítése és/vagy átalakítása a nagyszótár szempontjából mit jelent. Mivel a SZOTA1R eléggé rugalmasan alakítható, a nagyszótár pedig (már csak méreteinél fogva is) meglehetősen nagy tehetetlenséggel bír, ez utóbbi tervét adottnak veszem. Bár nem elképzelhetetlen, hogy ezek a tervek esetleg mégis módosulnak, úgy tűnik, hogy a nagyszótári munkálatokban következetesen érvényesíteni fogják a következő alapelveket:

(1) A munka empírikus alapját nem az eddigi szótárak, hanem összefüggő magyar nyelvű szövegek képezik.

(2) Automatizálandó minden olyan részfeladat, amit gazdaságosan automatizálni lehet.

Az (1) alatt betervezett több mint 10 millió szövegszó (amelynek egy része már rögzítésre is került) a számítógépesítést lényegében elkerülhetetlenné teszi. Véleményem szerint azonban hiba lenne (2)-t csupán szükséges rossznak tekinteni.

A magyar nyelv agglutináló természete miatt a szövegszavak szócikkekbe való csoportosításánál elkerülhetetlen egyes képzők vagy igekötők, de különösen a ragok és jelek leválasztása, tehát a morfológiai elemzés. Ennek automatizálását nem nehéz megoldani, a megoldás hatásfoka azonban nagyban függ az elemző által használt tőtártól. Ez részben mennyiségi kérdés (minél több tő szerepel a tőtárban, annál gyorsabb az elemzés), részben azonban minőségi: az elemzés annál jobb hatásfokú, minél több információt tartalmaz a tőtár az egyes tövek paradigmatikus alakjairól. A legfontosabb ilyen információ természetesen a szófaj.

A SZOTA1R szófajkódjai a Magyar Nyelv Értelmező Szótárának (a továbbiakban ÉrtSz.) szófajbesorolását tükrözik. A SZOTA1R tehát alkalmas arra, hogy a magyar lexikográfiának az ÉrtSz.-ben összefoglalt eredményeit a nagyszótár felé közvetítse, illetve annak munkálataiban felhasználhatóvá tegye. Sajnos a VégSz. (tehát eredetileg az ÉrtSz.) szófajkódjai nem teljesen felelnek meg a számítógép szabta precizitási és homogeneitási követelményeknek. Ez jól látszik a GyakSz. és a VégSz. szófajkódjainak összehasonlításából. Becslésem szerint a kódok legalább 2-3%-át módosítani kell majd. Célszerűnek tűnik nem az ÉrtSz., hanem az Értelmező Kéziszótár (a továbbiakban ÉKSz.) adataiból kiindulni - ez egyben a szókészlet kb. 15 ezer szavas bővülését is magával hozná.

A morfológiai elemzéshez azonban igen gyakran többre van szükség a durva szófajbesorolásnál. Nem elég tudni, hogy igével van dolgunk, azt is tudnunk kell, hogy például ikes-e. Ebben támpontot adhat a debreceni kód: a VégSz. (és így a SZOTA1R is) meglehetősen részletes információt tartalmaz a hangrendről, egyes toldalékokról, és paradigma-osztályba sorolást is ad. Ez az anyag azonban ismét inhomogén: ezt csak fokozta az az eljárás, hogy a tőszavak esetén a kódolók átvették az ÉrtSz. minősítéseit,de összetételek esetén saját nyelvérzékükre hagyatkoztak (ld. VégSz. 20-21. l.). A debreceni kód a magyar morfológia kutatóinak páratlanul érdekes adathalmazt kínál: érdemes lenne az újonnan bekerülő szavakat (tehát pl. a GyakSz. szavait) is

minősíttetni az eredeti kódolókkal, sőt kutatási szempontból az anyag akkor lenne igazán homogén, ha a tőszavakat is az ő nyelvérzékük szerint kódolnák le. Ez a fajta empírikus adatgyűjtés (amely véleményem szerint a VégSz. egyik legpozitívabb vívmánya) azonban kevéssé illeszthető be a nagyszótár írott nyelvre hagyatkozó munkálataiba, s csak igen kevéssé hasznos a morfológiai elemzés automatizálásában. Ehhez a feladathoz mindenképpen a debreceni kódnál jóval lényegesebb paradigma-kódokat kellene használni, például azokat, amelyeket Elekfi László Szókincsünk nyelvtani alakrendszere című munkájában az ÉKSz. egész anyagán végigvitt.

Bizonyos kódok, mint például a szóhossz vagy a CV-csontváz, automatikusan vagy majdnem automatikusan generálhatók. Más kódok, így például a szavak eredetét tükröző etimológiai kód előállítása azonban igen komoly emberi munkát igényel. E tevékenység automatizálására a közeli jövőben nem is gondolhatunk. A SZOTA1R azonban az ilyen kódok megállapításához is adhat segítséget azza, hogy közvetíti a magyar lexikográfia eddig elért eredményeit. Jelen formájában ugyan csak a Bárczi-féle Szófejtő Szótár adatait tartalmazza (ld. VégSz. 24. l.), de nem lenne nehéz kiegészíteni a TESz. adataival sem, hiszen ezek egy részét a debreceni munkacsoport már lekódolta. Ugyanez mondható a stílusminősítésekre is: célszerűbbnek tűnik az ÉrtSz. kódjait módosítani, mint az egész munkát újrakezdeni.

A nagyszótárnak természetesen nemcsak a szótan vagy az etimológia, hanem az egész magyar nyelvészet érdekeit kell szolgálnia: éppen ezért fontosnak tűnik, hogy szintaktikai jellegű információt is tartalmazzon. A legfontosabb talán az igei vonzatkeretek (elsősorban a kötelező bővítmények) megadása lenne. A SZOTA1R kódjai ehhez is kiindulási alapot adnak. Fontos lenne azonban a kódokat úgy kiterjeszteni, hogy ne csupán az első szótári jelentéshez tartozó vonzatkeret legyen megadva, és hogy mindenütt legyen kód (pl. névutóknál, mellékneveknél), ahol egyáltalán vonzatról beszélhetünk.

Végül ide tartozik a szemantikai kódok, tehát a jelentés kérdése is: az a véleményem, hogy a SZOTA1R itt is hasznosnak bizonyulhat. Ez az állítás meglepőnek tűnhet annak a fényében, hogy a VégSz. semmiféle szemantikai kódot nem tartalmaz, és hogy a ÉKSz. (de különösen az ÉrtSz.) értelmezései távolról sem tökéletesek. Hangsúlyozom azonban, hogy a SZOTA1R nem csupán adatok statikus halmaza, hanem egyben az adatokkal dolgozó (dinamikus) rendszer is. Miért fontos ez? A hagyományos szótárak értelmezéseit elsősorban azért nem lehet "intelligens" számítógépes rendszerekben felhasználni, mert körkörösek, a szemetet a hulladék, a hulladékot pedig a szemét segítségével definiálják.

Ezt a hibát csak akkor lehet elkerülni, ha a szótár készítői előre rögzítenek egy alapszókincset, és minden egyéb szót ennek segítségével értelmeznek. (Ez az eljárás persze nemcsak a számítógép, ahnem a nyelvet tanuló más anyanyelvű diák feladatát is hatalmas mértékben megkönnyíti.) Ilyen elven készült például a Longman Dictionary of Contemporary English (számítógépes felhasználását ld. Alshawi - Boguraev - Briscoe 1985); ennek alapszókincséből alakítottam ki egy ALAP DATA nevű file-t. A számítógépes környezet lehetővé teszi, hogy az értelmezések konzisztens voltát állandóan ellenőrizzük, és hogy a szójelentés kérdését világosan különválasszuk az "enciklopédikus" tudástól. Egy nyelvről való ismereteink jó részét szükségképpen a nyelv szótárának kell tárolnia: a XXI században, mire "A magyar nyelv nagyszótára" elkészül, már fontos lesz, hogy ezek az ismeretek ne csak az emberek, hanem a számítógépek számára is hozzáférhetők legyenek.

Irodalom

Alshawi, H. - Boguraev, E. - Briscoe, T. (1985): A dictionary support environment for real time parsing. In: Proceedings of the 2nd Conference of the European Chapter of the Association for Computational Lingusitics. 171-178.

Clements, N. - Keyser, S. (1983): CV Phonology. MIT Press, Cambridge, Mass.

Éltető L. (1985): Új adatbáziskezelő rendszer VM/CMS alatt. Információ - Elektronika. Megjelenés alatt.

Papp F. (1969a): A magyar nyelv szóvégmutató szótára. Budapest, Akadémiai.

Papp F. (1969b): Veenker ismertetése. Nyelvtudományi Közlemények 71: 190-194.

Prószéky G. (1985): Automatizált morfológaii elemzés a nagyszótári munkálatokban. Kézirat, MTA Nyelvtudományi Intézet.

Veenker, W. (1968): Verzeichnis der ungarischen Suffixe und Suffixkombinationen. Mitteilungen der Societas Uralo-Altaica 3. Hamburg.