Szótári adatbázis az akadémiai nagyszámítógépen

Kornai András
MTA Nyelvtudományi Intézet

1986 Június

0 Bevezetés

Ebben a munkában a Magyar Tudományos Akadémia IBM 3031-es számitógépén üzemelő szo1tári adatbázis (a továbbiakban: SZOTA1R) keletkezésének történetét (1. rész), jelenlegi állapotát (2. rész) és továbbfejlesztésének lehetőse1geit (3. rész) irom le. Ez úton szeretnék köszo2netet 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 Számitóközpont) Papp Ferenc (ELTE), Prószéky Gábor (OPKM), Tóth Péter (MTA SZTAKI) Vámos Tibor (MTA SZTAKI) – segitségük nélkül ez a rendszer nem jöhetett volna létre.

1 A mult

1983 nyarán Könyves Tóth Kálmán megemlitette Prószéky Gábornak, hogy a Papp Ferenc féle *a tergo* szótár (a továbbiakban VégSz., ld. Papp 1969a) alapját ke1pező un. 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 kivaló körülményeket biztositott a számitógépes nyelvészeti kutatásokhoz: az IBM 3031 a KGST legnagyobb számitógépe volt, és nemzetközi mércével is jó középgépnek számitott.

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 rögzitése jelenti. Az egyes cimszavak illetve cimszócsoportok kiválasztása ugyan jóval könnyebb volt, mintha az eredeti kiadványt kellett volna átlapozni, de ez a könnyebség nem állt arányban a befektetett munkával. A számitógépes 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 kináltak kiutat: úgy tűnt, hogy az Értelmező Szótár kishiján hatvanezer cimszavát lényege1ben munka nélkül újra számitógépes környezetben lehet majd tanulmányozni. Éppen ezért május 17-én egy taxival az egész 35 doboznyi anyagot átszállitottuk a SZTAKI-ba, és egy targoncán betoltuk a gépterembe. Néhány napot vártunk, hogy a kiszáradt kártyák a levegő nedvességét magukba szivják, majd megkezdtük a beolvasást. Az első tiz-tizenöt doboz minden baj nélkül lement, de később egyre to3bb kártyát kellett különvennünk: ezekro3l kézzel másolatot kellett csinálni, mert megrongálódtak. Egy ponton a beolvasott kártyákat tartalmazó file[1] túllépte a megengedett maximális méretet, és az egész olvasást ujra kellett kezdeni – az óriási file-méret még később is sok bajt okozott. Néhány éjszakai megfeszitett 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 kideritettük, hogy Papp Ferencék a Hollerith-szortgépes

[1] A file (ejtsd fájl) szót a szabványban is rögzitett magyar szakkifejezés ("adatállomány") mindeddig nem tudta kiszoritani a számitógépes szaknyelvből. Ennek oka valószinüleg nem csak az, hogy a magyar kifejezés négy szótaggal hosszabb, hanem az is, hogy a file szó értelmét (tkp. tetszóleges adatok egységbe fogott halmazáról, dossziéről van szó) az adatállomány szo1 jelentése nem adja vissza, sőt még csak meg sem közeliti.

rendezés érdekében egy sajátos kódot vezettek be: szerencsére a szokásos karakterek visszaállitását egy erre a célra irt program segitségével viszonylag egyszerűen meg lehetett oldani, mert a karakterkódok megfejtését megtaláltuk az egyik lyukkártyán. A végeredményt a magyar ékezetes magánhangzókra kifejlesztett u.n. Prószéky-kódban kaptuk meg: ebben az á megfelelője a1, (ugyanigy é helyett e1, és hasonlóan a hosszú i-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 számitógépen szabványos, de ennek ellenére gyakorlatilag korlátlanul bővithető. (Ezt a rugalmasságot én elsősorban a szanszkrit szövegek latin transzliterációjában fellépő mellékjelek kódolásában haszna1lom ki, de a módszer elvben minden latin alapu ábécénél alkalmazhato1. A régebbi korok magyar grafe1máinak kódolását ld. Prósze1ky 1985.)

Az immár jól olvasható anyagról kiderült, hogy a VégSz. anyagán kivűl mást is tartalmaz: az etimológiai szótár anyagát a-tól gy-ig, számos orosz ige egyfajta kódolását, továbbá valamit, amiről azóta sem tudtuk kideriteni, 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 feliratu dobozból kerültek elő. A szótári file-okat egyesitettem egy olyan file-ba, ami már az IBM-en használatos u.n. CMS file-formátumban volt: ez kapta a SZOTAR nevet.

A következő munkafázis az anyag rendezése volt: ezt nagyban megkönnyitette, hogy a kártyákon a kódok között sorszám is szerepelt. Ezt és a munkaszámot (ami minden kártyán ugyanaz volt) a későbbiekben eltávolitottam – az egyes rekordok igy 72 hosszuak lettek. Kb. másfélezer rekordot kellett kézzel kijavitani: az igy nyert SZOTA1R DATA file lényegében megegyezett a VégSz. nyomtatott változatával (az utóbbi egyes nyilvánvaló sajtóhibái is javitásra kerültek). A cimszavak mellett feltu2ntetett u.n. debreceni kódok részletes értelmezését a VégSz. előszavában megtalálja az olvasó: az eredeti anyag ezen túlmenően a szavak hosszára (nyomdai n-ben), stilusértékére, és eredetére vonatkozó információt is tartalmaz. Mutatványképpen álljon itt a "szocializmus előtti szóhasználat" minősitésű szavak listája:

ADO1PE1NZTA1R ADO1PRE1S ADO1TISZT ALJEGYZO3 ARATO1MUNKA1S ARATO1SZTRA1JK
A1LLAMKO2LTSE1GES A1LLAMSORSJA1TE1K A1LLAMSORSJEGY A1RUVA1LTO1 A1RVAPE1NZ
A1RVAU2GY BANKFIU1 BANKHA1Z BANKKO2LCSO2N BANKTO3KE BANKUZSORA BANKU2ZLET
BANKVEZE1R BANKZA1RLAT BA1LANYA BA1RO1I BA1RO1NO3 BA1RO1SA1G
BECSU2LETBI1RO1SA1G BENO3SU2L BETEGBIZTOSI1TA1S BE1LISTA BE1LISTA1S BE1LISTA1Z
BE1RCSE1PLE1S BE1RESGAZDA BE1RESLEGE1NY BI1RO1VISELT BOLSEVISTA BORDE1LY
BORDE1LYHA1Z BORDE1LYOS BOTOSISPA1N CI1MZETES CI1VIS CSELA1K CSELE1DHA1Z
CSELE1DKO2NYV CSELE1DLAKA1S CSELE1DLA1NY CSELE1DLE1PCSO3 CSELE1DNYU1ZA1S
CSELE1DSE1G CSELE1DSOR CSELE1DSZERZO3 CSELE1DSZOBA CSENDO3RO3RS CSENDO3RSE1G
CSENDO3RSORTU3Z DIA1KVEZE1R DI1JBIRKO1ZO1 DI1JNOK DI1SZDOKTOR DI1SZMAGYAR
DUGSEGE1LY EGYKERENDSZER EGYKE1Z ELCSEHESI1T ELEMISTA ELMAGYAROSI1T ELTOLONCOL
ENGEDE1LYES EXCELLENCIA1S E1RDEKHA1ZASSA1G E1VJA1RADE1K FAJMAGYAR FELA1R FEZO3R
FE1LPROLETA1R FIATALU1R FIZETE1SCSO2KKENTE1S FIZETE1STELEN FOLYAMODVA1NY
FO2LDBIRTOKOS FO3HIVATALNOK FO3ME1LTO1SA1GU1 GABONABE1R GARNISZA1LLO1
GAZDAIFJU1 GAZDAKO2R GAZDALEGE1NY GAZDATISZT GRO1FNE1 GRO1FNO3 GYA1MPE1NZ
GYA1RIPAROS GYA1RTULAJDONOS GYERMEKMENHELY HABILITA1L HADIMILLIOMOS
HADSEREGSZA1LLI1TO1 HA1ZBIRTOK HA1ZICSELE1D HA1ZISZOLGA HA1ZIU1R HA1ZMESTERNE1
HA1ZPARANCSNOK HELYSZERZO3 HENTESINAS HENTESLEGE1NY HENTESSEGE1D HERCEGI
HERCEGNO3 HERCEGSE1G HIVATALSZOLGA HU1SIPAROS ILLETME1NYHIVATAL IMPRESSZA1RIO1
INASISKOLA INGATLANIRODA INGYENHELY IPARISKOLA IPARKAMARA IPARMA1GNA1S
IPAROSINAS IPAROSTANULO1 IPARRAJZISKOLA IPARTESTU2LET IRODAIGAZGATO1 I1NSE1GADO1
I1NSE1GMUNKA I1NSE1GSEGE1LY JAVI1TO1INTE1ZET JA1TE1KKLUB JO1SZA1GIGAZGATO1
JO1TE1KONYKODIK KABINETIRODA KAMAT KARDPA1RBAJ KASZINO1TAG KASZI1RNO3 KASZNA1R
KATONAISKOLA KAUCIO1 KA1NTORTANI1TO1 KEGYDI1J KEGYDI1JAS KERESKEDO3SEGE1D
KERTE1SZSEGE1D KE1JNO3 KE1KHARISNYA KE1NYSZER- KO2LCSO2N KE1NYSZER- NYUGDI1JAZ
KICENZU1RA1Z KIMENO3NAP KISBE1RES KISEMBER KISTA1JGEROL KISTO3KE1S KOMORNA
KOMORNYIK KONZORCIUM KORMA1NY- FO3TANA1CSOS KORMA1NYLAP KORMA1NYSAJTO1
KORMA1NYTANA1CSOS KORONABIRTOK KORTES KORTESFOGA1S KORTESHADJA1RAT
KORTESKEDE1S KORTESKEDIK KOSZTKAMAT KO2NYVU2GYNO2K KO2ZALAPI1TVA1NY

KO2ZE1PBIRTOK KO2ZE1PBIRTOKOS KO2ZHIVATALNOK KO2ZRENDO3R KO2ZSE1GHA1ZA KO2ZSE1GTANA1CS KO2ZTISZTVISELO3 KULTUSZMINISZTER KULTUSZ- MINISZTE1RIUM KULTUSZTA1RCA KU1RIAI KU2LDVE1NY KU2LTERU2LET KVESZTOR KVESZTU1RA LEA1NYKERESKEDELEM LEA1NYKERESKEDO3 LELENCHA1Z LELENCU2GY LEVENTEOKTATO1 LIBE1RIA1S LO1KUPEC LUDOVIKA1S MAGA1NALKALMAZOTT MAGA1NBANK MAGA1NDETEKTI1V MAGA1NHIVATALNOK MAGA1NISKOLA MAGA1NNYOMOZO1 MAGA1NTISZTVISELO3 MAGA1NTITKA1R MARHAKERESKEDO3 MEGYEHA1ZA MENEDZSER MENTO3EGYESU2LET MINISZTERELNO2KSE1G MOZIS MUNKAADO1 MUNKANE1LKU2LI MUNKANE1LKU2LISE1G MUNKATA1BOR MUNKA1SBIZTOSI1TA1S MUNKA1SBIZTOSI1TO1 MUNKA1SEGYLET MUNKA1SELLENES MUNKA1SELO3ADA1S MUNKA1SKE1RDE1S MUNKA1SKIZA1RA1S MUNKA1SKO2R MUNKA1SLAP MUNKA1SNYU1ZO1 MUNKA1SSZTRA1JK NAGYBE1RLO3 NAGYBIRTOK NAGYBIRTOKOS NAGYIPAROS NAGYKAPITALISTA NAGYKERESKEDO3 NAGYME1LTO1SA1GU1 NAGYSA1GA NAGYVILA1GI NAPIDI1JAS NE1PKONYHA NE1PKO2R NE1VHA1ZASSA1G NO3EGYLET NYOMORNEGYED NYOMORTANYA OLA1H OLA1HSA1G ORSZA1GZA1SZLO1 O3FELSE1GE O3FENSE1GE O3ME1LTO1SA1GA O3NAGYME1LTO1SA1GA O3SKUTATA1S O3SPRO1BA PANAMA PANAMA1ZA1S PANAMA1ZIK PANAMISTA PARASZTNYU1ZO1 PARKETT PARVENU2 PA1RBAJ PA1RBAJDU2H PA1RBAJHO3S PA1RBAJKE1PES PA1RBAJKO1DEX PA1RBAJOZIK PA1RBAJSEGE1D PA1RBAJVE1TSE1G PA1RTHARC PA1RTKASSZA PA1RTVILLONGA1S PE1NZU2GYIGAZGATO1 PE1NZU2GY- IGAZGATO1SA1G PLUTOKRATA PLUTOKRA1CIA POLGA1RISTA POLGA1RO3RSE1G PO1TADO1 PROSTITU1CIO1 PROTEZSA1L RANGLISTA RANGOSZTA1LY RA1C RA1DIO1TA1RSASA1G REMUNERA1CIO1 RENDO3RBI1RO1 RENDO3R- KAPITA1NYSA1G RENDO3R- TISZTVISELO3 RE1SZVE1NYES SAJTO1FO3NO2K SEGE1DLEVE1L SEGE1DVIZSGA SEGE1LYDI1JAS SEGE1LYEGYESU2LET SUSZTERINAS SZAMA1RLE1TRA SZA1SZEGE1NYADO1 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 U1RIEMBER U1RIHA1Z U1RILA1NY U1RINO3 U1RISZOBA U1RLOVAS U1RNO3 U1RVEZETO3 VAGYONADO1 VAGYONDE1ZSMA VAGYONVA1LTSA1G VA1LTO1BE1LYEG VA1RMEGYEHA1ZA VERSENYTA1RGYALA1S VE1DLEVE1L VE1DO3LEVE1L VE1DO3O3RIZET VE1GELBA1NA1S VICEHA1ZMESTER VIRILISTA VOLONTO3R ZA1RDANO2VENDE1K ZUGBANKA1R ZUGISKOLA ZUGSAJTO1 ZUGSZA1LLI1TO1 ZSELLE1RHA1Z ZSELLE1RSOR ZSI1RO1 ZSU1RFIU1

2 A jelen

Az Éltető László által kifejlesztett adatbázis-kezelő rendszer (amelyet a SZOTA1R-on 1984 végén demonstráltunk) nem csupán a fentihez hasonló listák rendkivűl gyors összeálitását teszi lehetővé, hanem azt is, hogy az anyagot bővitsük, javitsuk. 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. pl. Clements – Keyser 1983) bevezetése. Egy célprogram segitségével minden szóhoz (például 'ILLEMTANA1R') egy új, a CV-csontvázat tartalmazó mezőt rendeltünk (a példában 'VCCVCCVCVVC'). A program természetesen nem tudott minden digráfról, trigráfról ill. hangzókiesésről automatikusan dönteni, igy a 'VI1ZZSUGA1R' típusú szavak CV-csontvázát kézzel kellett kijavitani. (Az összes kétes esetet, tehát mintegy 15 ezer szót át kellett nézni, de szerencsére csak néhány százat kellett kijavitani.)

A második fontos változtatást az tette lehetővé, hogy Füredi Mihály végleges formába öntötte a Gyakorisági Szótár (a továbbiakban GyakSz.) számitógépes 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 kialakitott F(rekvencia) mezőben csak egy egyszámjegyű kódot tüntettünk fel. Ez 0 akkor, ha a szó nem szerepel a GyakSz.-ban; 1 akkor, ha 1 gyakoriséggal szerepel; 2 akkor, ha többször szerepel, de ugyanabban az anyagrészben; 3 akkor, ha kétszer szerepel, de különböző anyagrészekben; 4 akkor, ha a statisztikai eszközökkel kialakitott un. módositott 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 pozicióban szerepel a kód, és a hatodikon kezdődik a kód jelentésének a leirása), a rendszer egy újabb file-ban tárolja: ez az un rekordleirás. Esetünkben ez a következőképpen fest:

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

     ENTRYNEV        TIPUS SZAM   HELY  HOSSZ OFFSET

  1: KO1D              N     1       0     2
  2: JELENTE1S         M     1       5    55
  3: MEGJEGYZE1S       X     1      60    20
  4: KO1D_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 tipus), de jelentése bármilyen alfanumerikus karaktert tartalmazhat (mixed, azaz M tipus). A megjegyzés rovatban akár speciális karaktereket is használhatunk (extra, azaz X tipus). Ha a kódra és jelentésére egyszerre akarunk rákérdezni, azaz a rekordokat az elsőtől a hatvanadik pozicióig tekintjük, akkor egy olyan mezőre van szükségünk, ami már másutt is emlitett pozíciókból épül fel (overlay, azaz O tipus).

Az F-ben tárolt információ a gyakoriságról meglehetősen durva, de ezért cserébe igen megbizható 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 függenek a választott mintától, hiszen már egyetlen ivnyi adat hozzáadása számos szó abszolút gyakoriságát emelheti 0-ról 1-re vagy 1-ről 2-re – a nagyobb abszolút gyakorisági szavak *relativ* gyakorisága természetesen nem fog lényegesen megváltozni.

Átvettünk a GyakSz.-ból néhány olyan kódot is (szófaj, T-kód, homonimia-kód), ami az egyes homonimák azonositását könnyiti meg: tekintve, hogy a homonimák beosztása a két anyagban nem ugyanolyan, ezek összefésülése csak kézi munkával, esetről esetre haladva lesz megvalósitható. 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 szerepeltek, de a GyakSz.-ban igen.

A SZOTA1R file rekordleirása 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:	TO3TI1PUS	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	O	1	0	1	
20:	SZO1_2_BETU3	O	1	0	2	
21:	SZO1_3_BETU3	O	1	0	3	
22:	SZO1_4_BETU3	O	1	0	4	
23:	SZO1_5_BETU3	O	1	0	5	
24:	SZO1_10_BETU3	O	1	0	10	
25:	SZO1_12_BETU3	O	1	0	12	
26:	SZO1FAJ	O	3	33	1	1
27:	FO3SZO1FAJ	O	1	33	1	
28:	ATERGO	O	1	0	31	
29:	ATERGO_1	O	1	0	31	
30:	ATERGO_2	O	1	0	31	
31:	ATERGO_3	O	1	0	31	
32:	ATERGO_4	O	1	0	31	
33:	ATERGO_5	O	1	0	31	
34:	NOMRAG	O	1	42	6	
35:	C1TERGO	O	1	60	31	
36:	C2TERGO	O	1	60	31	
37:	C3TERGO	O	1	60	31	
38:	C4TERGO	O	1	60	31	
39:	C5TERGO	O	1	60	31	
40:	CVTERGO	O	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	O	1	0	110	
46:	C1BETU3	O	1	60	31	
47:	C2BETU3	O	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 techikai okokból van szükség: ezek segitségével lehet *a tergo,* illetve röviditett (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 átalakitani, 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 összefüggésben vizsgálom.

A nagyszótári munkálatok a magyar lexikográfia, sőt, talán az egész magyar nyelvtudomány ezidáig legnagyobb vállalkozását jelentik. Ezt nem csupán az erre a célra betervezett hatalmas pénzösszegek (melyek egy részéből a Nyelvtudományi Intézet kifejezetten erre a célra dedikált számitógép vásárlását tervezi), hanem az összességében több mint 100 évet átfogó munkafolyamat is jól mutatja. E tanulmány keretei nem teszik lehetővé, hogy a vállalkozás eddigi sikereiről és jövőbeni terveiről részletesen irjak (nem is érzem magam hivatottnak erre) – megelégszem annak bemutatásával, kódról kódra, hogy a SZOTA1R egyes mezőinek kiegészitése és/vagy átalakitása a nagyszótár szempontjából mit jelent. Mivel a SZOTA1R eléggé rugalmasan alakitható, a nagyszótár pedig (már csak méreteinél fogva is) meglehetősen nagy tehetetlenséggel bir, ez utóbbi tervét adottnak veszem. Bár nem elképzelhetetlen, hogy ezek a tervek esetleg mégis módosulnak, úgy tűnik, hogy a nagyszótári munkálatokban következetesen érvényesiteni fogják a következő alapelveket: 1. A munka empirikus 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 rögzitésre is került) a számitógépesitést lényegében elkerülhetetlenné teszi. Véleményem szerint azonban hiba lenne 2-t csupán szükséges rossznak tekinteni.

A magyar nyelv agglutináló természete miatt a szövegszavak szócikkekbe való csoportositásánál 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 közvetitse, illetve annak munkálataiban felhasználhatóvá tegye. Sajnos a VégSz. (tehát eredetileg az ÉrtSz.) szófajkódjai nem teljesen felelnek meg a számitógép szabta precizitási és homogeneitási követelményeknek. Ez jól látszik a GyakSz. és a VégSz. szófajkódjainak összehasonlitásából. Becslésem szerint a kódok legalább 2-3%-át módositani 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 igy 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. minősitéseit, de összetételek esetén saját nyelvérzékükre hagyatkoztak

(ld. VégSz. 20-21.o.). A debreceni kód a magyar morfológia kutatójának páratlanul érdekes adathalmazt kinál: érdemes lenne az újonnan bekerülő szavakat (tehát pl. a GyakSz. szavait) is minősittettni 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 empirikus adatgyűjtés (amely véleményem szerint a VégSz. egyik legpozitivabb vivmánya) azonban kevéssé illeszthető bele a nagyszótár irott 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 cimű munkájában a ÉKSz. egész anyagán végigvitt.

Bizonyos kódok, mint például a szóhossz vagy a CV-csontváz, automatikusan vagy majdnem automatikusan generálhatók. Más kódok, igy például a szavak eredetét tükröző etimológiai kód előállitása azonban igen komoly emberi munkát igényel. E tevékenység automatizálására a közeli jövőben nem is gondolhatunk. A SZOTA1R azonban az ilyen kódok megállapitásához is adhat segitséget azzal, hogy közvetiti a magyar lexikográfia eddig elért eredményeit. Jelen formájában ugyan csak a Bárczi-féle Szófejtő Szótár adatait tartalmazza (ld. VégSz. 24.o.), de nem lenne nehéz kiegésziteni a TESz. adataival sem, hiszen ezek egy részét a debreceni munkacsoport már lekódolta. Ugyanez mondható a stilusminősitésekre is: célszerűbbnek tűnik az ÉrtSz. kódjait módositani, 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 bővitmények) 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 állitás 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" számitógépes rendszerekben felhasználni, mert körkörösek, a szemetet a hulladék, a hulladékot pedig a szemét segitségével definiálják. Ezt a hibát csak akkor lehet elkerülni, ha a szótár készitői előre rögzitenek egy alapszókincset, és minden egyéb szót ennek segitségével értelmeznek. (Ez az eljárás persze nemcsak a számitógép, hanem a nyelvet tanuló más anyanyelvű diák feladatát is hatalmas mértékben megkönnyiti.) Ilyen elven készült például a Longman Dictionary of Contemporary English (számitógépes felhasználását ld. Alshawi – Boguraev – Briscoe 1985) – ennek alapszókincséből alakitottam ki egy ALAP DATA nevű file-t. A számitógépes 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 számitógépek számára is hozzáférhetőek legyenek.

4 Irodalom

Alshawi, H. – Boguraev, B. – Briscoe, T. 1985: A dictionary support environment for real time parsing. In: Proc. 2nd Conference of the European Chapter of the Association for Computational Linguistics, 171-8

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

Éltető L. 1985: Uj adatbáziskezelő rendszer VM/CMS alatt. Információ – Elektronika, megjelenés alatt.

Papp F. 1969a: A Magyar Nyelv Szóvégmutató Szótára. Akadémiai

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

Prószéky G. 1985: Automatizált morfológiai elemzés a nagyszótári munkálatokban. Kézirat, MTA Nyelvtudományi Intézet

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