# Készülő szótárak mint adatbázisok

Két szerkesztés alatt álló szótár példáján szeretném bemutatni, milyen előnyöket nyújthat a számítógép azoknak a lexikográfusoknak, akik szótáraikat úgy készítik, hogy nem csupán okos írógépként használják a számítógépet, hanem ezzel egyidejűleg egyszersmind szótári adatbázist is létrehoznak. Egyik példánk a francia–magyar együttműködéssel készülő francia–magyar/magyar–francia szótár, a másik néhai Vázsonyi Endre Túl a Kecegárdán. Calumet-vidéki amerikai magyar szótár-a, amelyet a szerző halála (1986) után Kontra Miklós fejezett be és szerkesztett meg.

## 1. Számítógépes szótár – szótári adatbázis

Ma már szinte minden könyv, így a szótárak kiadásához is segítségül hívják a számítógépet. Már a szerzők maguk is többnyire számítógépen írják, szerkesztik műveiket. A nyomdászok rendszerint ezek kinyomtatott verziójából állítják elő a végleges, kiadásra kerülő formátumot. A szerzők vagy a nyomdászok által gépre vitt szöveg azonban általában nem alkalmas arra, hogy a nyomtatáson kívül más célra, például elektronikus verzió kiadására is felhasználják. Ha ilyen módon gépre vitt szótárban szeretnénk keresni a különböző szótári elemek között (például a szavak etimológája vagy használati minősítése szerint), kereséseink vagy sikertelenek lennének, vagy oda nem tartozó elemeket is tartalmaznának, amelyek közül csak fáradságos munkával válogathatnánk ki a releváns adatokat.

Ha azonban adatbázisként rögzítünk egy szöveget/szótárat, az annyit jelent, hogy az egyes szerkezeti elemeket (az adatbázis mezőit) valamilyen módon megjelöljük, azonosítjuk. Ismerünk olyan adatbáziskezelő programokat (pl. relációs adatbáziskezelők), amelyeknél az egyes mezők hossza, a bennük tárolható adatok típusa jól meghatározott, s a kitöltendő mezők száma állandó. Ezek bizonyos célokra kiválóan megfelelnek, azonban szöveges adatok — így szótárak — kezelésére nem alkalmasak, hiszen a szótárban az egyes szócikkek hossza, a bennük előforduló elemek száma rendkívül változó. E probléma megoldására született az ún. nyelvtan által

definiált adatbázis fogalma (Gonnet–Tompa 1987). Nyelvtannak ez esetben azt a szabálygyűjteményt nevezzük, amely leírja, milyen elemek lehetnek az adatbázisban, azok hogyan ágyazódhatnak egymásba, mely elemek kötelezőek, melyek ismétlődhetnek stb. Ha egy szótár szerkesztésének kezdetén megfogalmazzuk a szerkesztői utasítás alapján a szótár „nyelvtanát", akkor biztosítható, hogy az előre meghatározott elveket valamennyi szócikkíró betartsa. Ezen túlmenően, már a szerkesztés alatt számos olyan ellenőrzési lehetőséget nyújthat a számítógép, amelyre másképp nem, vagy csak igen fáradságos munkával lenne lehetőségünk. Ezek közül kívánok itt néhányat bemutatni.

## 2. A francia–magyar/magyar–francia szótár

A szótár a párizsi III. egyetem Centre Interuniversitaire d'Études Hongroises és a szegedi József Attila Tudományegyetem francia tanszékének együttműködésében készül: a magyar–francia rész Párizsban Szende Tamás irányításával, a francia–magyar rész Szegeden Pálfy Miklós irányításával. A munkálat tudományos vezetője Jean Perrot professzor. Az MTA Nyelvtudományi Intézete a számítógépes megvalósításhoz nyújt szakmai segítséget.

A magyar szótárak közül ez az első, amely a szerkesztés kezdetétől a nyelvtan által definiált adatbázis koncepciójának felhasználásával készül. Az első szerkesztési utasítás elkészülte után definiáltuk a szócikkek nyelvtanát, amelyet később a tapasztalatok alapján módosítottunk, továbbfejlesztettünk (1. ábra).

```
Diction ((art|ren)+)
art     (ent, (bgr*|bls+), lfg?, ifs?)
ent     (vdt, pho?, mor?, cgr?, rct?, mae?)
bgr     (cgr, mae?, bls+, lfg?, ifs? )
bls     (gen?, rct?, mae?, lig?, ids?, beq*, (rct?, (exp, trd+)*)*, lfg?)
beq     (ids?, (eqv, gen?)+, mae?)
exp     (mae?, ids?)
trd     (mae?, ids?)
lfg     ((exp,trd+)+)
mae     (dds?, rdl?, lig?)
ren     (vdt, rvd)
```

| | | |
|---|---|---|
| art | article | szócikk |
| ren | renvoi | utalás |
| ent | entrée | szócikkfej |
| vdt | vedette | címszó |
| bgr | bloque grammatical | grammatikai egység |

| | | |
|---|---|---|
| bls | bloque sémantique | szemantikai egység |
| lfg | locution figée | állandósult szókapcsolat |
| ifs | information supplémentaire | kiegészítő információ |
| ent | vedette | címszó |
| pho | phonétique | kiejtés |
| mor | morphologie | morfológiai információ |
| cgr | catégorie grammaticale | szófaj |
| rct | rection | vonzat |
| mae | marque d'emploi | használati minősítések |
| gen | genre | nem, genus |
| lig | limitation géographique | földrajzi minősítés |
| dds | domain de spécialité | szakmai minősítés |
| rdl | régistre de langue | stílusminősítés |
| ids | indication sémantique | jelentésmegszorítás |
| beq | bloque équivalent | ekvivalens blokkja |
| egv | équivalent | ekvivalens |
| exp | exemple | példa |
| trd | traduction | fordítás |
| rvd | renvoi vedette | utaló szócikk |

**1. ábra**

A fenti ábrában a nyelvtan baloldalán lévő elemek a jobboldali elemeket tartalmazhatják, az ott leírt sorrendben. Kérdőjellel az opcionális elemeket jelöljük, „+" követi azokat, amelyek legalább egyszer előfordulhatnak, de több is lehet belőlük, „*"-gal a nem kötelező, de esetleg többször is előforduló elemek vannak jelölve. A jelöletlen elemek kötelezően egyszer fordulnak elő.

Az így definiált adatbázisok számítógépes rögzítésére kialakult egy egységes, szabványos jelölési rendszer, az ún. SGML (Standard Generalized Markup Language). Ennek leglényegesebb tulajdonsága, hogy az egyes mezők (szerkezeti elemek) elejét a mező nevének rövidítése jelzi „< >" zárójelpárban, végét pedig ugyanez egy „/" jellel bővítve. Így a szócikket az <ART> </ART> jelek, a szócikk fejet a <ENT> </ENT> jelek zárják közre. Egy példaszócikk a szótár francia–magyar részéből a 2. ábrán látható.

```
ART><ENT><VDT>agricole</VDT> <CGR>adj</CGR></ENT>
<BLS><BEQ><EQV>földművelő:</EQV> </BEQ>
<EXP>peuple/population ~ </EXP> <TRD>földművelő
nép/népesség</TRD> </BLS> <BLS> <BEQ>
<EQV>mezőgazdasági:</EQV> </BEQ> <EXP>pays/coopérative
~ </EXP> <TRD>mezőgazdasági ország/szövetkezet;</TRD>
<EXP>outils/machines/produits/travaux agricoles</EXP>
```

<TRD>mezőgazdasági szerszámok/gépek/termékek/munkák;</TRD>
<EXP>lycée ~ </EXP> <TRD><mezőgazdasági
szakközépiskola>;</TRD> <EXP>ingénieur ~ </EXP>
<TRD>mezőgazdász;</TRD> <EXP>petite/grande exploitation
~ </EXP> <TRD>kis/nagygazdaság</TRD></BLS> </ART>

**2. ábra**

Ez a formátum az ember számára igen nehezen áttekintehető, a számítógép számára azonban így egyértelműen meghatároztuk az egyes mezők elejét végét, a különböző elemek egymásbaágyazottságát. Nem mindegy például, hogy egy stilisztikai minősítés a szócikk fejrészében, vagy az ekvivalensben, esetleg egy fordításban szerepel.

Ebből a formából egy program segítségével automatikusan állítható elő az élvezhető nyomtatott formátum (3. ábra).

**agricole** *adj* 1 földművelő: **peuple/population** ~ földművelő nép/népesség **2** mezőgazdasági: **pays/coopérative** ~ mezőgazdasági ország/szövetkezet; **outils/machines/produits/travaux agricoles** mezőgazdasági szerszámok/gépek/termékek/munkák; **lycée** ~ <mezőgazdasági szakközépiskola>; **ingénieur** ~ mezőgazdász; **petite/grande exploitation** ~ kis/nagygazdaság

**3. ábra**

Külön előnye az ilyen formában való tárolásnak, hogy a végleges nyomtatási képet elég közvetlenül a nyomtatás előtt eldöntenünk, az átalakító program különböző variációk kipróbálására is lehetőséget adhat, majd a végleges változatot automatikusan előállítja.

A szótár 1994. őszéig elkészült anyagán próbakereséseket végeztünk, most ezekből mutatok be néhány eredményt.

| | magyar–francia | francia–magyar |
|---|---|---|
| VDT | 12 131 | 6 625 |
| | a,b,c,d,e,g,gy,h,j | a,c,d,e |
| BLS | 14 980 | 12 312 |
| BGR | 763 | 1 862 |
| EQV | 21 106 | 13 128 |
| LFG | **582** | 507 |

**4. ábra**

A fenti összehasonlító táblázat a két szótári rész főbb elemeinek számát mutatja be. Mint látjuk, a francia–magyar (továbbiakban F–M) szótár általunk tesztelt része kb. fele annyi szócikket tartalmazott, mint a magyar–francia (M–F), a feldolgozott kezdőbetűk számával összhangban. Ehhez képest meglepőnek tűnik, hogy a szemantikai blokkok száma a két részben csaknem azonos, a grammatikai blokkok száma pedig éppenséggel nagyobb a F–M részben! Ez feltehetően elsősorban annak köszönhető, hogy a francia szavaknak sokkal nagyobb része többszófajú, ebből adódik aztán a szemantikai blokkok „elszaporodása" is. Szintén érdekes tapasztalni, hogy az állandósult szókapcsolatok száma a két szótárfélben gyakorlatilag azonos.

| | | magyar–francia | francia–magyar |
|---|---|---|---|
| DDS | | 2 919 | 2 797 |
| ebből | ENT | 2 127 | 580 |
| | BGR | 91 | 459 |
| | BLS | 776 | 2 173 |
| | BEQ | 7 | 13 |
| | EXP | 91 | 212 |
| | TRD | 6 | 4 |

**5. ábra**

A szakmai minősítések összehasonlító táblázatát láthatjuk az 5. ábrában. Megint csak az tűnik fel, hogy az összes ilyen minősítés száma gyakorlatilag megegyezik; a M–F részben a szavak jelentős részében a szócikk fejrészében találhatjuk ezeket, míg a F–M-ban inkább a szemantikai blokkon belül. Ezt a tényt csak részben magyarázza a grammatikai és ebből következőleg a szemantikai blokkok nagyobb száma. Mint kiderült, a F–M-ban kevésbé helyeztek súlyt arra, hogy az olyan szavaknál, ahol a minősítés a szócikk egészére vonatkozik, a fejrészben helyezzék el a minősítést. Éppen ebből az összehasonlításból vehette észre a F–M szerkesztője, mi a jelentősége egy ilyen döntésnek.

A felsorolt lehetőségek bármelyikéről automatikusan listát készíthetünk, például kiválogattathatjuk a géppel az összes „Chim" (kémia) minősítésű szót, vagy az összes olyan szócikket, amelynek fejrészében szakmai minősítés szerepel. Egy ilyen listából láthatunk szemelvényt a 6. ábrán.

```
>> DDS within region ENT
>> 580 matches
abrogeable    <CGR>adj</CGR> <DDS>Jur</DDS></ENT>
acronyme      <CGR>nm</CGR> <DDS>Ling</DDS></ENT>
```

admonition <CGR>nf</CGR> <DDS>Jur</DDS></ENT>
2 affleurement <CGR>nm</CGR> <DDS>Techn</DDS></ENT>
akkadien <DDS>Hist</DDS></ENT>
allitération </VDT> <CGR>nf</CGR><DDS>Litt</DDS></ENT>
ammoniaque <CGR>nf</CGR> <DDS>Chim</DDS></ENT>

**6. ábra**

A következő összehasonlító táblázatban a stílusminősítések megoszlását láthatjuk. Ebben is hasonló jelenségek figyelhetők meg, mint a szakmai minősítések használatában. Így azt állapíthatjuk meg, hogy az egyes szótári részek legalább önmagukban következetesek.

| | | magyar–francia | francia–magyar |
|---|---|---|---|
| RDL | | 2 609 | 2 595 |
| ebből | ENT | 1 206 | **203** |
| | BEQ | 310 | **463** |
| | EXP | 237 | **413** |
| | TRD | 258 | **469** |
| | LFG | 136 | **211** |

**7. ábra**

A stílusminősítésekről is különféle csoportosítású listák készíthetők, itt például a választékos (soutenu) minősítésű szavakból láthatunk egy rövid részletet.

>> soutenu within region RDL
>> 461 matches
akármint .. <RDL>soutenu</RDL> </ENT> <BLS> <IDS> (mindegy, hogy)
álokoskodás .. </CGR> <RDL>soutenu</RDL> </ENT><BLS>
árulkodik .. </RCT> <RDL>soutenu </RDL><IDS>(vmire vall/utal)
avul .. <RDL>soutenu </RDL> </ENT> <BLS> <BEQ> <EQV> vieillir;
bealkonyodik .. /CGR><RDL>soutenu </RDL> </ENT> <BLS> <BEQ>
búbánat .. </CGR><RDL>soutenu </RDL> </ENT> <BLS> <BEQ>
csapodár .. </EQV><RDL>soutenu </RDL> </BEQ> </BLS> </ART>
csorbít .. </CGR><RDL>soutenu fig </RDL> </ENT> <BLS> <BEQ>

dalol .. </CGR><RDL>soutenu </RDL> </ENT> <BLS> <IDS>
(énekel)
deres .. </CGR><RDL>soutenu </RDL> <BLS> <EXP> ~ haj

**8. ábra**

A fenti csoportosításokon túlmenő, a szótári adatbázisok által nyújtott lehetőség, hogy egy szó összes előfordulását kikereshetjük a szótár egészéből, illetve annak bármely mezőjéből. A 9. ábrában láthatjuk, hogy először kikerestük a *faire* ige összes előfordulását a M–F részből, majd azokat, ahol az ekvivalensek között fordul elő a *faire*.

>> faire
>> 937 matches
>> faire within region EQV
>> 403 matches
>> pr sample
áthajt ..<EQV>faire traverser qc à qc/qn
autózgat ..<EQV>faire de la voiture
bedolgoz ..<EQV>faire pénétrer dans
bekakál ..<EQV>faire dans son froc</EQV
bemesél ..<EQV>faire avaler qc à qn:
beszólít ..<EQV>faire entrer:
bevásárol ..<EQV>faire des/les/ses courses:
bolondozik ..<EQV>faire le fou;
céloz ..<EQV>faire allusion à qc/qn:
csicsikál ..<EQV>faire dodo
csúszkál ..<EQV>faire des glissades:
dorbézol ..<EQV>faire la bringue
édeleg ..<EQV>faire sa cour à qn
fektet ..EQV>(faire) coucher qn qpart;
felmos ..<EQV>faire revenir qn de son évanouissement avec de l'eau..
felsül ..<EQV>faire chou blanc;

**9. ábra**

Számomra meglepően sok ilyen magyar címszót találtunk (12 000-ből 400!), itt vannak mindazok az igék, amelyeknek nincs francia igei megfelelőjük. Ebből a listából kiválasztottam egy-két olyan szót, amely a szótár másik felében már elvben megtalálható lehet (10. ábra).

>> dodo within region VDT
>> one match
>> pr
<ART><ENT><VDT>*dodo* </VDT><CCGR>nm </CCGR></ENT>
<BLS> <IDS>(langage enfantin) </IDS><BEQ><IDS>(a gyermeknyelvben)
</IDS><EQV>alvás; </EQV></BEQ><BEQ><EQV>csicsiskálás </EQV>
<RDL>langage enfantin: </RD..

>> dorbézol within region EQV
>> one match
>> pr
<ART><ENT><VDT>*débauche* </VDT><CCGR>nf </CCGR></ENT>
<BLS> <IDS>(vice) </IDS><BEQ><EQV>*dorbézolás*/kicsapongás: </EQV>
</BEQ><EXP>mener une vie de ~ </EXP><TRD>kicsapongó életet él;
</TRD><EXP>partie de ~ </EXP><TRD>orgia; </TRD><EXP>exciter
des mineurs à la ~ </EXP><TRD>fiatalkorúakat megront </TRD><LFG>
<EXP># faire une petite ~ <RDL>fam </RDL></EXP><TRD>kissé kirúg a
hámból <RDL>fam </RDL></TRD></LFG></BLS><BLS> <IDS>(usage
déréglé) </IDS><EXP> ~ d'imagination </EXP><TRD>korlátlanul

>> fektet within region EQV
>> 2 matches
*aliter* ..> <RCT>SOUVENT AU PASS</RCT> <IDS>(un malade)</IDS>
<BEQ><IDS>(beteget)</IDS> <EQV>ágyba *fektet*;</EQV> <EQV>ágynak dönt:</EQV></BEQ> <EXP>une mauvaise grippe l'avait alitée pendant quinze jours</EXP> <TRD>egy csúnya influenza két hétre ágynak döntötte; </TRD> <EXP>malade alité</EXP> <TRD>fekvőbeteg;</TRD>
<EXP>être/rester alité (depuis longtemps)</EXP> <TRD>(hosszú ideje) nyomja az ágyat</TRD></BLS></BGR> <BGR> <CCGR>v pron</CCGR>
<BLS> <IDS>(malade)</IDS> <BEQ><IDS>(beteg)</IDS> <EQV>ágynak dől/az ágyat nyomja</EQV> </BEQ> </BLS> </BGR> </ART>
<ART><ENT><VDT>alité</VDT> <DDS>Mét..
*carénage* ..T><CGR>nm </CGR> </ENT><BLS> <DDS>Navig </DDS>
<IDS>(action) </IDS><BEQ><EQV><hajó oldalra *fektetése*/megdöntése javítás céljából> </EQV></BEQ></BLS><BLS> <DDS>Navig </DDS><IDS>
(lieu) </IDS><BEQ><EQV>szárazdokk/sólya </EQV></BEQ></BLS>

**10. ábra**

A *dodo* címszó a F–M részben főnévként szerepel. A *dorbézol* igét az ekvivalensek között keresve a *débauche* szócikket találtam meg, azaz nem ugyanazt a megfelelőt, amelyet a M–F-ben láthattunk (*faire la bringue*). A

*fektet*-et keresve az ekvivalensek között szintén nem találtam meg a *coucher* szócikket, csupán az *aliter* és a *carénage* címszókat, de ennek inkább az lehetett az oka, hogy a *coucher* még nem szerepelt a kiválasztott mintában.

A másik irányból indulva is kísérletet tettem a címszavak és ekvivalensek kölcsönös megfelelésének ellenőrzésére. A F–M rész alábbi listájából vizsgáltam a dőlt betűvel szedett szavakat (11. ábra).

accumuler ..<EQV>(fel-/össze)gyűlik:</EQV>
*affinement* ..<EQV>finomítás/*csiszolás*</EQV>
aligner ..<EQV>felsorakozik;</EQV>
analytique <EQV>pszichoanalitikus</EQV>
*calme* ..<EQV>*békés* ..
caractéri ..<EQV>jellemzést ad vkiről/vmiről:
déceler ..<EQV>vmi vmiről árulkodik
dégorger ..<EQV>kitisztít
*disconvenance* ..<EQV>*aránytalanság*;
*enrégimenter* ..<EQV>*beléptet vkit*

>> affinement within region EQV
>> no match
>> csiszolás within region VDT
>> one match
<ART><ENT><VDT>csiszolás</VDT> <CGR>n </CGR> </ENT>
<BLS><BEQ><EQV>polissage </EQV><GEN>m; </GEN></BEQ>
<BEQ><IDS>(koronggal) </IDS><EQV>meulage </EQV><GEN>m;
</GEN></BEQ> <BEQ><IDS>(habkővel/üvegpapírral)
</IDS><EQV>ponçage </EQV><GEN>m; </GEN></BEQ>
<BEQ><IDS>(arany; fém) </IDS> <EQV>brunissage
</EQV><GEN>m; </GEN></BEQ> <BEQ><IDS>(kő/márvany)
</IDS><EQV>égrisage </EQV><GEN>m; </GEN></BEQ>
<BEQ><IDS>(fényesre) </IDS><EQV>lustrage </EQV><GEN>m;
</GEN></BEQ> <BEQ><IDS>(parketta) </IDS><EQV>passage
</EQV><GEN>m </GEN><EQV>à la paille de fer </EQV></BEQ>
</BLS><BLS><RDL>fig </RDL><IDS>(stílusz)
</IDS><BEQ><EQV>peaufinage </EQV><GEN>m;
</GEN><EQV>fignolage </EQV><GEN>m
</GEN></BEQ></BLS></ART>

>> enrégimenter within region EQV
>> no match
>> beléptet within region VDT

```
>> one match
<ART><ENT><VDT>beléptet </VDT></ENT><BGR><CGR>v tr
</CGR><BLS><IDS>(belépésre kényszerít) </IDS><BEQ><EQV>faire
entrer; </EQV> <EQV>obliger qn à entrer: </EQV></BEQ>
<EXP>beléptették a szakszervezetbe </EXP> <TRD>ils l'ont fait entrer au
syndicat </TRD></BLS></BGR>
<BGR><CGR>v intr </CGR> <BLS><IDS>(lóháton)
</IDS><BEQ><EQV>entrer à cheval </EQV> <IDS>(au pas)
</IDS></BEQ> </BLS></BGR></ART>
```

**11. ábra**

A keresett szavak közül egyedül a *calme – békés* szópár fordult elő mindkét részben egymás megfelelőjeként. Az *aránytalanság – disconvenance* szópárt a M–F-ban nem sikerült megtalálnom, az *affinement* szó az ekvivalensek között nem szerepelt, a *csiszolás* szónak pedig számos ettől eltérő ekvivalensét láthatjuk a szócikkben.

Az *enrégimenter* szót szintén nem találtam meg az ekvivalensek között, a *beléptet* szó megfelelői között pedig megint csak más megoldások szerepelnek.

Úgy gondolom, ebből a néhány példából jól látható, milyen sokféle ellenőrzési lehetőséget biztosít a szerkesztőknek már munka közben az adatbázis formátumban való bevitel. Természetesen a szótár leendő használóinak is rendelkezésére fognak állni ugyanezek a lehetőségek, amennyiben annak elektronikus verzióját veszik igénybe.

### 3. Az amerikai magyar szótár

Ezt a szótárat az Amerikai Egyesült Államokban dolgozó Vázsonyi Endre kezdte el szerkeszteni saját gyűjtései alapján az ott élő magyarok sajátos nyelvhasználati szokásairól. Halála után az ő gyűjteményének felhasználásával Kontra Miklós fejezte be a szótár szerkesztését.

Számítógépes szempontból az anyag bevitelének módja kevésbé volt ideális, mint a M–F/F–M esetében, mivel a szerkesztő akkor az Egyesült Államokban tartózkodott, folyamatos konzultációra kevés lehetőségünk volt. Így az anyag első változatában „fél-SGML" formátumban lett rögzítve, ami ugyan lényegesen jobb, mintha a szövegszerkesztőt egyszerűen csak írógépként használták volna, de ez a forma számos átalakítást tett szükségessé, amely hibák forrásává vált.

<SZO><CIM>abstéz, abstóz,
abtéz<\CIM><FAJ1>fn<\FAJ1><JE1>'egyemeletes ház emeleti
része'<\JE1><MO1>(<*upstairs* <FAJ2>hsz<\FAJ2> 'fent,
emeleten')<\MO1> <GY1>GyÁ<\GY1><PL1>Abstézen laktak [Sné], Még
azt mondják: „Gyere, John, az abstézre" [HM], Abstózon voltak mindig burdosok
[Nné], Abtézen volt egy rúm [FJ]<\PL1><VO1>Vö. dánstéz<\VO1><\SZO>

**12. ábra**

Mitől „fél-SGML" a fenti, ránézésre a másik szótárhoz igen hasonló leírási mód? A leglényegesebb és legtöbb hibaforráshoz vezető eltérés: az előforduló elemek egy része nincs tagolójellel azonosítva. A MO rövidítésű modellben kötelezően előforduló angol megfelő például csak kurzívval van jelölve. Mivel azonban kurziválást a szerkesztő más esetekben is alkalmazott, ez az elem nem volt automatikusan azonosítható. A fenti példában nem látható, de ugyanilyen problémát okozott a homonimaszámok és jelentésszámok jelölése is, a homonimaszám csak a szám felemelésével, a jelentésszám pedig csak félkövér szedéssel volt megkülönböztetve. Szintén nem volt szerencsés az adatközlők jelölése: „[ ]", ezeket a zárójeleket viszont csak erre a célra használta a szerkesztő, így tagolójellé való átalakításuk egyszerű és egyértelmű volt.

A kereséshez és a nyomtatáshoz szükséges konverzióhoz zavaró és szükségtelen lett volna a tagolójelek szócikkeken belüli sorszámának megtartása, ezért ezeket az első konverzió során megszüntettük.

Mivel a keresőprogram jelenlegi verziója az általunk használt számítógépen az ékezetes betűket nem tudja kezelni, a konverziónál az ékezetes karaktereket átalakítottuk egy sajátos kódolási konvenció szerint (ún. Prószéky kód), ahol az ékezeteket az angol ábécé betűi után írt számok jelölik. (pl. á=a1, é=e1, ö=o2, ő=o3 stb.) A tagolójeleket egységes angol rövidítésből származó jelekre cseréltük, ahol pedig lehetett, ott az SGML által javasolt rövidítést vettük át. A fenti szócikk konverzió utáni formáját a 13. ábrán láthatjuk, az átalakított szótár nyelvtanával együtt.

<ART> <LEM>abstelz, abstolz, abtelz</LEM> <SEN>
<POS>fn</POS> <EQV>'egyemeletes halz emeleti relsze'</EQV>
<MOD> (<ENG> <upstairs</ENG> <POS>hsz</POS> 'fent,emeleten')
</MOD> <FRQ>GyA1</FRQ> <EXA>Abstelzen laktak
<INF>Snel</INF>, Melg azt mondjalk: „Gyere, John, az abstelzre"
<INF>HM</INF>, Abstolzon voltak mindig burdosok <INF>Nnel</INF>,
Abtelzen volt egy rulm <INF>FJ</INF> </EXA>Vo2.
<CMP>dalnstelz</CMP> </SEN> </ART>

```
dic     (art+)
art     (lem, (sen+|xrf))
sen     (snu?, pos, eqv, xrf?, mod?, frq?, exa?, cmp?, rem?, cmp?)
exa     ((inf*,cmp*,eng*)*)
lem     (hom*)
eqv     (eng?)
mod     (eng,pos*,hom*)*)
eng     (hom*)
cmp     (hom*)
xrf     (hom*)
rem     ((eng*, cmp*, inf*,hom*)*)
```

Ahol:

| | | |
|---|---|---|
| dic | dictionary | szótár |
| art | article | szócikk |
| lem | lexeme | címszó |
| pos | part of speech | szófaj |
| xrf | cross reference | utalás |
| snu | sense number | jelentésszám |
| eqv | equivalent | ekvivalens |
| mod | model | átadó nyelvi modell |
| eng | English | eredeti angol szó |
| frq | fequency | gyakoriság |
| cmp | compare (cf) | vesd össze |
| rem | cultural remark | kulturális megjegyzés |
| inf | informant | adatközlő |
| hom | homonym number | homonimaszám |

**13. ábra**

A konvertált szótárat a WRITERSTATION program segítségével ellenőriztük. A program az SGML formátumú szövegek bevitelét segíti elő olymódon, hogy jelzi, hol talált az előre meghatározott nyelvtantól eltérő elemet. Az ellenőrzött és kijavított adatbázist azután ugyanazzal a lekérdező programmal vizsgáltuk, amellyel a M–F/F–M szótárat is.

Megállapítottuk, hogy a szótár 1159 szócikket és ugyanennyi címszót tartalmaz, ezek közül 894-nek van értelmezése, a többi utaló címszó. Az értelmezéssel bíró szócikkekben összesen 942 átadó nyelvi modell szerepel, egyikből sem hiányzik az angol eredeti szó.

Megnéztük, hány címszónak van 3-nál több jelentése: 23 ilyet talált a program, ezek közül 3-nak 4 jelentése, 1-nek 5 jelentése, és 1-nek 7 jelentése van (ez a biznisz címszó).

Ellenőriztük a szótárban használt szófajmegjelöléseket olymódon, hogy felsoroltuk az ismerteket, és kerestük azt a halmazt, amelyik nem a felsorolt elemek valamelyikét tartalmazza a POS mezőben. E lista (14. ábra) segítségével könnyen kiszűrhetők és javíthatók a szerkesztő esetleges következetlenségei.

```
enikájnd <POS> ált névm</POS> <EQV> 'mindenféle'</EQV> <..
enivé    <POS> ksz</POS> <EQV> 'mindenesetre, különben'
genszt   <POS> vsz</POS> <EQV> 'ellen, ellene'</EQV>
hú       <POS> névm</POS> <EQV> 'ki,kicsoda'</EQV>
hukit    <POS> állandósult szókapcsolat</POS> <EQV>
jú       <POS> névm</POS> <EQV>'te'</EQV>
lacó     <POS> htl tőszn</POS> <EQV> 'sok'</EQV>
mébi     <POS> módsz</POS> <EQV> 'talán'</EQV>
óver     <POS> ik</POS> <EQV> 'túl-'</EQV> <MOD>
peda     <POS></POS> <EQV> 'fizetés'</EQV> <EXA>
plenti   <POS> htl tőszn</POS> <EQV> 'sok'</EQV>
ranol    <POS></POS> <EQV> 'választással betöltendő'
súr      <POS> módsz</POS> <EQV> 'biztos'
```

**14. ábra**

Hasonló ellenőrzéseket végeztünk a gyakorisági kódok között: megkaptuk azon címszavak listáját, amelyek gyakorisági kódjának jelentése egyelőre ismeretlen (K, M),* de talán e lista alapján rekonstruálható az első szerkesztő szándéka (15. ábra).

```
braunsugo .. <FRQ> K</FRQ> <EXA> Braunsugort tesz a tetejébe a..
briccs ..    <FRQ> K</FRQ> <EXA> Az ő bridzsén akartam én
bucser ..    <FRQ> K</FRQ> <EXA> Kitanulta a bucsert, vett
cirkulál ..  <FRQ> K</FRQ> <EXA> Nem cirkulálódik a vér
csekk ..     <FRQ> K</FRQ> <EXA> A bátyám csekkjére
dempt ..     <FRQ> M</FRQ> <EXA> Abban a bányában dempt nem
diggol ..    <FRQ> M</FRQ> <EXA> Bediggoltam a kokszrakás
ekszmarí ..  <FRQ> K</FRQ> <EXA> Ekszmaríner volt <INF>
es ..        <FRQ> M</FRQ> <EXA> Leejtette az est <INF>
eszid, es .. <FRQ> M</FRQ> <EXA> Az ólomnak az eszidben
```

**15. ábra**

* Vázsonyi Endre ötféle gyakoriság jelölő rövidítést használt: R(itka), Gy(akori), Á(ltalános), valamint: K és M. Ez utóbbiak feloldását a sajtó alá rendező Kontra Miklós nem tudta megtalálni Vázsonyi hagyatékában.

A következő listában az adatközlők között előfordult „gyanús elemekből" láthatunk egy részletet (16. ábra).

248139, ..<INF> egy állam neve</INF> ) ..
218469, ..<INF> Egy nyolcéves gyerek. F..
191159, ..<INF> Elin</INF> , Főburdos =..
176771, ..<INF> Fahrenheit</INF> ') fív..
311680, ..<INF> FD</INF> , A csont elmuf..
241322, ..<INF> FD</INF> , A nép mind e..
118109, ..<INF> 42</INF> , A dípísták..
54608, ..<INF> 51</INF> </EXA> </SEN> <..
117453, ..<INF> 51???</INF> </EXA> </SEN..
335506, ..<INF> A garyi Verhovay Egylet ..
182996, ..<INF> „Gyermekszáj"-történet..

**16. ábra**

A szótár koherenciájának ellenőrzésén túlmenően arra is lehetőségünk van, hogy a programmal ellőállíttassuk az angol modellként előfordult szavak és a hozzájuk tartozó címszavak listáját (17. ábra).

| | |
|---|---|
| anytime | enitájm |
| boss | főbász |
| design | dizejn |
| displaced person | dípí, d .. |
| dizziness | diziség |
| dizzy | dizi |
| doctor | doktorolt |
| double | dóbelez, .. |
| double | dóbelez, .. |
| elemózsia | alimózsi |
| fire-chief | fájerbász |

**17. ábra**

Az ellenőrzések és javítások után egy újabb konverziós program segítségével átalakítjuk a szótárat kinyomtatandó formára (18. ábra).

**abstéz, abstóz, abtéz**

fn 'egyemeletes ház emeleti része' **M:** (< *upstairs* hsz 'fent, emeleten') **Gy:** GyÁ □ Abstézen laktak [Sné], Még azt mondják: „Gyere, John, az abstézre" [HM], Abstózon voltak mindig burdosok [Nné], Abtézen volt egy rúm [FJ] Vö. **dánstéz**

**18. ábra**

## 4. A felhasznált programokról

Munkánk során több programot használtunk. A már említett WRITERSTATION szövegszerkesztő kifejezetten SGML formátumú dokumentumok számítógépre rögzítésének megkönnyítésére készült. Ha ennek segítségével akarunk bevinni egy szöveget, először definiálnunk kell a szöveg nyelvtanát, majd megadnunk a velük végzendő műveleteket. (Pl. a MOD jel helyén jelenjen meg egy félkövér M és egy kettőspont, az INF helyén a szögletes zárójel, az egyes mezők jelenjenek meg különböző színekkel, hogy jobban elkülönüljenek stb.) Ezután a program szerkesztés közben folyamatosan „vezeti a kezünket", a képernyő alsó sorában mindig láthatjuk, milyen szerkezeti elemben, a struktúra mely részében járunk éppen. A lehetséges jeleket pedig az ALT és egy számbillentyű lenyomásával beilleszthetjük; a billentyűk aktuális sorszáma szintén az alsó sorban látható. Ezzel a módszerrel az adatbázisba bevitt szerkezeti hibák mennyisége minimálisra csökkenthető.

Az Amerikai magyar szótár eredetileg nem ezzel a szövegszerkesztővel lett lerögzítve, hanem WordPerfecttel. Ezzel is tökéletes SGML formátumú szöveget lehetett volna előállítani, csupán az lett volna fontos, hogy minden szerkezeti elem jól meg legyen különböztetve tagoló jellel. Ez egyúttal azt is jelenti, hogy SGML formátumú szöveg előállításhoz nem feltétlenül van szükségünk kifejezetten e célra szánt eszközre, voltaképpen bármilyen szövegszerkesztővel előállíthatunk ilyeneket. Csupán arra kell figyelemmel lennünk, hogy a szövegszerkesztő által nyújtott tipográfiai lehetőségeket **n e v e g y ü k i g é n y b e** ! Azaz ne használjunk különféle betűtípusokat és egyéb a szöveget széppé tevő eszközt, hanem **m i n d e n t t a g o l ó j e l e k k e l j e l ö l j ü n k**. A tagolójelek hibás gépelése is elkerülhető, ha billentyű-makrók segítségével visszük be őket. Ennek a megoldásnak egyetlen hátránya a WRITERSTATION-ben való rögzítéshez képest, hogy sokkal könnyebben tévedhetünk a szerkezeti elemek elhelyezésében, mert csupán szemmel tudjuk ellenőrizni, vajon minden szükséges elem benne van-e a szócikkben, és minden a megfelelő helyen van-e. Előnye viszont, hogy nem kell megvásárolnunk az igen drága WRITERSTATION szoftvert, használhatjuk saját, jól bevált szövegszerkesztőnket.

A szótárak lekérdezésére a PAT (Gonnet 1987, Pajzs 1994a, 1994b) programot használtuk, amelyet eredetileg az Oxford English Dictionary kezelésére fejlesztettek ki, ezért kiválóan alkalmas az ilyen formában rögzített adatbázisok elérésére.

A program jelenlegi verziójában (PAT 4.0) igen gyorsan kereshetjük bármilyen szó vagy karaktersorozat előfordulását, és meghatározhatjuk azoknak a tagoló jelekkel ellátott mezőknek a halmazát, ahol a keresést végezni akarjuk. (Így kereshettük pl. a szakmai minősítéseket csak a szócikk fejrészében, vagy egyes szavakat az ekvivalensek között stb.) A kapott eredmények bármelyikét kimenthetjük és felhasználhatjuk bármilyen szövegszerkesztőből, esetleg kinyomtathatjuk. Bizonyos „gyermekbetegségei" vannak azért a jelenlegi verziónak. Egyelőre egyszerre legfeljebb két mező tartalmát tudjuk kilistázni, pl. az ekvivalenst és a hozzátartozó címszót, és még azt is csak töredékesen, csupán a címszó első 10 karakterét. Az eredmények rendezésének lehetősége is szűkös. Hiányzik még a hasonló jellegű lekérdező programok által rendszerint automatikusan előállított szólista, amely a szövegben előforduló valamennyi szóalak előfordulását szokta tartalmazni ábécérendben, mellettük feltüntetve az előfordulás számát. Reméljük, hogy a hamarosan hozzánk kerülő új verzió, amely az Informatikai Infrastruktúrális pályázatból e célra elnyert új SUN gépen fog működni, a fenti problémákra, vagy legalább egy részükre megoldást kínál.

Az általunk használt programokon kívül nagyon sok olyan szoftver van, amely SGML szövegek bevitelét és/vagy lekérdezését támogatja. Számuk évről évre nő, az általuk nyújtott lehetőségek is folyamatosan bővülnek, mivel a kiadók és egyéb szótárak, lexikonok előállításával foglalkozó cégek mindinkább erre a rögzítési–kiadási formára állnak át.

* * *

A szótárak adatbázisként történő rögzítése tehát homogén, könnyen kezelhető és módosítható anyagot eredményez. Ugyanakkor felmérhetetlen változást eredményezhet a lexikológiai, a fonetikai, a morfológiai kutatások terén is, mivel ezzel az eljárással hatalmas anyagon tudunk sokirányú, egységes szempontú, az esetlegességet kiküszöbölő vizsgálatokat végezni. Végül, de nem utolsósorban, ebből a formából automatikusan előállítható a szótár nyomtatott és elektronikus formában való kiadásra szánt változata.

PAJZS JÚLIA

## Irodalom

GONNET, G. (1987), PAT — An efficient text searching system. University of Waterloo, Centre for the New OED.

GONNET, G. – TOMPA, F. (1987), Mind your Grammar: a New Approach to Modelling Text. University of Waterloo, Centre for the New OED.

PAJZS J. (1990), Számítógép és lexikográfia. (Linguistica, Series A: Studia et Dissertationes, 4.) Budapest.

PAJZS J. (1994a), A számítógépes nagyszótári korpusz felhasználásának lehetőségei. MNy 3: 287–302.

PAJZS J. (1994b), A magyar irodalmi és köznyelv nagyszótárának számítógépes megvalósítása. Megjelenés előtt az egri Magyar Nyelvészkongresszus kötetében.

## Electronic dictionaries in the works

by JÚLIA PAJZS

This paper aims to demonstrate the possibilities offered by compiling dictionaries in SGML (Standard Generalized Markup Language) format. Examples from two ongoing projects (a new Hungarian–French/French–Hungarian dictionary, and a "Hunglish" dictionary) are used to show the many ways of selecting specific fields from the computerized texts, and the various procedures for controlling the coherence of the dictionaries.