Próbák és példák a Magyar értelmező kéziszótár (2. kiadás, 2003) rejtett információinak feltárására

Mártonfi Attila

MTA–ELTE Nagyszótári kutatócsoport
Budapest VI., Benczúr u. 33. 1068
rumci@nytud.hu

Kivonat. A tudásfeltárás, illetve ennek részeként az adatbányászat az információtechnológia divatos területei, melyek jellemzően az üzleti adatbázisok hasznosítására irányulnak, személete, eszköztára azonban – legalábbis részben – alkalmazható szótári adatbázisokra is. A VégSz.-ból nyomdatechnikai okok miatt kimaradt betűjegyben mért hosszúság, jelentésszám, etimológia, valamint szócikkfejben adott stílusminősítés pótlásánál az ÉKsz.[2] XML-változata segítségével teljesebb és korszerűbb adattábla hozható létre, ugyanis ez naprakész etimológiai információt nyújt a magyar szókészlet legtágabb köréről, valamint megtalálható benne a Magyar nemzeti szövegtár-beli abszolút gyakorisági érték is. Az így létrehozott relációs adatbázisból egyszerű lekérdezésekkel előállíthatók a különféle etimológiájú, lexikai minősítésű, szófajú vagy jelentésszámú szóhalmazok szótári, valamint – jelentős újdonságként – szöveggyakorisági mutatói. Az adatbányászat eszköztárával feltárhatók a fenti paraméterek közt fennálló rejtett mintázatok asszociációs szabályok kinyerése útján.
Kulcsszavak: lexikográfia, tudásfeltárás, etimológiai statisztika, Magyar értelmező kéziszótár

1. Bevezetés

A tudásfeltárás, illetve ennek részeként az adatbányászat az információtechnológia divatos területei, melyek jellemzően az üzleti adatbázisok hasznosítására irányulnak. Mivel azonban a cél – jelesül nagy adatbázisokból minél több rejtett adat, ismeretlen mintázat gépi úton történő kinyerése – lényegében a tudományos kutatás legáltalánosabb céljának tekinthető, így személete, eszköztára legalábbis részben alkalmazható szótári adatbázisokra is (a szótári adatbázisok mérete rendszerint nagyságrendekkel kisebb lévén az adatbányászat elsődleges területén előforduló hatalmas üzleti adatbázisoknál, a műveletek eszközigénye lényegesen kisebb, a kinyerhető információ azonban természetszerűleg szűkebb körű).

Az első jelentős magyar szótári adatbázis a VégSz. [3], illetve az ebből létrejött PC-s adatállomány (BUT). A szóvégszótár alapjául szolgáló adatbázis, a DT2 négymezőnyi többletet tartalmazott a papíron megjelent változathoz képest, ezek: a betűjegyben mért hosszúság, az ÉrtSz.-beli [2] jelentésszám, etimológia a SzófSz. [1] alapján, valamint az ÉrtSz. szócikkfejben adott stílusminősítése – ezek nyomdatechnikai okokból nem kerültek végül bele a papírszótárba, s így az ebből készült PC-s adatbázisba.

2. Az adatok átalakítása

A Magyar értelmező kéziszótár új kiadása [5], korszerű szótári munkálathoz méltó módon először XML-dokumentumként készült el, s bár grammatikai információi, melyek a VégSz. gerincét jelentik, lényegesen szegényesebbek, megfelelő átalakításokkal a fenti hiányok pótlásánál teljesebb és korszerűbb adattábla hozható létre. Korszerűbb, mert az ÉKsz.[2] naprakész etimológiai információt nyújt a magyar szókészlet legtágabb köréről, és teljesebb, ugyanis a szótár szófaji és lexikai minősítésén, valamint a kidolgozott jelentések számán kívül az adatbázisban minden címszónál megtalálható a Magyar nemzeti szövegtár-beli abszolút gyakorisági érték is, valamint gépi úton kódolható a fonémaszámban, illetve a szótagszámban mért szóhossz is.

Szükséges volt tehát az XML-dokumentumból egy olyan adattáblát létrehozni, mely egyszerű formában szolgáltatja a szükséges információkat, hogy a többé-kevésbé rejtett információk kinyerhetők legyenek. A konverziót követően (hiszen más struktúrában más adathibák tűnnek elő) el lehetett végezni néhány adattisztítási műveletet is. Az átalakított és megtisztított adattábla 72 444 rekordot tartalmazott, rekordonként 10 mezővel.

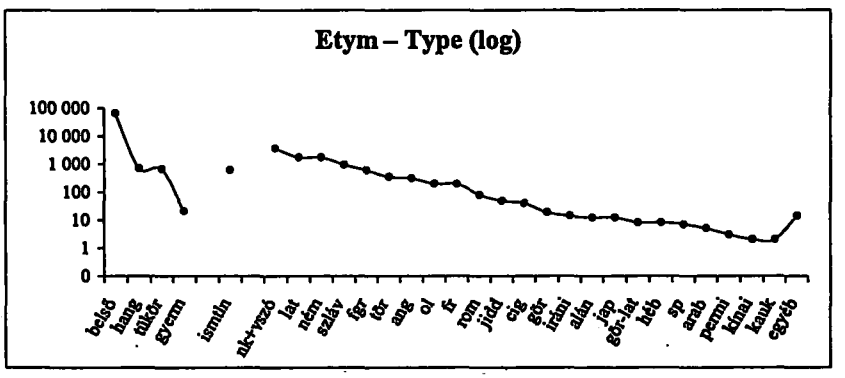
1. táblázat. Az adatbázis néhány rekordja

Azonosító	Lemma	Split	Syll	Phon	Freq	Usg	Pos	Sens	Etym
3053	asszonykerülő	☑	5	11	0		mn fn	1	
3054	asszonykéz	☑	3	8	21		fn	2	
3055	asszonykormány	☑	4	11	0	ritk tréf	fn	1	
3056	asszonymunka	☑	4	10	1	nép	fn	2	
3057	asszonynéni	☑	4	9	0	rég	fn	1	
3058	asszonynép	☑	3	8	62	nép	fn	1	
3059	asszonynév	☑	3	8	33		fn	1	
3060	asszonyos	☐	3	7	34		mn	2	
3061	asszonypajtás	☑	4	11	7	biz tréf	fn	1	
3062	asszonyrokon	☑	4	10	1		fn	1	
3063	asszonyság	☐	3	8	296		fn	3	
3064	asszonytárs	☐	3	9	15		fn	2	
3065	asztag	☐	2	5	92		fn	2	szláv
3066	asztal	☐	2	5	16627		fn	5	szláv
3067	asztalbontás	☑	4	11	4	vál	fn	1	

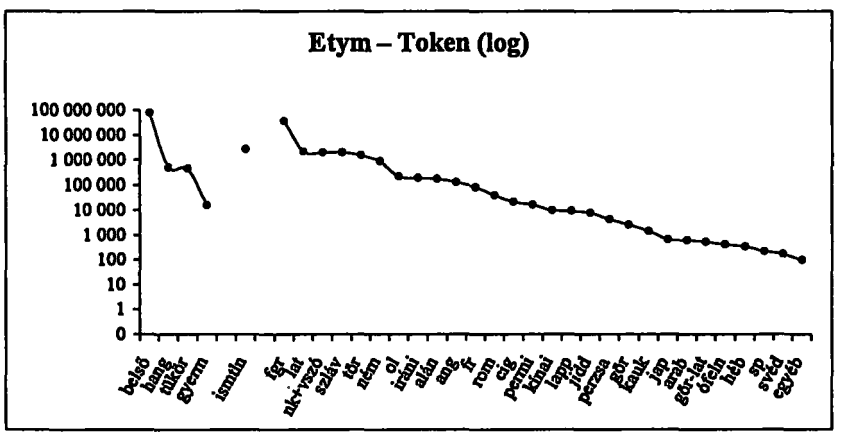
Az egyes mezők a következő információkat tartalmazzák: **Azonosító** – a rekord egyedi azonosítója; **Lemma** – a címszó főváltozata; **Split** – annak jelölője, hogy található-e | vagy ~ a címszóban (ennek megléte morfológiai tagoltságra utal, hiánya esetén lehet a szó morfológiailag tagolt vagy tagolatlan); **Syll** – a címszó szótagokban mért hossza; **Phon** – a címszó fonémákban mért hossza (a hosszú mássalhangzók kettőnek, a hosszú magánhangzók egynek számítanak); **Freq** – a Magyar nemzeti szövegtár-beli gyakorisági érték; **Usg** – a szócikkfejben szereplő lexikai minősítések, szóközzel elválasztva; **Pos** – a szófaji minősítések, szóközzel elválasztva; **Sens** – a jelentések száma; **Etym** – az etimológia, tömörített formában (lényegében az átadó nyelv vagy nyelvcsalád neve, esetenként a szó keletkezési módja).

3. Egyszerű lekérdezések

Az így létrehozott relációs adatbázisból egyszerű lekérdezésekkel elő lehetett állítani a különféle etimológiájú, lexikai minősítésű, szófajú vagy jelentésszámú szóhalmazok szótári, valamint szöveggyakorisági mutatóit (tekintettel azokra a mezőkre, amelyek több lexikai, illetve szófaji minősítést is tartalmaznak, n minősítés esetén a k-adik minősítés $\frac{k}{1+2+\mathrm{K}+n}$ pontot kapott – ez $n=k=1$ esetén szerencsére éppen 1)[1]. Efféle szöveggyakorisági mutatók korábban megfelelő adatbázis-, illetve korpuszháttér híján nem voltak számíthatók; a szótári gyakoriságok Papp Ferenc korábbi forrásokon alapuló hasonló vizsgálataival való összevetésre adnak lehetőséget ([3], [4]).



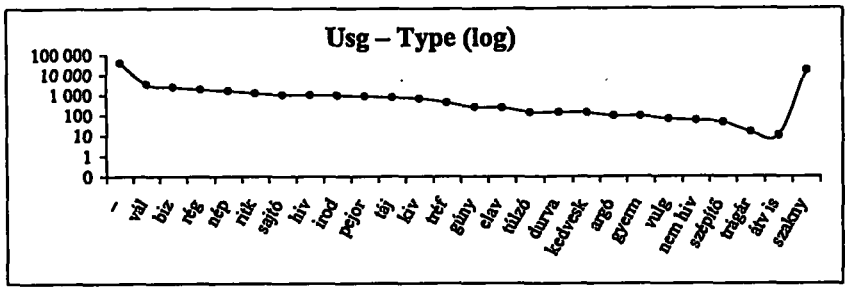
1. ábra. A különféle etimológiájú szavak szótári gyakorisága (logaritmikus skálán)[2]



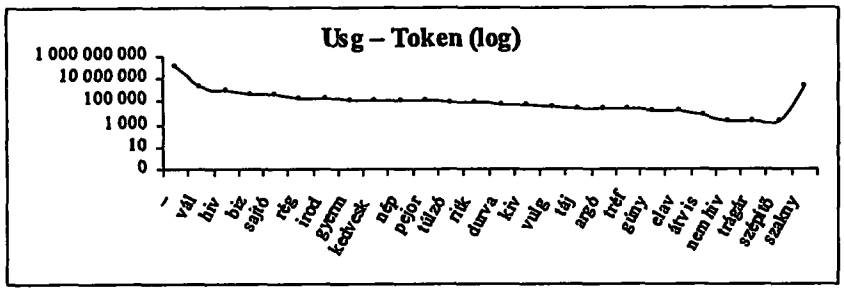
2. ábra. A különféle etimológiájú szavak szöveggyakorisága (logaritmikus skálán)

[1] A szöveggyakorisági értékek a számítás lényegéből adódóan pontatlanok, hiszen az egyes szövegszók szófajának vagy a lexikális minősítéseknek valódi eloszlásáról nincsen adatunk.

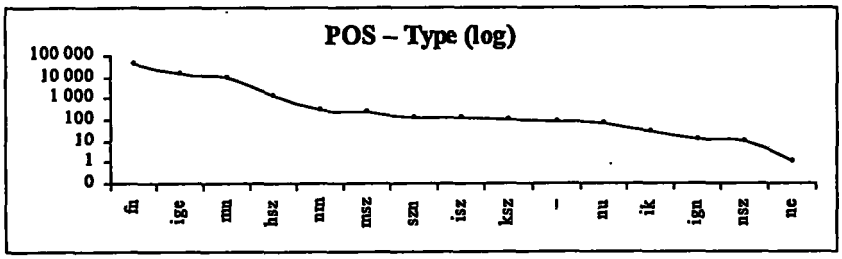
[2] Az etimológiai minősítés nélküli lemmák belső keletkezésűként számítottak.



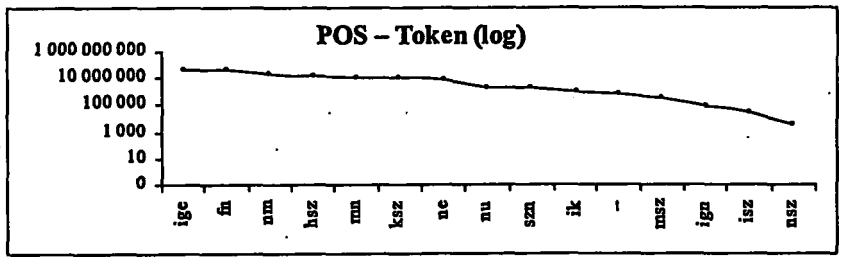
3. ábra. A különféle lexikai minősítésű szavak szótári gyakorisága (logaritmikus skálán)[3]



4. ábra. A különféle lexikai minősítésű szavak szöveggyakorisága (logaritmikus skálán)

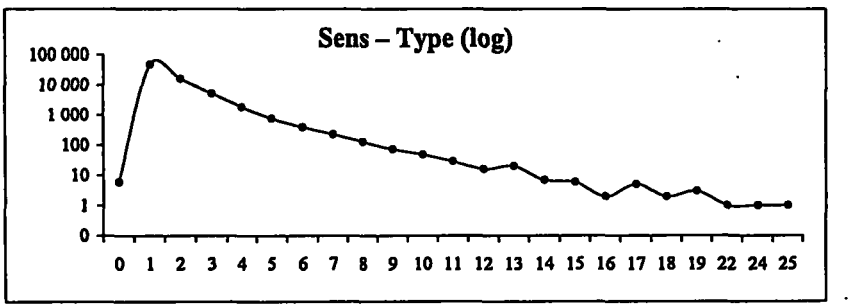


5. ábra. A különféle szófaji minősítésű szavak szótári gyakorisága (logaritmikus skálán)

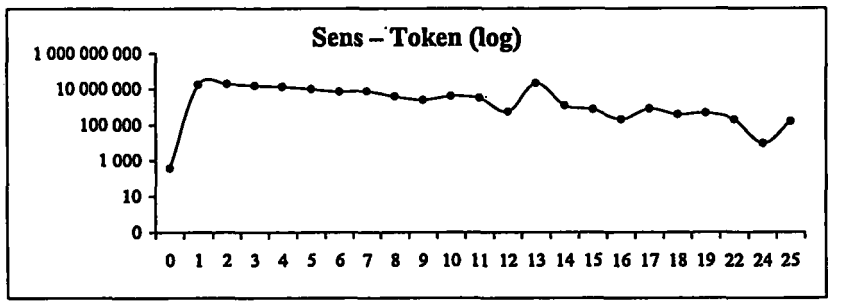


6. ábra. A különféle szófaji minősítésű szavak szöveggyakorisága (logaritmikus skálán)

[3] A szaknyelvi (tehát nagybetűs kezdetű) minősítéseket egységesen *szakny*-nak tekintettük.



7. ábra. A különféle jelentésszámú szavak szótári gyakorisága (logaritmikus skálán)



8. ábra. A különféle jelentésszámú szavak szöveggyakorisága (logaritmikus skálán)

4. Asszociációs szabályok

Az adatbányászat eszköztárával azonban – a fenti paraméterek közt fennálló rejtett mintázatok feltárását megcélzandó – ennél izgalmasabb elemzéseket is el lehetett végezni, asszociációs szabályok kinyerése segítségével. A rejtett mintázatok feltárásakor – több lexikai vagy szófaji minősítés esetén – csupán az első helyen állót vettük figyelembe.

Kiválogatva azokat a párosokat, amelyek tartója[4] legalább 10, összesen 7015 szabályjelölt állt elő, ezek közül a további vizsgálatok csupán az 1%-nál (tehát 724-nél) nagyobb tartójú 254 párt érintették, melyek közül csupán 103 volt legalább az egyik irányban legalább 67%-os valószínűségű szabály.

Azok az $A \rightarrow B$ asszociációs szabályok, amelyek esetében az A gyakorisága alacsony, ellenben a B gyakorisága magas, nem igazán beszédesek – habár ilyenből sok van. A legnagyobb gyakoriságú B-k a 'belső keletkezésű', illetve az 'egyjelentésű': a fenti 103 párból 71-et érintenek.

Bizonyos értelemben semmitmondó a fonémában, illetve szótagszámban mért szóhosszúság közti összefüggés, hiszen ez is triviális, beszédes is azonban, hiszen a magyar szótaghosszúságról állít valamit. E tárgyban a következő asszociációs szabályok adód-

[4] Tartó = az együttes előfordulás gyakorisága.

tak:[5] 5 fonéma ⇒ 2 szótag (8%; 92%); 3 fonéma ⇒ 1 szótag (1%; 87%); 7 fonéma ⇒ 3 szótag (11%; 81%); 10 fonéma ⇒ 4 szótag (9%; 79%); 8 fonéma ⇒ 3 szótag (12%; 77%); 4 fonéma ⇒ 2 szótag (2%; 73%); 13 fonéma ⇒ 5 szótag (2%; 68%); 12 fonéma ⇒ 5 szótag (3%; 67%). Egyiknek sem éri el a megfordítása az 50%-ot.

Habár a 'főnév' is igen nagy gyakoriságú, magas tartója és valószínűsége miatt érdemes megemlíteni a szakny ⇒ fn (20%; 87%) szabályt, illetve azokat a szabályokat, amelyek azt állítják, hogy a nagyon kis szöveggyakoriságú szavak jó eséllyel főnevek:[6] 0 db ⇒ fn (5%; 77%); 1 db ⇒ fn (3%; 72%); 2 db ⇒ fn (2%; 69%); 5 db ⇒ fn (1%; 68%); 6 db ⇒ fn (1%; 67%); 4 db ⇒ fn (1%; 67%). Megfigyelhető, hogy a hosszabb szavak szintén nagyobb valószínűséggel főnevek: 14 fonéma ⇒ fn (1%; 73%); 6 szótag ⇒ fn (3%; 71%); 13 fonéma ⇒ fn (2%; 71%); 12 fonéma ⇒ fn (4%; 70%); 5 szótag ⇒ fn (8%; 68%); 11 fonéma ⇒ fn (6%; 67%). A *rég* minősítésű szavak is igen gyakran főnevek: rég ⇒ fn (1%; 67%). Az etimológia szintén összefüggésben áll a főnévi szófajjal: szláv ⇒ fn (1%; 87%); nk+vszó ⇒ fn (4%; 81%); ném ⇒ fn (2%; 75%); lat ⇒ fn (2%; 70%) – ezen jelenség oka feltehetőleg az, hogy a főnév a legnyíltabb (tulajdonképpen az egyetlen teljesen nyílt) szóosztály, tehát a szókölcsönzések leginkább ezt érinti.

A fejben lexikálisan nem minősített lemmák[7] körében is megfigyelhetünk néhány szabályosságot. Természetes, hogy a többjelentésű címszavak szócikkfejében gyakran nincs lexikai minősítés: 5 jelentés ⇒ nincs lexikai minősítés (1%; 97%); 4 jelentés ⇒ n. lex. min. (2%; 93%); 3 jelentés ⇒ n. lex. min. (6%; 88%); 2 jelentés ⇒ n. lex. min. (17%; 77%). Összehasonlításképpen a monoszémák esetében: 1 jelentés ⇒ n. lex. min. (27%; 42%). Érdekes, szintén a lexikai minősítés hiányával kapcsolatos, nagy tartójú szabály: ige ⇒ n. lex. min. (16%; 71%) – melyet vélhetőleg az igék poliszemantikus hajlama indukál. Hasonló oka lehet az 1 szótag ⇒ n. lex. min. (2%; 67%) szabálynak.

5. Összefoglalás

A fentiekben áttekintettük, hogy melyek azok az információtípusok, amelyek rejtve maradnak egy szótárban, és bemutattunk néhány példát arra, hogy ezek miként nyerhetők ki.

Irodalom

1. Bárczi Géza: Magyar szófejtő szótár. Egyetemi Nyomda, Budapest (1941)
2. Bárczi Géza–Országh László (szerk.): A magyar nyelv értelmező szótára I–VII. Akadémiai Kiadó, Budapest (1959–1962)
3. Papp Ferenc (szerk.): A magyar nyelv szóvégmutató szótára. Akadémiai Kiadó, Budapest (1969)
4. Papp Ferenc: A debreceni thészaurusz. Linguistica. Series C. Relationes 11. MTA Nyelvtudományi Intézet, Budapest (2000)
5. Pusztai Ferenc (szerk.): Magyar értelmező kéziszótár. Akadémiai Kiadó, Budapest (2003[2])

[5] A zárójelen belül először a tartó, majd a szabály valószínűsége áll százalékban.

[6] A szabályok értelmezéséhez fontos tudni, hogy a főnevek adják a szócikkek 59%-át.

[7] Ezek a szócikkek 56%-át adják.