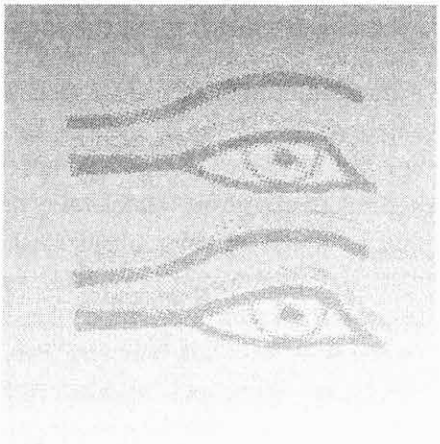


Gyakorisági szótárak

MAGYARORSZÁGI HELYZETKÉP

LENGYELNÉ MOLNÁR Tünde

A gyakorisági szótárak szerepe a tartalmi feltárásban

Napjainkban egyre nagyobb szerepet kap a tartalmi feltárás automatizálása. Egyik legaktuálisabb felhasználási területe az információkereső rendszerek hatékonyságának növelése, továbbá az internetes keresőoldalak keresőmotorjának fejlesztése.

Véleményem szerint azonban nem csak az interneten megjelent információk áttekintése jelent gondot, mert a kutató-, oktató akkor is nehéz helyzetben van, ha a saját tudományterületén a nyomtatásban megjelent publikációkat szeretné figyelemmel kísérni. Erre a problémára megoldást jelenthetnek a referátumok, ám ezek előállításának határt szab a készítő dokumentalisták véges kapacitása. E helyzet javítását szolgálná saját kutatásom, mely egy magyar nyelvű offline kivonatoló program készítésére irányul.

A kivonatkészítés automatizálásának lényegét a szöveg legfontosabb mondatainak meghatározásában látom, ennek előállítása a nyelvészet eredményeinek felhasználásával statisztikai alapon történik. Az eljárás kiindulópontját a szöveg szavai képezik, illetve a szótövekből felállítható szógyakorisági lista. Ezen a listán kell meghatározni azt a tartományt, mely a szöveget jellemző szignifikáns szavakat tartalmazza. Zipf-törvénye szerint a szignifikáns kifejezések köre a gyakorisági lista adott tartománya, ami szakterületenként változik, de minden egyes tudományágban igaz, hogy nem a lista eleje és nem is a vége.[1] A szignifikáns szavak listájának meghatározása során figyelembe kell venni az adott terület kifejezéseiből készített gyakorisági szótárakat. A lista elején található szavak képezik a tiltott szavak listáját (egy matematikai példatár esetén a feladat, és példa szavak nem lesznek szignifikánsak, hiába magas az előfordulási számuk), a „középmezőny" pedig megadja a területre jellemző szókincset.

Hazai készítésű gyakorisági szótárak

A magyarországi gyakorisági szótárak többnyire szépirodalmi szövegek alapján készültek egy-egy szerző műveiből, egy-egy kódex szavaiból képzett gyakorisági adatokkal-, vagy pedig a magyar nyelv egészére vonatkoznak.* E szótáraknak fontos történeti értékük van, továbbá segítenek az adott korszak szókincsét feltárni, értelmezni,** és sok más területen szolgálnak kutatások, fejlesztések alapjául. A felhasználók között nemcsak irodalomelméleti és nyelvészeti kutatókat találhatunk, hanem a pszichológiában, filozófiában, illetve különböző történeti kutatásokban érdekeltek is hasznosítani tudják. Meg kell említeni a gyermekek szókincsére irányuló vizsgálatokat, melyekből több esetben önálló gyakorisági szótárak is készültek – lásd később –, illetve alapul szolgálnak a gyermekek fejlettségi szintjének meghatározásakor. A gyakorisági szótárak egyik legjelentősebb alkalmazási területének a nyelvoktatás tekinthető, ahol az elsajátítandó szavak meghatározása – a köznyelv szókincsét összesítő – gyakorisági szótárak adatainak figyelembe vételével történik.

Az elmúlt években megnőtt a gyakorisági szótárak készítése iránti elkötelezettség és szerencsére egyre több szótárt építő vállalkozással találkozhatunk, azonban szakterülethez kapcsolódó átfogó gyakorisági szótár továbbra sem létezik magyar nyelven.

A következőkben áttekintem a Magyarországon meglévő gyakorisági szótárakat, részletesen bemutatva az adataikat és ezek elérési módjait, illetve a korpusz, vagyis az alapul szolgáló szöveganyag nagyságát.

Szógyűjtés a két világháború között

Az első gyakorisági szótárak megjelenését a gyorsírás használata tette szükségessé. A gyakoribb szavak gyorsírásban alkalmazott jelét egyszerűvé kellett tenni, míg a ritkább előfordulású szavak hosszabb jelölést is kaphattak. *Nemes Zoltán* a parlamenti gyorsírók munkáját könnyítette meg **A magyar parlamenti nyelv leggyakoribb szavai** című művével, mely 1933-ban jelent meg. Szótárában 40 977 szót dolgozott fel és a 25-nél többször előforduló szavakat vette alapul. A 200 parlamenti beszédrészletből előálló szótárban a szavak alfabetikus és gyakorisági sorrendben is megtalálhatók, sőt az összetett szavakból különálló táblázat is készült. Ha megnézzük a gyakorisági sorrendben közölt adatait láthatjuk, hogy a szerző előbb az alapszó ragozott formáinak számát adja meg *(3. oszlop)*, majd a továbbképzések nélküli előfordulások számát közli *(4. oszlop)*, végezetül a továbbképzett alakok és azok ragozott formáinak (ez az érték *zárójelben* található) előfordulását is megismerhetjük.

Sorszám	Alapszó	Az alapszó gyakorisága		Továbbképzések és megjegyzések
		Továbbképzéseivel együtt	Továbbképzések nélkül	
1	2	3	4	5
...				
20	van	268	237	-nak 31
	köz	261	120	-ött 45 (40); -ség 28 (21); -vet-16. –Még: közel 19; közép 13
	miniszter	253	219	-ium 26 (26)[2]

* Ezenkívül az iskolások-, újságok szókincséből készültek gyakorisági szótárak.

** A legtöbb szótár a gyakorisági adatok megadásán túl példamondatokkal segíti a szó értelmezését. (Lásd a későbbiekben ismertetésre kerülő szótárak esetében)

A gyakorisági szótárakra nem jellemző, hogy az összetételekben szereplő szavakból külön szótárat készítsenek, ezért talán érdemes ebből is megtekinteni egy részletet:

	Összesen	Ebből előfordul			
		önállóan	előtagban	középtagban	utótagban
...					
miniszter	219	121	38	–	60
ad	217	97	36	8	82
visel	209	9	–	102	98[3]

Nemes Zoltán munkájából azt is megtudhatjuk, hogy a leggyakoribb 111 szó az összes szó több mint 50%-át képzi. Munkája – mivel parlamenti szövegeken alapszik – szakszótárnak tekinthető, de követőre csak 50 évvel később talált.

Következő műve a beszélt nyelv szóstatisztikájának felállítását célozta meg. Az 1941-ben megjelenő **Szóstatisztika egymillió szótagot felölelő újságszövegek alapján** című könyve már nagyobb korpusszal dolgozott, pontosan* egymillió szótag alapján készült. Alapja a Budapesten akkor megjelent összes (26) napilap egy-egy száma, melyből a szerző havonta egyet, esetleg kettőt dolgozott fel két éven keresztül, így biztosítva a rövid életű hírek ismétlődésének kizárását.

A könyv fő része a 25-nél gyakrabban előforduló egyszerű szavak** gyakorisági sorrendben, illetve az 5-nél többször előforduló szavak alfabetikus sorrendben történő közlése. Emellett azonban számos egyéb táblázatban ismerhetjük meg a magyar és idegen szavakra vonatkozó statisztikákat szókezdő betűik, szótagszámuk, illetve előfordulásuk gyakorisága szerint csoportosítva. Példaként nézzük meg a 15 –18. leggyakoribb szót:

	Önállóan	Összetett szó			Összesen
		Elő-	Közép-	Utó-	
...					
Magyar	1746	461	15	71	2293
Nagy	1417	578	21	117	2133
volt (ige)	2011	-	-	4	2015[4]

A szerző 7 év alatt cédulázó módszerrel dolgozta fel a saját kezűleg gyűjtött anyagot, melynek csupán az átlapozása kereken 15 óráig tartott.[5]

Szótárak a közelmúltból

A gyorsírás mellett jelentős volt a gyakorisági szótárak nyelvtanulásban történő felhasználása, ahol a nyelvi tankönyvek összeállítói – optimális esetben – ilyen szótárakat használnak annak eldöntésére, hogy az adott nyelv esetenként több tízezres szókincséből melyik az az 1500 szó, amit tudni kell egy középfokú nyelvvizsgához, vagy amelynek a tanításával kell kezdeni egy első osztályos kisiskolás képzését. Magyarországon erre a célra a mai napig a Füredi–Kelemen szótárat veszik alapul.

* A szerző előbb elhagyta a minta egy millió szótagon felüli részét, majd végül pótolta a feldolgozás során lekopott szótagokat, hogy tartsa a kitűzött pontos értéket.

** A nem összetett és az összetett szavak alkotóira bontásával nyert szavak.

Füredi Mihály és Kelemen József nagyszabású munkája **A mai magyar nyelv szépprózai gyakorisági szótára** az 1965–1977 közötti időszak szépirodalmát dolgozta fel statisztikai mintavétel alapján. Munkájuk egy részét kézzel végezték, majd a feldolgozás számítógéppel zárult.

A szótár gyakoriságuk szerint csökkenő sorrendben közli a szóalakokat és lexémákat, melyek együttes megjelenítése megkönnyíti az egyes lexémák előfordulásainak áttekintését. A csökkenő sorrend alapját módosított gyakorisági adatok képezik, melyek az abszolút előfordulás mellett a szóródás értékét is figyelembe veszik. Így hiába van egy adott szóalaknak megfelelő számú előfordulása, ha az alapul szolgáló 258 író közül csak néhány használta, és az előfordulása nem egyenletesen oszlik el a művek között, alacsonyabb módosított gyakorisági értéket kap.

Nézzük meg hogyan épül a csökkenő gyakoriság szerinti lista a szótárban[6]. *(lásd táblázat)*

Ha figyelembe vesszük, hogy a munka egy része kézzel készült, lenyűgöző az adatok sokasága és jól használható formában történő megadása, melyet sok szempontú közreadással még alaposabbá tettek a szerzők. A műben kiegészítésként – módosítás nélküli – abszolút gyakorisági adatok alapján is közlik a szóalakokat, illetve a gyakorisági szótár részletes statisztikai adatokat is tartalmaz a szófajok, igealakok, magánhangzók, grafémák gyakoriságára vonatkozóan.

rangérték*	szuperlexémák, lexémák, szóalakok	homonímia kód**	szófaj kód***	műszám****	abszolút előfordulás	diszperzió százalék*****	módosított gyakoriság******
A	B	C	D	E	F	G	H
	…						
16	**tud**		**10**		**2462**	**96,26**	**2369,87**
	tud		10	258	2333	95,65	2231,46
	tudom				405	93,29	377,84
	tudja				339	91,59	310,47
	tudta				310	90,76	281,36
	tud				172	94	163,39
	tudod				173	93,31	161,42
	tudott				160	91,56	146,49

* Az A oszlopban lévő rangérték a leggyakoribb szavak sorszáma

** 1-4. szótári homonima, 5, kérdő-; 6. vonatkozó-; 7. határozatlan-; 8. általános-; 9. emfatikus névmás.

*** 1x ige, 2x főnév, 3x mn, 4x szn, 5x némás, 9x kötőszó. Részletesen: Füredi Mihály-Kelemen József: A mai magyar nyelv szépprózai gyakorisági szótára. Budapest: Akad. K. 1989. p. XXI.

**** Hány műben fordul elő a szóalak (max. 258 lehet)

***** Megmutatja mennyire egyenletes a szövegszó, vagy lexéma eloszlása (alapul a fél millió vizsgált szó szárezres részösszesítései szolgálnak)

****** Az abszolut gyakoriság diszperziós értéke

Papp Ferenc szövégmutató szótára és *A mai magyar szépprózа gyakorisági szótár* anyagából készült egy közös adatbázis is: a **Debreceni tezaurusz**[7], melynek online elérése (http: //romanid.freeweb.hu/DT2/) azonban a gyakorisági adatokat nem közli.

Nagyobb tudományterület szakkifejezéseit hosszú időn keresztül nem dolgozta fel senki más, viszont egy-egy író szóhasználatából készültek gyakorisági szótárak.

A legkiemelkedőbb a **Petőfi-szótár,**[8] mely Petőfi összes versére kiterjed és 22 719 önálló szócikket tartalmaz. Mivel a készítők célja, nemcsak a szavak érthetővé tétele, hanem Petőfi teljes nyelvezetének, egyéniségének érzékeltetése, ezért a szerzők egy írói szótárat készítettek. A mű 4 kötete az MTA Nyelvtudományi Intézetében jött létre, ahol egy munkacsoport több mint három évtized alatt[9] állította össze az anyagát hagyományos cédulázó anyaggyűjtéssel. A szótárban az egyes szavak statisztikája a szóalakok adatait is tartalmazza, megadva minden egyes előforduláshoz tartozó gyakorisági értéket, illetve az előfordulások helyét (a zárójelben található számokhoz egy forrásjegyzékben található a kódszámok feloldása).

füzet fn 3|-ét 1; -ébe 1; -re 1

1. 'összefűzött kisebb kéziratköteg':
   *A meglevő s eddig még ki nem nyomtatott költemények egy füzetre terjedő kézirata. (P 113/7 : 638) +7 : *508
2. 'kisebb lapszámú, fűzött nyomtatvány (kiadványsorozat egy száma)':
   a Tavasz második füzetébe majd összeszedem katonakoromból megmaradt alexand-rinus verseimet (P 72/5 :11)[10]

A negyedik kötet végén találjuk meg a szavak gyakorisági sorrendben lévő jegyzékét, mely a sorrenden, szóalakon és gyakorisági értékek megadásán kívül homonimák esetén a szófaj közlésével segíti az értelmezést (azonos szófaj esetén csak egy sorszám utal a több jelentésre).

...	40.	szép	1281
		vagy	1278
		ember	1260
		szív[2]	1224[11]

Juhász Gyula költői nyelvének szótára nemcsak a költő életművéből készült, hanem az összes (1373) verse is feldolgozásra került *Benkő László* 1972-ben megjelentetett írói szótárában.

A 12 000 szócikk ebben a műben is értelmezésre kerül, továbbá megtalálhatjuk a szerző kiegészítéseit (pl. alábbi idézetben: *ritk*án előforduló *rég*ies szó; *megszem*élyesítés, mely *hasonl*atban fordul elő).

ábráz ts ige (ritk)
(Képzőművészeti alkotás) ábrázol: A falu tornya. Különb szívemnek, mint a cifra torta, Mely habból ábráz téged, kölni dóm! 905/56 (vleg ironikus rég) megszem hasonl[12]

A szótárban nincs külön gyakorisági lista, hanem a gyakorisági értékek a szócikkek mellett vannak feltüntetve (az egynél többszöri előfordulás esetén).

Csokonai színműveiből készítette *Jakab László* és *Bölcskei András* a **Csokonai-szókincstár I.** kötetét,* melyet a szerzők nem tartanak írói szótárnak, mert a szócikkek jelentéseit nem fogalmazzák meg, hanem példákkal szemléltetik. Ezért kapta a mű a szókincstár elnevezést.[13] A majdnem 10 000 feldolgozott szócikknek közlik a szófaját, idegen kifejezés esetén megtudjuk mely nyelvből származik a szó, továbbá megtaláljuk a szócikk előfordulásaiknak számát, és versrészleteket a kifejezés illusztrációjaként.

A szókincstár először alfabetikus rendben közli az egyes szócikkek elemzését, zárójelben közölve az előfordulásának abszolút értékét, mint például:

* A Csokonai szótár és egyéb kódexek teljes anyaga elérhető a Számítógépes Nyelvtörténeti Adattárban: http://mnytud.arts.klte.hu/sznyta.htm

gyengeség fn (2)
Még eddig soha nagyobb gyengeséggel nem szerettelek tégedet (10/80). Ki kárhoztatná az Emberek, az Istenek kőzőtt az Égen, a' Földőn az én Gyengeségemet? (12/475)[14]

Majd a szókincstár végén a 10-nél többször előforduló szócikkek gyakoriságuk sorrendjükben is feltüntetésre kerülnek. Összesen 1042 szóból áll az itt található lista.

Sorszám	Szó	Szófaj	Előfordulás száma
⋮	.	.	.
38.	szép	mn/fn	251
39.	isten	fn	250
40.	mond	i	248
41.	lát	i	246[15]

A szótár nemcsak az író szókincsét összesítő, értelmező mű, hanem a 18. század végére jellemző beszédnyelv gyakorisági szótárának tekinthető.
A mű 1993-ben jelent meg nyomtatott kiadásban, és az eddig tárgyalt szótárak közül ez az első, melynek teljes anyaga online módon is letölthető a Magyar Elektronikus Könyvtár adatbázisából (http://www.mek.iif.hu/porta/szint/tarsad/irodtud/csokonsz/).
Ugyancsak Jakab László és Bölcskei András nevéhez fűződik a **Balassi-szótár**, mely Balassi verseit és Szép magyar komédiájának a szókincsét tartalmazza, 4735 szócikket képezve belőlük. Ez már írói értelmező szótár, melyből nemcsak a szavak jelentéseit ismerhetjük meg, hanem a Balassi művekben található szólások magyarázatát is megadják a készítők.
A műben a következőképpen épül fel egy szócikk (mely 8-szor fordult elő Balassi műveiben):
„felség fn (8)
1. Uralkodó. *Mert ő az, akinek hatalmában az Ég, Neki enged tenger, Menny s földi kerekség, Segélli övéit, mint mennyei* ~ (99/39).
2. Hatalom, magas tisztség. *Nabugodonozort hét esztendőre Vivéd ismét királyi ~ére* (92/38).
3. <Isten megszólítása.> *Én Istenem, ebben ne vesszen el vérem, Őrizz meg gonosztól, ~édet kérem!* (65/12)
4. Fenség, magasztosság. *Az Szentháromságnak első személye, Atyaisten dicsőséges ~e, Mind ez széles világnak teremtője, Tekints reám, ilyen veszett szegénre!* (92/2)
~ (92/42), ~edet (1002/68), ~ednek (102/27), ~re (93/8).[16]

A forrás megjelölés a feldolgozott 138 vers egyikére, illetve annak sorára utal, továbbá K betűvel jelzi, ha a Komédia adatairól van szó, illetve még N betűvel is találkozhatunk, ha a Komédia nyomtatott verziója szolgált forrásul.
A leggyakoribb 938 szócikk gyakorisági statisztikáját a szótár elején találjuk meg, azonban sajátos módon nem közlik a szócikkeket, csak a sorszámot, az abszolút gyakoriságot, és a halmozott gyakoriság értékét:

Sorszám	Abszolút gyakoriság	Halmozott gyakoriság
1.	1241	1241
2.	676	1917
...		
578-634.	9	29564
635-719	8	30244[17]

Ha mi szeretnénk megtudni, hogy az előbb idézett felség szó hol helyezkedik el a gyakorisági listán, akkor vissza kell fejtenünk, hogy a 8 abszolút gyakorisággal rendelkező szavak az 578-634. helyen állnak.
A teljes anyag 2000-ben készült el Debrecenben, és ezen szótár online elérését is lehetővé tették (http://www.mek.iif.hu/porta/szint/tarsad/irodtud/balassi/).
2004-ben jelent meg *Beke József* **Zrínyi-szótára**, a következő olyan írói szótár magyar nyelven, mely teljességre törekszik a szerző műveinek figyelembe vétele során. Zrínyi Miklósnak nemcsak a verseit, és prózai műveit dolgozza fel, hanem a levelei és a könyveiben található glosszák is be-

letartoznak a gyakorisági adatok korpuszába, melynek eredményeként 6882 címszót állítottak elő, és annak 127 504 előfordulását dolgozták fel.
A gyakorisági táblázat a mű elején található, melyben a gyakorisági adaton túl a szótárbeli első szófaj is feltüntetésre kerül.

„ ... 26. tud 580 i
27. török 540 fn
28. vitéz 529 mn"[18]

A címszavak szótárában különlegesség, hogy megadják az adott szó megtalálható-e az Értelmező Kéziszótárban. Ha nem, ezt a szó előtti *-al jelölik. A forrásmegjelölésre rövidítéseket használnak (pl.: *Le 170:12* A 319 Zrínyi levél közül a 170-ediknek 12. sorából való az idézet), melyek feloldása természetesen megtalálható a szótárban.

„*királyképe 3 fn (király képe 1)
‚a király képviselője, helyettese, királyi biztos' [egy levelet] győri királyképe uramnak írtam, azt is méltóztassék Kegyelmed megküldeni űkegyelmének Le 170:12; .. elrontá szép reménségünket egynihány koszos német, győr[i] király képétül voltak elküldve Le 232:25; Bi 187 (vö. TESz. II. 448; NySz. II. 203)"[19]

Egy kis érdekesség – mely ezen idézet tükrében már nem is tűnik olyan meglepőnek –, hogy a gyakorisági adattáblában a 4. helyen a kegyelmed szó áll, melyet csak az ‚a', ‚és', ‚nem' szavak előznek meg.
Egyelőre a Zrínyi-szótár zárja a teljes életműre kiterjedő magyar írói szótárak sorát, viszont írók legjelentősebb művéből (műveiből) készült gyakorisági szótárat többet is ismerünk.

király fn **41** (vö. **kiskirály)**

Az elsőt *Szily Kálmán* készítette Mikes Kelemen *Törökországi leveleiből* 1906-ban.*
Arany János Toldi c. elbeszélő költeményét *Pásztor Emil* dolgozta fel, és munkájának eredményét, a **Toldi-szótár**-at 1986-ban jelentette meg. A szótár a Toldi trilógiából a legismertebb Toldi című első rész szókincsét tartalmazza, melyből 3059 szócikk állt elő. Némelyik szócikket grafikával is szemléltetnek.

1. 'monarchikus állam férfi uralkodója' (minj nélkül): Felmegyek Budára bajnok katonának, Mutatok valamit ottan a királynak, Olyat, a mi nem lesz bátyám szégyenére (6:15); (8:2, 3);..."[20]

A függelékben található gyakorisági szótár a 10-nél többször előforduló szavakat tartalmazza:

„ ... 7. nagy 85
8. de 75
9. Miklós 71"[21]

Katona József Bánk bán című drámája a keletkezésekor is régies nyelvezetű volt, melynek értelmezése mára még nehezebb feladat, hiszen sok benne található szó nem része a mai szókincsnek, vagy jelentése módosult. Beke József **Bánk bán-szótára** átsegíti az olvasót a nyelvi értelmezés nehézségein, így joggal ajánlják a mű bevezetőjében színészeknek, tanároknak, diákoknak segédeszközként. A szótár felépítésében a Balassi Bálint-szótárhoz hasonlít a legjobban, itt is megtaláljuk az Értelmező Kéziszótárral való párhuzamot, és a csillag jelet azon szavaknál melyek nem találhatók meg a Kéziszótárban. Beke József még bevezetett 3 új jelölést: a népies (nép), a tájjellegű (táj) és a ma már szokatlan szavak (ritk) jelzésére. Az egyes idézetek esetén nemcsak a szöveget olvashatjuk el, hanem megtudjuk azt is, hogy ki mondta (esetleg kinek) az adott idézetrészletet, sőt Beke József feloldotta az idézetekben szereplő névmásokat, utalásokat is és [] jelben megtalálhatjuk, hogy melyik szereplőre vonatkozik a szöveg.

„zsivány 2 fn

* készült „II. Rákóczi Ferencék hamvainak hazahozatala alkalmából... (Az emlékkiadás utószavaként jelent meg, de ugyanabban az esztendőben a Magyar Nyelv című szakfolyóiratban is kiadták.)"
Kerényi Ferenc: Az írói szótárak hasznos voltáról http://forras.rkk.hu/0505/kerenyi.html

‚törvényen kívül élő rabló; szegénylegény': ez'/Asszonynak [=G-nak] a ' hatalma bűntetlen / teszi azt, mint a ' közönséges zsivány/ talán fizetne életével is?! **B** 2:262; Vigyázz, hogy egyj / zsivány, tömött erszényeiddel együtt, / ne lopjael nagylelkűségedet **T** B-nak 3:265"[22]
(Az utóbbi idézetet Tiborc mondja Bánknak a 3. jelenet 265. sorában).

A Bánk bán-szótárban 2882 címszót találhatunk, melyek közül a 20-nál többször előfordulókról készített a szerző gyakorisági listát.

„ …	10.	de	122
	11.	Melinda	111
	12.	oh, óh	10"[23]

Jakab László és Bölcskei András nemcsak Csokonai műveiből készített szókincstárat, hanem 2003-ban **Egy XVI. századi emlékirat szókincstára** címmel jelentették meg újabb gyakorisági szótárukat. A két szótár felépítése hasonló, itt is megtaláljuk a címszavak szófajait, gyakorisági adatait, illetve a szó eredetét, majd példamondatokkal történő magyarázatot. Az emlékirat több mint 15 000 kifejezéséből 2172 címszó készült. Az előző Jakab-Bölcskei művekhez hasonlóan a gyakorisági lista a szótár végén, a gyakorisági adatok pedig az egyes szócikkek mellett is megtalálhatók (zárójelben):
„gyülekezet fn (4)
Merth Ew nekyek annal Egyeb Sem Kell woth, az gywlekewzethbe Bp gywlnek Ky Ky mynd fegyweres Kezzel (27/10).
gywlekewzethyk (35/28), gywlekewzetewth (25/26, 35/5)."[24]

A szótár teljes anyaga letölthető.
(http://mek.oszk.hu/01200/01220/)

Nem csak a szépirodalom terén létezik gyakorisági szótár. Meg kell említeni, hogy van **Orosz –_magyar kémiai gyakorisági szótár**, amely *Egyed László* szerkesztésében 1984–ben jelent meg, azonban ennek felépítése jelentős mértékben eltér az eddig ismertetett szótáraktól. A szerző célja az orosz szaknyelv elsajátításának segítése volt, így az egyetemi oktatásban használt kémiai szöveg alapján kiválasztotta a területre legjellemzőbb 2500 szót, és a szótár ezen leggyakoribb orosz kifejezések jelentését adja meg, de nem tartalmaz gyakorisági adatokat.
„pl. АКТИВАТОР – aktiválószer; -anyag, aktivátor, gyorsító"

Nem képeznek különálló tudomány területet, de még speciálisnak tekinthetők a gyermek beszéltvagy írott nyelvét feldolgozó szótárak. Többet könyv formájában is megjelentettek, illetve sok kutatás épül ezen szótárak anyagára.
Cs. Czachesz Erzsébet és *Csirik János* a **10–16 éves tanulók írásbeli szókincsének gyakorisági szótára** címmel írt, 2002-s munkája a tanulók 2170 írásbeli fogalmazását elemezte, így közel 600 000 szó alkotta szótáruk forrását. A szótárat sokoldalú közreadás jellemzi. A szavak alfabetikus (bár folyamatos írásmódú) közlését a leggyakoribb 1000 szó feltüntetése követi, majd a leggyakoribb 600 szót szófajonként is megtalálhatjuk. Jól használható a könyv, ha a kutatásunk valamely korosztályra irányul, ugyanis évfolyamonként is feltüntették a szerzők a leggyakoribb szavakat, sőt itt is van szófajonkénti bontás. Minden esetben a szócikk, a szófaj kódja és az abszolút gyakoriság értéke kerül feltüntetésre (sorszám nélkül). Sorszámot a mű végén található gyakorisági eloszlások megadásánál találunk csak, ahol viszont a szócikk nem kerül feltüntetésre, helyette kumulált gyakorisági értékkel egészítették ki az adatsort.
Nézzünk meg egy részletet a leggyakoribb 1000 szó listájából:

„ …	család	N	1340	
	dolog	N	1329	
	érdekes		A	1274"[25]

Majd a gyakorisági eloszlások nézegetésével tudhatjuk meg, melyik helyre is kerültek az előző szavak a gyakorisági listán:

„ ... 64 1340 299588
65 1329 300917
66 1274 302191"[26]

A szótár teljes anyaga elérhető interneten, a Szeged Korpusz adatbázisába bedolgozva (http://www.inf.u-szeged.hu/projectdirs/hlt/).

A siket gyermekek beszédkészségének javítására készült Beszédmester szoftver előállítása közben jött létre az első osztályosok olvasástanításában leggyakrabban előforduló 2000 szó szótára. *Bácsi János* és *Kerekes Judit* összefoglalója 2003-ban jelent meg a kapott eredményekről: **Az első osztályos olvasókönyvek szóanyagából készült gyakorisági szótár: „Van szó"** címmel, melyből megtudhatjuk, hogy 15 iskola első osztályosainak szókincsét vették alapul. 12 226 szóalak gyakorisági értékének vizsgálata után határozták meg azt a 10-nél többször előforduló 1953 szót, mely a szoftverük adatbázisát képzi (a leggyakoribb ige a „van", a leggyakoribb főnév pedig a „szó"). Mivel a munka célja nem egy gyakorisági szótár összeállítása volt, ezért csak a leggyakoribb igék, állat- és tulajdonnevek, növények listája került publikálásra.

Igék		
1.	van	(1/2952/2,55204)
2.	olvas	(3/896/0,77460)
3.	mond	(5/827/0,69939) [27]

Az értékek feloldása nem történik meg, csak következtetni lehet, hogy az első szám az összesített (csoportosítás nélküli) gyakorisági listában elfoglalt helyett mutatja, majd az abszolút- és végül a relatív gyakorisági értékek kerültek feltüntetésre.

Az újságok szövegének vizsgálatából nem csak Nemes Zoltán készített gyakorisági szótárat, 1986-ban került kiadásra *Cs. Czachesz* Erzsébet és *Csirik* János oktatási célból készített **Újságnyelvi gyakorisági szótára**. A feldolgozásuk alapját 14 különböző témakörökben kiadott (Pl. Népszabadság, Füles, Ludas Matyi stb.), de legalább 200 000 feletti példányszámmal megjelenő újság 1-1, de teljes anyagú példánya képezte. Szótárukban az 5-nél többször előforduló szavakat közlik alfabetikus, majd gyakoriság szerinti sorrendben, azonban mindez folyamatos szövegként teszik, megadva a szócikket, szófajkódot, és az abszolút gyakoriság értékét.

„ ...méter F 159 így K 156 cél F 153 család F 153 magas M 153 éppen H 152..."[28]

Ha kíváncsiak vagyunk arra, hogy a *'család'* szó milyen helyezést kap a gyakorisági listán, a külön megadott gyakorisági eloszlás táblázatát felkeresve megkapjuk, hogy 119-edik leggyakoribb szava az újságoknak.

„ ... 118 156 77962
119 153 78421
122 152 79029"[29]

Akár az írói szótárak, akár a gyermeknyelvi szótárak kiadási éveit vizsgáljuk is meg, láthatjuk, hogy egyre nagyobb aktivitás figyelhető meg a szótár építők körében. A gyakorisági adatok iránti keresletet mutatja az is, hogy a **Magyar értelmező kéziszótár**at a 2003-s kiadástól kezdődően kiegészítették gyakorisági mutatókkal. A szóhasználat gyakorisági értékeinek meghatározásához a 150 millió szót tartalmazó Magyar Nemzeti Szövegtárat vették alapul. Az Magyar értelmező kéziszótár az eddigi gyakorisági szótáraktól eltérően nem egy abszolút gyakorisági értéket közöl, hanem egy ötös skálán rangsorolja a szavakat.

1. rangsorba kerül az első kétezer leggyakoribb szó
2. rangsort kap 2001–10 000 közötti leggyakoribb szó
3. helyre 10 001–30 000 leggyakoribb szó
4. helyre 30 001–60 000 leggyakoribb szó került

A rangsor 5. helyére kerültek azok a szavak, melyek nem tartoznak a 60 000 leggyakoribb szavunk közé.

A szótárban a következő formában találkozhatunk ezen adatokkal:
„gyermektelen ○ mn Akinek nincs gyermeke"[30]

A szerzők a gyakorisági adatbázis részletes adatait egy különálló gyakorisági szótár formájában szeretnék kiadni.[31]

Gyakorisági szótárak a 21. század elején

Napjaink egyik legnagyobb szótárépítő munkája a Budapesti Műszaki és Gazdaságtudományi Egyetem Média Oktató és Kutató Központjának **Szószablya** projektje, mely 2003 májusában zárult. A munka magában foglal egy mindenki számára ingyen elérhető helyesírás-ellenőrző, szótövező és morfológiai elemző programot, illetve a Szószablya Gyakorisági Szótár létrehozását, mely köznyelvi írott szöveget dolgozott fel. A gyakorisági szótár alapjául szolgáló webkorpusz forrását az origo+vizsla 18 millió mentett weboldala szolgáltatta.
A 18 millió weboldal közül 3,5 millió weboldal tartalmazott szöveget, mely 1486 millió szövegszavas forrást jelent. A gyakorisági értékek vizsgálatához szükséges a készítők weboldal szűrési módszerének áttekintése:
A 3,5 millió weboldal közül kiszűrték a nem magyar dokumentumokat, mégpedig azon oldalakat meghagyva, ahol az ismeretlen szavak száma 40% alatti (így 3,125 millió oldal maradt).
Következő lépésben az ismeretlen szavak száma nem haladhatja meg a 8%-t, melynek hatására kiestek a nem ékezetes betűket tartalmazó oldalak.
Majd képeztek egy még szűkebb korpuszt, melyben a hibás szavak száma nem haladhatja meg a 4 %-t, és ennél a „határnál már csak azok a lapok maradnak meg, amik kevesebb hibát tartalmaznak, mint egy átlagos nyomtatásban megjelenő szöveg".[32] Ez a küszöb felel meg a köznyelvi szókincsnek.

A szótár mérete

Korpusz	Oldalak (millió)	szövegszó (millió)	szóalak (millió)
Teljes	3,5	1486	19,1
40%	3,125	1310	15,4
8%	1,918	928	10,9
4%	1,221	589	7,2[33]

A szótár anyaga letölthető a http://mokk.bme.hu/eszkozok/webkorpusz/index_html/view címről, egy regisztrálás után. A készítők a 4%-s küszöbnek megfelelő majdnem 600 millió szó-cikkből álló korpuszból készült gyakorisági szótár letöltését teszik lehetővé. Nézzünk meg a leggyakoribb 20 szó adatait:

Sorrend	Szócikk	Teljes korpuszon	40% korpuszon	8% korpuszon	4% korpuszon
1	A	91905116	89013997	65868724	43050309
2	Az	32836302	32483167	24383179	16189184
3	És	26216677	26038935	19288013	12398358
4	A*	18411067	18160755	13490177	8867291
5	Hogy	16291361	16118673	11902102	7654795
6	Is	15782647	14271378	10097022	6285293
7	Nem	14228895	14018094	9918516	6151477
8	Egy	9574168	9446422	6440405	3746912
9	Az*	6550604	6488579	4802398	3179607
10	The	6421561	326080	45627	9377

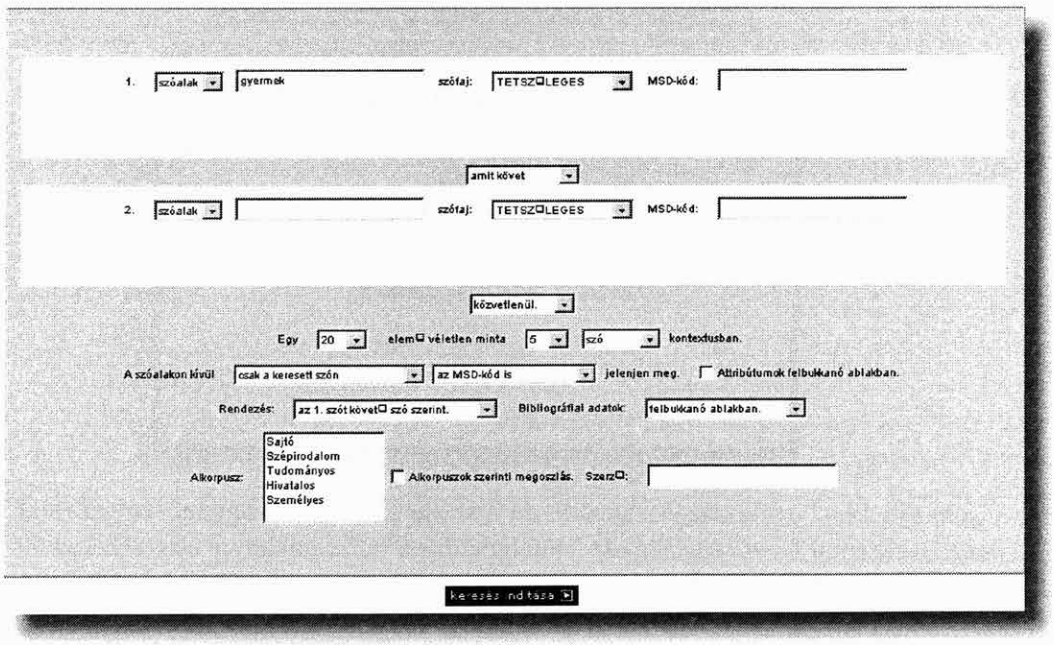
(A táblázat a köv. oldalon folytatódik.)

11	Meg	6347261	6239032	4316594	2739125
12	De	5265870	4691559	3103363	1799351
13	Csak	4589178	4523654	3156043	1919291
14	Vagy	4518453	4447767	3329093	2319038
15	Van	4491634	4388915	2857807	1711747
16	Volt	3915837	3873285	2776863	1653452
17	Of	3899809	290642	58256	15077
18	Már	3848697	3833176	2728287	1625065
19	#	3808091	3397264	2733003	2229385

A szavakat megvizsgálva látható, hogy az angol nyelvű oldalak kiszűrése a teljes korpuszon még nem valósult meg, azonban a 4%-s korpusznál található értéket vizsgálva már nem jelennek meg.
Az elkészült munka a korpusz nagyságával, és az adatok nyílttá tételével egyedülálló.

Befejezés

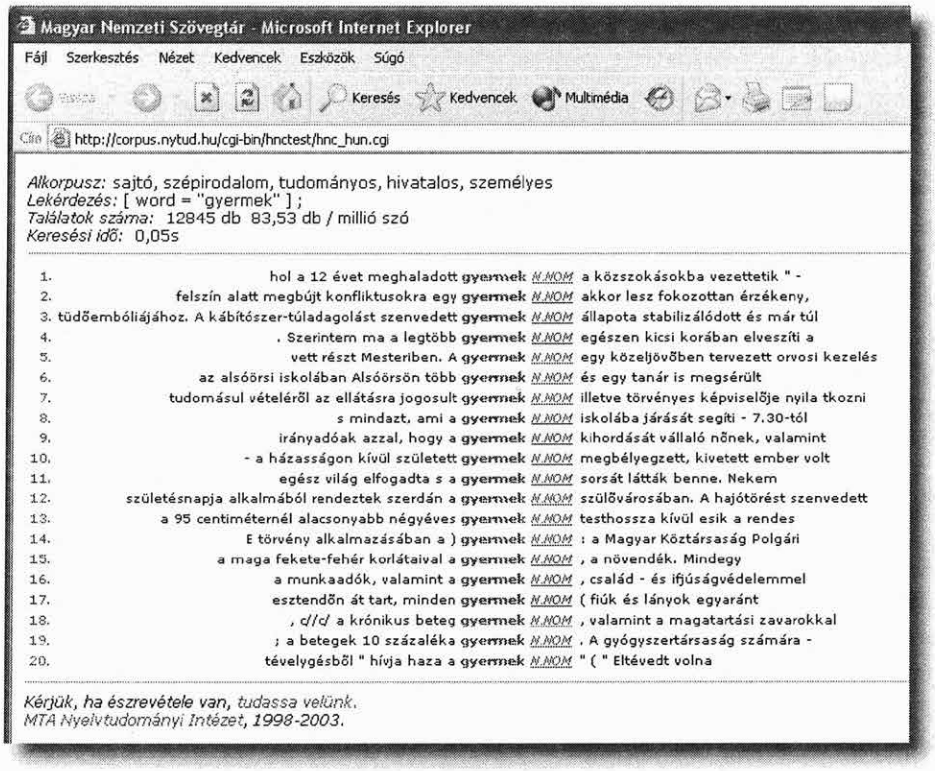
Mint ahogy említettem és az eddig ismertetett szótárak is mutatják, hogy a szépirodalomtól eltérő tudományághoz kapcsolódó átfogó gyakorisági szótár még nem készült magyar nyelven. Ezért mindenképp szeretném megemlíteni a **Magyar Nemzeti Szövegtárat**,* melyet 2002-re készített el a Magyar Tudományos Akadémia Nyelvtudományi Intézete. 150 millió szavas korpuszával magyar nyelven példa nélküli. A szövegtár a magyar írásos nyelvhasználatot reprezentálja, a forrásként szolgáló szöveganyaga öt nagy területről származik: a sajtó online kiadásaiból 75 millió szót vettek alapul, a szépirodalmat a Digitális Irodalmi Akadémia anyagai képviselik, és 15 millió szó került a korpuszba erről a területről, 20–20 millió szót választottak a tudományos prózát képviselő Magyar Elektronikus Könyvtár adatbázisából, a hivatali nyelvet a minisztériumok, önkormányzatok anyagai képviselik, a személyes közlést pedig az Index.hu Törzsasztal anyagából választották.[34]



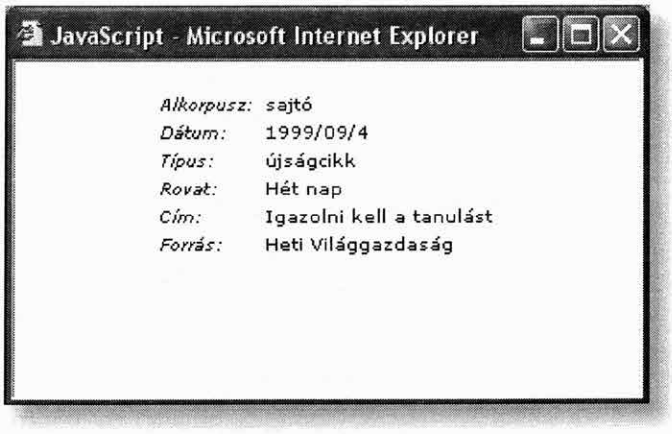
* http://corpus.nytud.hu/mnsz/

A szövegtárat bárki használhatja, aki regisztrálja magát weboldalukon, ahol a keresőkifejezés szófaját, megjelenési helyét is korlátozhatjuk.

A keresés eredményeként a kifejezés összes előfordulásáról kapunk egy-egy mondat részletet:



A rendszer ezen felül közli a forrás adatait is:



Igaz, nem tartalmaz gyakorisági adatokat, de a szótár készítések alapjául szolgál (már a Magyar értelmező kéziszótár is ezt vette alapul). Mint láthattuk a szövegtár az egyes szócikkekhez eltárolja a forrást is, így véleményem szerint nemcsak átfogó gyakorisági szótár alapja lehet, hanem az 5 terület mindegyikéhez rendelkezik elegendő korpusszal, így forrása lehetne egy szépirodalmi-, hivatali-, és sajtó terültére vonatkozó gyakorisági szótárnak is a köznyelvin kívül.

A kutatások számának növekedése mindenképp biztató. A jövőben véleményem szerint, egyre több gyakorisági szótárral fogunk találkozni, melyek közül néhánynak az alapját talán a Magyar Nemzeti Szövegtár szolgáltatja. Ahogy növekszik a tartalmi feltárásra való igény, illetve az itt elért technikai eredmények, úgy fognak nőni a korpuszok méretei, és az ezeket felhasználó projektek.

Irodalom

1. HORVÁTH Tibor: A könyvtártudomány és információtudomány alapjai. In: Könyvtárosok kézikönyve. 1. Alapvetés. Szerk. Horváth Tibor, Papp István. Budapest, Osiris, 1999. p. 56

2. NEMES Zoltán: A magyar parlamenti nyelv leggyakoribb szavai. Szeged, Szeged Városi Nyomda és Kiadó Rt., 1933. p. 28.

3. Lásd előző p. 42.

4. Lásd előző p. 55.

5. NEMES Zoltán: Szóstatisztika egymillió szótagot felölelő újságszövegek alapján. In: Az Egységes Magyar Gyorsírás Könyvtára 190. Szeged, 1941. p. 8.

6. FÜREDI Mihály – KELEMEN József: A mai magyar nyelv szépprózai gyakorisági szótára. Budapest, Akadémiai Kiadó, 1989. p. 5.

7. Szóvégmutató szótár. Szerk. Papp Ferenc. Budapest, Akadémiai Kiadó, 1968.

8. Petőfi Sándor életművének szókészlete 1–4. Szerk. J. Soltész Katalin, Szabó Dénes, Wacha Imre; Gáldi László irányításával. Budapest, Akadémiai Kiadó, 1973–1987.

9. WACHA Imre naplója alapján http://www.inaplo.hu/na/2000_10/070.htm

10. Petőfi-szótár. Első kötet A–F. Szerk. J. Soltész Katalin, Szabó Dénes, Wacha Imre. Budapest, Akadémiai Kiadó, 1973. p. 1160.

11. Petőfi-szótár. Negyedik kötet Sz–Zs. Szerk. J. Soltész Katalin, Szabó Dénes, Wacha Imre. Budapest, Akadémiai Kiadó, 1987. p. 785.

12. Juhász Gyula költői nyelvének szótára. Szerk. Benkő László. Budapest, Akadémiai Kiadó, 1972. p. 37.

13. JAKAB László – BÖLCSKEI András: Csokonai-szókincs-tár 1. Csokonai színművei szókincsének szövegszótára és adattára. Debrecen, 1993. http://mnytud.arts.klte.hu/sorozat/csoksz/csokbev.htm

14. Lásd előző

15. Lásd előző

16. JAKAB László – BÖLCSKEI András: Balassi-szótár. Debrecen, KLTE, Piremon, 2000. p. 137.

17. Lásd előző p. 18.

18. Zrínyi-szótár. Szerk. Beke József. Budapest, Argumentum, 2004. p. 35.

19. Lásd előző p. 469.

20. PÁSZTOR Emil: Toldi-szótár. Budapest, Tankönyvkiadó, 1986. p. 135.

21. Lásd előző p. 264.

22. BEKE József: Bánk bán-szótár. Katona József Bánk bán c. drámájának szókészlete. Kecskemét, Kecskeméti Lapok, 1991. ([Kiskunfélegyháza], Z-P Formular Kft.) p. 338.

23. Lásd előző p. 339.

24. http://mek.oszk.hu/01200/01220/

25. CS. CZACHESZ Erzsébet – CSIRIK János: 10–16 éves tanulók írásbeli szókincsének gyakorisági szótára. [Szeged], BIP, 2002. p. 104.

26. Lásd előző p. 254.

27. BÁCSI János – KEREKES Judit: Az első osztályos olvasókönyvek szóanyagából készült gyakorisági szótár: „Van szó". In.: Módszertani közlemények, 43. évf. 2. sz. p. 53.

28. CSIRIKNÉ CHACHESZ Erzsébet – CSIRIK János: Újságnyelvi gyakorisági szótár. 1–2. Szeged –Budapest – Debrecen, 1986. p. 235.

29. Lásd előző p. XLIII.

30. Magyar értelmező kéziszótár. Budapest, Akadémiai Kiadó, 2003. p. 462.

31. Lásd előző p. XV.

32. http://mokk.bme.hu/eszkozok/webkorpusz/index_html/view

33. http://mokk.bme.hu/eszkozok/webkorpusz/index_html/view

34. Magyar értelmező kéziszótár. Budapest, Akadémiai Kiadó., 2003. p. XIV.

További irodalom

FÜREDI Mihály: Metainformációk előállítása: a kivonatolás szempontjai. In: Tudományos és Műszaki Tájékoztatás, 51. évf. 2004. 12. sz.
http://tmt.omikk.bme.hu/show_news.html?id=3781&issue_id=457

JAKAB László – BÖLCSKEI András: Egy XVI. századi emlékirat szókincstára. http://mek.oszk.hu/01200/01220/

JAKAB László – KISS Antal: Az Apor-kódex ábécérendes adattára http://www.mek.iif.hu/porta/szint/egyeb/szotar/apor/apor.mek

JAKAB László – KISS Antal: A Festetics-kódex ábécérendes adattára http://www.mek.iif.hu/porta/szint/egyeb/szotar/apor/apor.mek

JAKAB László — KISS Antal: A Guary-kódex ábécérendes adattára http://mek.oszk.hu/01200/01293/

Magyar szókincstár. -© Tinta, 1998 © MorphoLogic, 1999 http://www.altavizsla.hu/kereses/d.pl?url=http:%2F%2Ftext.goliat.hu%2Fcgi-bin%2Fkereses.cgi%3FSubmit4.x=28%26Submit4.y=11%26KERESES=gyakorisági+szótár

A Magyar Tudományos Akadémia Nyelvtudományi Intézetének tudományos beszámolója a 2002. évről
http://www.nytud.hu/adm/jelentcel.html

A Miskolci Egyetem Központi Könyvtárának tájékoztatója. In: Online Híradó, 1995. Január.
http://217.20.130.18/cgi-bin/cache.cgi?ID=11240514&URL=http://www.lib.uni-miskolc.hu/kvtrol/publ/oh/51/online51.htm&KERESES=gyakorisági,szótár

NÉMETH László: Szószablya gyakorisági szótár
http://www.szoszablya.hu/pub/szotar/web2.0/LEIRAS.txt

Számítógépes Nyelvtörténeti Adattár
http://mnytud.arts.klte.hu/sznyta.htm

UNGVÁRY Rudolf: A tartalom szerinti információkeresés az Interneten 1. Indexelőszolgáltatások. In: Tudományos és Műszaki Tájékoztatás, 47. évf. 2000. 1. sz.
http://tmt.omikk.bme.hu/show_news.html?id=1624&issue_id=15

ZSOBRÁK Róbert: Isz siveöl tégaz hazuj avagy Mit tud a nyelvstatisztika?
http://www.sulinet.hu/tart/cikk/ag/0/21182/1

Az Akadémiai Könyvtár alapításának 180. évfordulója alkalmából tanulmány jelent meg a Magyar Tudomány 2006. 3. számában. A Körmendy Kinga és Mázi Béla által írt dolgozat (A Telekiek alapítványa) eredeti dokumentumok alapján mutatja be a Tudós Társaság kezdeti napjait és a könyvtár létrehozásának körülményeit. Webes elérhetősége: http://www.matud.iif.hu/06mar/13.html
(Bánhegyi Zsolt, Katalist, 2006. márc. 20.)