# A szövegkorpuszok szókincsének összehasonlítása szótári címszójegyzék felhasználásával – neologizmusok és archaizmusok detektálása

Kiss Gábor[1], Kiss Márton[1]

[1] TINTA Könyvkiadó, Budapest
{kissgabo, kissmarci}@tintakiado.hu

**Kivonat:** A Magyar Történeti Korpusz (TK) és a Magyar Nemzeti Szövegtár (MNSz) összehasonlítása egy más irányú lexikográfiai feladat elvégzésének melléktermékeként jött létre a TINTA Könyvkiadóban. Az elsődleges feladat az Értelmező szótár+ (ÉrtSz+ [1]) címszavainak gyakorisági mutatóval való ellátása volt. A gyakorisági mutatók meghatározásához felhasználtuk mindkét magyar szövegkorpuszt. Az elsődleges feladat elvégzése során megvizsgáltuk, hogy az ÉrtSz+ 15.850 címszava előfordul-e, és ha igen, milyen gyakran a fenti két magyar szövegkorpuszban külön-külön. A két korpuszból kinyert gyakorisági adatok segítségével (súlyozást is alkalmazva) állapítottuk meg az egyes címszók gyakorisági osztályát, azaz soroltuk be az 5 gyakorisági osztály valamelyikébe.

## 1 A vizsgálat előzménye

A Magyar Történeti Korpusz (TK) és a Magyar Nemzeti Szövegtár (MNSz) összehasonlítása egy más irányú lexikográfiai feladat elvégzésének melléktermékeként jött létre a TINTA Könyvkiadóban. (Mint ismeretes, a TK-t a Magyar Nagyszótár munkálatai során építették fel, és ez a korpusz 18., 19. és 20. századi magyar szövegrészleteket tartalmaz, míg az MNSz-t a 20. század végén keletkezett elsősorban sajtónyelvi, irodalmi szövegek alkotják.)

Az elsődleges feladat az Értelmező szótár+ (ÉrtSz+ [1]) címszavainak gyakorisági mutatóval való ellátása volt. A gyakorisági mutatók meghatározásához felhasználtuk mindkét magyar szövegkorpuszt[1]. Az utóbbi évtizedekben a nemzetközi szótárirodalomban az angol értelmező szótárak nyomán elterjedt a címszavak gyakoriságának jelölése. A magyar szótárirodalomban a Magyar értelmező kéziszótár (ÉKsz.[2] [2]) közli elsőként a címszavak gyakoriságát.

Az ÉrtSz+ sajátosan izgalmas színt képvisel a magyar, de a nemzetközi szótárkínálatban is, mert erőteljesen túllép az értelmező szótár szokásos funkcióin. Az ÉrtSz+ a szómagyarázó funkció mellett – mint a szótár az alcímében is jelzi – Értelmezések, példamondatok, szinonimák, ellentétek, szólások, közmondások, etimológiák, nyelv-

[1] Ezúton is köszönetet mondunk az MTA Nyelvtudományi Intézetének a két korpusz szokásos felhasználási módját meghaladó vizsgálat engedélyezéséhez.

használati tanácsok és fogalomköri csoportok szerint is átfogó módon dolgozza fel címszóállományát. Sőt a szótár még a címszó gyakoriságát is feltünteti egy ötfokozatú skálán.

Mint említettük, az ÉrtSz+ címszavai gyakoriságának meghatározásához felhasználtuk a TK-t és az MNSz-t. A feladat elvégzése során megvizsgáltuk, hogy az ÉrtSz+ 15.850 címszava előfordul-e, és ha igen, milyen gyakran a fenti két magyar szövegkorpuszban külön-külön. A két korpuszból kinyert gyakorisági adatokat – leegyszerűsítve mondva – átlagoltuk, majd súlyozást is alkalmazva állapítottuk meg az egyes címszók gyakorisági osztályát, azaz soroltuk be az 5 gyakorisági osztály valamelyikébe. A kapott eredmény elemzését követően, a gyakorisági mutatót néhány esetben szubjektív nyelvérzékünk alapján módosítottuk.

## 2 A vizsgálat

Szövegkorpuszok összehasonlítására nincs általánosan elfogadott módszer. A matematikai logika nyelvére lefordítva: két nagy, de véges számú, ismétlődő diszkrét elemeket is tartalmazó halmazt kell összevetni. A halmazokban nemcsak az egyes elemek, azaz szavak megléte vagy hiánya a fontos, hanem az is jellemző, hogy egy-egy szó hányszor fordul elő egy-egy szövegkorpuszban.

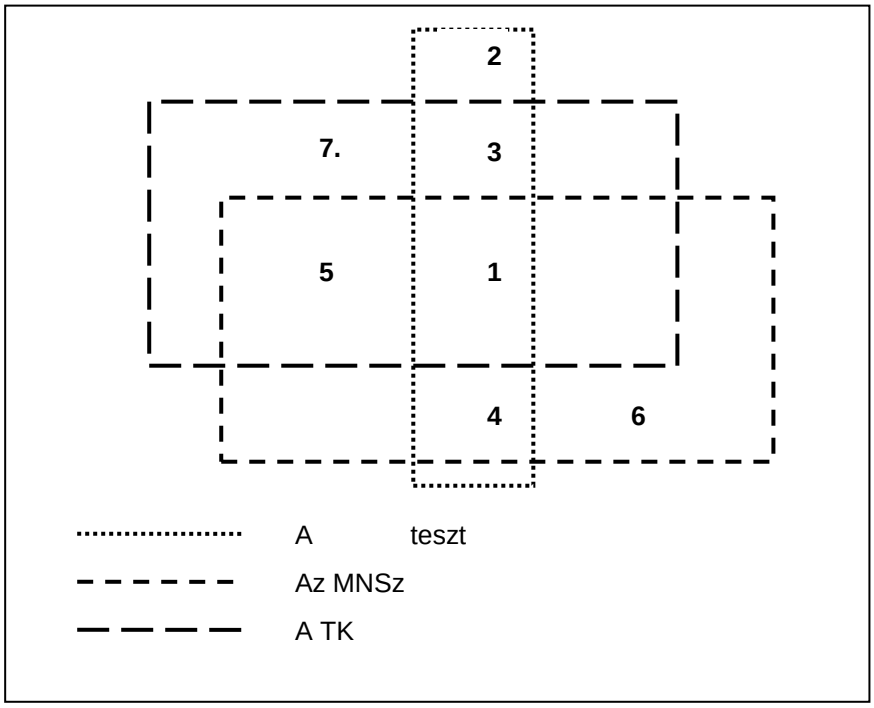
Mivel a gyakorlatban szinte lehetetlen két vizsgált nagy szövegkorpusz minden egyes szavának az összevetéséből keletkezett eredmény kiértékelése, az ÉrtSz+ címszavainak gyakorisági mutatójának a meghatározása megmutatott egy praktikusnak és helyesnek tűnő eljárást ahhoz, hogy két vagy több szövegkorpusz szókincsét miként lehet összevetni. Ugyanis a munkálat rávilágított arra, hogy egy jól megválasztott minta, vagy más néven tesztszólistának a szavait felhasználhatjuk a korpuszok szókincsének összevetésére, összehasonlítására. Természetesen ennek a tesztszólistának a hossza az összehasonlítandó korpuszok szókincséhez képest nem lehet se túl kicsi, se túl nagy. Az ÉrtSz+ címszavainak gyakorisági osztályokba sorolása során úgy tapasztaltuk, hogy a szótár 15.850 címszavának a két korpuszban való előfordulása és az előfordulások gyakorisága jól jellemzi és sajátosan leírja a fenti két korpuszt. Így az ÉrtSz+ címszójegyzékét joggal tekinthetjük vizsgálatunk tesztszólistájának.

Az ÉrtSz+ címszavait különös gonddal válogatták össze a szerkesztők. A címszólista összeállításáról így vallanak: „Milyen szavakat tartalmaz címszóként a szótár? Elsősorban az úgynevezett alapszókincs elemeit, amelyeket a leggyakrabban és legtöbben használunk, ezek között egyaránt vannak fogalomszók (...) és formaszók (...). Az alapszavak mellett sok a tantárgyi szakszó, olyanok, amelyekkel a diákok az irodalom-, a nyelvtan-, a történelem-, a matematika-, a fizikaórán találkozhatnak. (...) Megtalálható a szótárban több olyan régi szó, amelyek már csak irodalmi és történelmi szövegekben olvashatók (...). ezeken kívül vannak benne új, legtöbbször a modern technikával kapcsolatos szavak (...) Helyet kaptak a szótárban olyan szavak is, amelyeket Magyarországon nem használunk, de a határon túli magyarok életéhez hozzátartoznak.'' [1: VIII].

Két korpusz (MNSz és TK) szókincsének (szavainak) és a tesztszólistaként használt ÉrtSz+ címszavainak összevetése során a következő elméleti esetek lehetségesek:

1. A tesztszólista egy adott szava mindkét korpuszban előfordul valahányszor.
2. A tesztszólista egy adott szava nem fordul elő egyik korpuszban sem.
3. A tesztszólista egy adott szava a TK korpuszban előfordul valahányszor, míg az MNSz-ben nem.
4. A tesztszólista egy adott szava az MNSZ-ben előfordul valahányszor, míg a TK-ban nem.
5. A tesztszólistában nem szereplő szó mindkét korpuszban előfordul.
6. A tesztszólistában nem szereplő szó az MNSZ-ben előfordul, de a TK-ban nem.
7. A tesztszólistában nem szereplő szó a TK-ban előfordul, de az MNSZ-ben nem.

A Magyar Nemzeti Szövegtár (MNSz), a Magyar Történeti Korpusz (TK) és a tesztszólista szavainak lehetséges viszonya:




Vizsgálatunk körébe természetszerűleg csak a 1−4. pontok tartoztak, az 5−7. pontok esetei kívül estek figyelmünkön. Mérésünket 2006 májusában végeztük. Ebben az időben az MNSz 111.746.000 szövegszó nagyságú, míg a TK 8.897.000 szövegszó terjedelmű volt. A mérés elején az ÉrtSz+ címszavai közül eltávolítottuk az 1-nél nagyobb homonimaindexszel ellátott címszavakat. A homonimák nélküli tesztszólista az eredetileg 15.810 címszóból álló lista helyett 15.010 szó hosszúságú lett. Ezt követően egy kis robotprogram lekérdezte a tesztszólistát alkotó 15.010 szó előfordulási gyakoriságát a két korpuszban. A MNSz és a TK korpuszok a szövegszavaik korábbi

morfológiai elemzése során − ha kis mértékben is − eltérő morfológiai elemzési technikát alkalmaztak, ezért az egyes esetekhez tartozó szavak számát 10-es értékre kerekítettük.

# 3 A vizsgálat eredményei

## 3.1 Hasonlósági mutatók

Számszerűsítve a következő eredményt kaptuk a teszt szólista 15.010 szava és a vizsgált két korpusz szavainak viszonyára:

1. A tesztszólista 15.010 szavából 14.290 szó előfordult mind a TK-ban, mind az MNSz-ben; 95,20%
2. A tesztszólista 15.010 szavából 45 szó nem fordult elő egyik szövegkorpuszban sem; 0,30%
3. A tesztszólista 15.010 szavából 670 szó csak az MNSz-ben fordult elő, a TK-ban nem szerepelt; 4,46%
4. A tesztszólista 15.010 szavából 5 szó csak a TK-ban fordult elő, az MNSz-ben nem szerepelt; 0,03%.

Az adatokból látszik, hogy a két magyar szövegkorpusz szókincse hasonló, jelentős átfedést mutat egymással. A tesztszólistának használt ÉrtSz+ címszójegyzéke a szótár jellegéből és funkciójából adódóan jobban illeszkedik a MNSz-hez, mint a TK-hoz.

## 3.2 Neologizmusok

Az MNSz jellegéből adódóan joggal feltételezhetjük, hogy azok a szavak, amelyek csak az MNSz-ben fordulnak elő, jól jellemzik a 20. század végét, illetve az ezredfordulót, és így ezek a magyar szókincs neologizmusai közé tartoznak. Feltételezésünk igazolására közreadjuk annak az 50 leggyakoribb szónak a listáját, amelyek az MNSz-ben előfordulnak, de a TK-ban nem találhatók meg (a szó melletti szám a szó MNSz-beli előfordulását mutatja).

Neologizmusok, az MNSz leggyakoribb korfestő szavai:

*közszolgálati* 6764
*munkáltató* 6263
*honlap* 5383
*euró* 4515
*közterület* 4415
*sportág* 4204
*CD* 3890
*piacgazdaság* 3743
*e-mail* 3728
*szoftver* 3611
*környezetvédelem* 3592
*digitális* 3590
*szia* 3475
*drog* 3399
*parkoló* 3373
*atomerőmű* 3217
*elsőfokú* 3184
*internet* 3057
*közalkalmazott* 3008
*foci* 2541
*bevásárlóközpont* 2476
*kosárlabda* 2323
*világháló* 2294
*társasház* 2234
*tömegközlekedés* 2051
*informatika* 1993
*globalizáció* 1833
*videó* 1726

*bróker* 1566
*multi* 1378
*természetvédelem* 1136
*AIDS* 1079
*telefax* 1060
*hobbi* 1055
*tizenéves* 970
*rajzfilm* 955
*mobiltelefon* 912
*tévécsatorna* 873
*interaktív* 860
*kemping* 852
*versenyszféra* 803
*éllovas* 794
*elektronika* 789
*sci-fi* 761
*papírforma* 755
*hardver* 735
*lakópark* 693
*akciófilm* 666
*sikerdíj* 636
*bulvárlap* 635

### 3.3 Archaizmusok

Ezt követően felmerül a kérdés, hogy a fenti neologizmusokhoz hasonlóan a magyar nyelv archaizmusait is kigyűjthetjük-e a két korpusz szóanyagának az összevetéséből? Mivel vizsgálatunkban a TK-ban nem találtunk jelentős számban olyan szavakat, amelyek csak abban fordulnak elő, és nem találhatók meg a MNSz-ben – direkt módon nem gyűjthetünk archaizmusokat a TK-ból. Azonban adódik a gondolat, hogy talán archaizmusnak tekinthetők azok a szavak is, amelyek arányaiban jóval többször fordulnak elő a TK-ban, mint az MNSz-ben. A gondolat életrevalónak tűnik, bizonyításképpen alább közreadjuk az első 50 olyan szót, amelyek jelentősen többször fordulnak elő a TK-ban, mint az MNSz-ben (a szót követő szám az előfordulási arányt mutatja).

Archaizmusok, azok a szavak, amelyek arányaiban a TK-ban jelentősen többször fordulnak elő, mint az MNSz-ben:

*aprószentek* 15
*asztag* 12
*beszély* 15
*billikom* 45
*borong* 13
*bőszült* 20
*burnót* 24
*csepű* 14
*csibuk* 28
*csigáz* 99
*csöngettyű* 13
*dicső* 13
*divatozik* 22
*dragonyos* 16
*epeszt* 36
*fejkötő* 20
*findzsa* 21
*gondola* 16
*hajdankor* 23
*hanga* 23
*hazámfia* 64
*hevenyében* 15
*honfi* 28
*honn* 47
*horgany* 50
*ispán* 16
*játszi* 26
*kebel* 16
*kegyed* 25
*komika* 30
*komorna* 24
*kopja* 20
*korhely* 13
*ködmön* 13
*mál* 28
*málé* 11
*mente* 14
*messzely* 28
*midőn* 34
*nadály* 18
*okuláré* 11
*orozva* 27
*pacsuli* 20
*patvar* 14
*sarjadék* 17
*sólya* 14
*süveg* 13
*szövétnek* 11
*szüle* 18
*tekintetes* 18
*téns* 54
*tragika* 21
*urambátyám* 20
*vágta* 23
*várta* 109
*vitézkötés* 13

### 3.4 A magyar szókincs magja

Mind a két vizsgált korpuszban előfordul a tesztszólista szavainak 95,2%-a. Mint láttuk, nem mindegy azonban, hogy a közös szavak előfordulásának mi az aránya. A következőkben közreadunk mintaképpen 50 olyan gyakori szót, amelyek előfordulási aránya megegyezik vagy közel azonos mindkét szövegkorpuszban. Az 50 mintaszó mindegyikét az ÉrtSz+ leggyakoribb címszavai közül választottuk, azaz mindegyik az első gyakorisági kategóriába tartozik a szótár öt kategóriájából. A szavak mögött álló szám az MNSz-ben és a TK-ban lévő előfordulások aránya. Ha a szám nagyobb egynél, akkor arányaiban az MNSz-ben fordult elő többször a szó, ellenkező esetben a TK-ban gyakoribb.

Megállapíthatjuk, hogy ezek a szavak fontosak, a magyar szókincs középpontjában, magjában helyezkednek el, hiszen használatuk több mint két évszázadon át gyakori és állandó intenzitású a korpuszok adatai alapján.

Állandó intenzitású és gyakori szavak a magyar szókincs magjából:

*ad* 1,13
*csinál* 0,99
*elmegy* 0,87
*esik* 1,15
*este* 0,92
*férfi* 1,17
*fiatal* 0,90
*gondol* 1,03
*győz* 0,86
*három* 0,99
*hat* (szn) 1,02
*hisz* 0,85
*hoz* 1,00
*idő* 1,13
*ismer* 0,89
*jut* 1,06
*katona* 0,88
*két* 1,17
*kevés* 0,91
*könnyű* 0,83
*könyv* 1,03
*magas* 0,92
*magyaráz* 1,00
*mond* 0,83
*nagyon* 1,18
*nap* 1,03
*négy* 1,17
*nehéz* 0,93
*név* 1,20
*nyár* 1,09
*ok* 1,13
*olvas* 0,88
*óra* 1,17
*orvos* 1,02
*rendel* 0,85
*rossz* 1,03
*sok* 1,03
*század* 0,89
*széles* 0,89
*szeret* 1,10
*szó* 0,92
*szolgál* 0,92
*tart* 1,18
*tartozik* 1,04
*teremt* 1,15
*tud* 1,19
*út* 1,12
*város* 1,13
*vesz* 0,95
*víz* 0,87

## 4 Összegzés

Összegzésként elmondhatjuk, hogy újszerű vizsgálatunk bebizonyította, hogy két szövegkorpusz szókincsének összehasonlítása eredményesen elvégezhető egy kisebb alkalmas tesztszólista – akár egy szótár címszójegyzékének − segítségével.

Két különböző jellegű szövegkorpuszból hatékonyan gyűjthetők ki egy megfelelő tesztszólista segítségével a korpuszokra külön-külön is jellemző szavak, a mi esetünkben neologizmusok és archaizmusok. A kigyűjtött neologizmusok a 20. század végé-

nek, az ezredfordulónak a korfestői, míg az archaizmusok jól jellemzik a 18. század vége, illetve a 19. század magyar szókincsét.

Ugyanakkor megállapítható, hogy a két korpusz közös elemei közül azok, amelyek előfordulása magas és az előfordulások aránya közel azonos az egyes korpuszokban, a magyar szókincs magját képezik, így köznyelvi szótárak címszójegyzékének összeállításához eredményesen használhatók.

A két magyar szövegkorpusz összehasonlítása során nyert eredmények rávilágítanak arra is, hogy több és különböző típusú magyar szövegkorpuszra lenne szükség, mert a szövegkorpuszok összevetésével jellegzetes csoportokat alkotó szavak gyűjthetők egybe többé-kevésbé automatikusan. A korpuszokból kigyűjtött szócsoportok alkalmas kiindulásai lehetnek a magyar szókészlet különböző szempontú szótári munkálatainak.

## Hivatkozások

1. Előry V. (főszerk.): Értelmező szótár+. Értelmezések, példamondatok, szinonimák, ellentétek, szólások, közmondások, etimológiák, nyelvhasználati tanácsok és fogalomköri csoportok. TINTA Könyvkiadó, Budapest (2007)
2. Pusztai F. (főszerk.): Magyar értelmező kéziszótár[2]. Akadémiai Kiadó, Budapest (2003)