

KISS GÁBOR – KISS MÁRTON

**A Magyar Történeti Korpusz és a Magyar Nemzeti Szövegtár
szókincsének összehasonlítása az *Értelmező szótár+*
címszójegyzéke felhasználásával**

Ce travail présente la comparaison du lexique du Corpus historique du hongrois (TK) et du Corpus national du hongrois (MNSz), fondée sur la liste des vedettes (15.850 mots) de l'*Értelmező szótár+* (ÉrtSz+) (Dictionnaire unilingue du hongrois+). Grâce à cette méthode, nous avons pu sélectionner de façon efficace les néologismes et les archaïsmes caractéristiques de ces deux corpus différents. La comparaison nous a aussi permis de constater que les éléments communs les plus fréquents des deux corpus constituent le noyau central du lexique hongrois.

I. 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+; Eöry, 2007) címszavainak gyakorisági mutatóval való ellátása volt. A gyakorisági mutatók meghatározásához felhasználtuk mindkét magyar szövegkorpuszt¹. Az utóbbi évtizedekben a nemzetközi szótáriradalomban az angol értelmező szótárak nyomán elterjedt a címszavak gyakoriságának jelölése. A magyar szótáriradalomban a *Magyar értelmező kéziszótár* (ÉKsz.²; Pusztai, 2003) közli elsőként a címszavak gyakoriságát².

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

² Azt, hogy az ÉKsz.² címszavainak gyakorisági besorolása nem volt problémamentes, és a feladat elvégzése után kapott eredmény sem hibátlan, jelzi, hogy kétségtelenül a legritkább magyar szavak közé tartozó *ball* szó, amely a szótár szerint a 'A földrengések erősségének egyik mértékegysége', közepes gyakoriságú besorolással szerepel a kiadványban.

Az ÉrtSz+ sajátosan izgalmas szint 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, nyelvhasználati 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 szövegtörzsben külön-külön. A két törzsből 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.

II. A vizsgálat

Szövegtörzsek 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 szövegtörzsben.

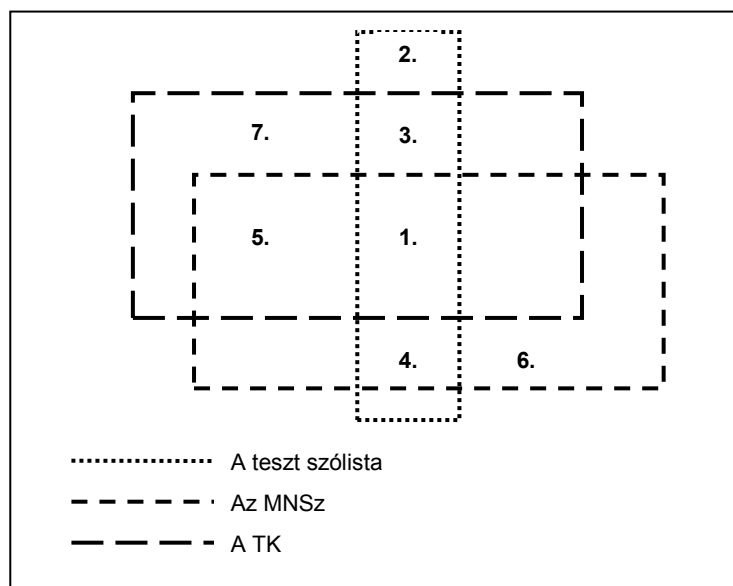
Mivel a gyakorlatban szinte lehetetlen két vizsgált nagy szövegtörzs minden egyes szavának az összevetéséből keletkezett eredmény kiértékelése, az ÉrtSz+ címszavainak gyakorisági mutatójának a meghatározása megmutatott egy praktikusnak és helyesnek tűnő eljárást ahhoz, hogy két vagy több szövegtörzs szókincsét miként lehet összevetni. Ugyanis a munkálat rávilágított arra, hogy egy jól megválasztott minta, vagy más néven *teszt szólistának* a szavait felhasználhatjuk a törzsek szókincsének összevetésére, összehasonlítására. Természetesen ennek a teszt szólistának a hossza az összehasonlítandó törzsek 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 törzsben való

előfordulása és az előfordulások gyakorisága jól jellemzi és sajátosan leírja a fenti két korpuszt. Így az ÉrtSz+ címszójegyzékét joggal tekinthetjük vizsgálatunk teszt szólistájának.

Az ÉrtSz+ címszavait különös gonddal válogatták össze a szerkesztők. A címszólista összeállításáról így vallanak: „Milyen szavakat tartalmaz címszóként a szótár? Elsősorban az úgynevezett alapszókinccs elemeit, amelyeket a leggyakrabban és legtöbbször 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.” (Eőry, 2007: VIII).

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

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



A Magyar Nemzeti Szövegtár (MNSz), a Magyar Történelmi Korpusz (TK)
 és a *teszt szólista* szavainak lehetséges viszonya

Vizsgálatunk körébe természetesen 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 *teszt szólista* az eredetileg 15.810 címszóból álló lista helyett 15.010 szó hosszúságú lett. Ezt követően egy kis robotprogram lekérdezte a teszt szólistát 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.

III. A vizsgálat eredményei

a) Hasonlósági mutatók

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

1. A teszt szólista 15.010 szavából 14.290 szó előfordult mind a TK-ban, mind az MNSz-ben; 95,20%
2. A teszt szólista 15.010 szavából 45 szó nem fordult elő egyik szövegtörzsben sem; 0,30%
3. A teszt szólista 15.010 szavából 670 szó csak az MNSz-ben fordult elő, a TK-ban nem szerepelt; 4,46%
4. A teszt szólista 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 szövegtörzs szókinése hasonló, jelentős átfedést mutat egymással. A teszt szólistának használt ÉrtSz+ címszójegyzéke a szótár jellegéből és funkciójából adódóan jobban illeszkedik a MNSz-hez, mint a TK-hoz.

b) 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 szókinés 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):

<i>közszolgálati</i> 6764	<i>internet</i> 3057	<i>tizenéves</i> 970
<i>munkáltató</i> 6263	<i>közalkalmazott</i> 3008	<i>rajzfilm</i> 955
<i>honlap</i> 5383	<i>foci</i> 2541	<i>mobiltelefon</i> 912
<i>euró</i> 4515	<i>bevásárlóközpont</i> 2476	<i>tévécsonna</i> 873
<i>közterület</i> 4415	<i>kosárlabda</i> 2323	<i>interaktív</i> 860
<i>sportág</i> 4204	<i>világháló</i> 2294	<i>kemping</i> 852

<i>CD</i> 3890	<i>társasház</i> 2234	<i>versenyszféra</i> 803
<i>piacgazdaság</i> 3743	<i>tömegközlekedés</i> 2051	<i>éllovas</i> 794
<i>e-mail</i> 3728	<i>informatika</i> 1993	<i>elektronika</i> 789
<i>szoftver</i> 3611	<i>globalizáció</i> 1833	<i>sci-fi</i> 761
<i>környezetvédelem</i> 3592	<i>videó</i> 1726	<i>papírforma</i> 755
<i>digitális</i> 3590	<i>bróker</i> 1566	<i>hardver</i> 735
<i>szia</i> 3475	<i>multi</i> 1378	<i>lakópark</i> 693
<i>drog</i> 3399	<i>természetvédelem</i> 1136	<i>akciófilm</i> 666
<i>parkoló</i> 3373	<i>AIDS</i> 1079	<i>sikerdíj</i> 636
<i>atomerőmű</i> 3217	<i>telefax</i> 1060	<i>bulvárlap</i> 635
<i>elsőfokú</i> 3184	<i>hobby</i> 1055	

Neologizmusok, az MNSz leggyakoribb korfestő szavai

c) 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 találhatóak 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 *arányosan* 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).

<i>aprószentek</i> 15	<i>hanga</i> 23	<i>midőn</i> 34
<i>asztag</i> 12	<i>hazámfia</i> 64	<i>nadály</i> 18
<i>beszély</i> 15	<i>hevenyében</i> 15	<i>okuláré</i> 11
<i>billikom</i> 45	<i>honfi</i> 28	<i>orozva</i> 27
<i>borong</i> 13	<i>honn</i> 47	<i>pacsuli</i> 20
<i>bőszült</i> 20	<i>horgany</i> 50	<i>patvar</i> 14
<i>burnót</i> 24	<i>ispán</i> 16	<i>sarjadék</i> 17
<i>csepű</i> 14	<i>játszi</i> 26	<i>sólya</i> 14
<i>csibuk</i> 28	<i>kebel</i> 16	<i>süveg</i> 13

<i>csigáz</i> 99	<i>kegyed</i> 25	<i>szövétnék</i> 11
<i>csöngettyű</i> 13	<i>komika</i> 30	<i>szüle</i> 18
<i>dicső</i> 13	<i>komorna</i> 24	<i>tekintetes</i> 18
<i>divatozik</i> 22	<i>kopja</i> 20	<i>téns</i> 54
<i>dragonyos</i> 16	<i>korhely</i> 13	<i>tragika</i> 21
<i>epeszt</i> 36	<i>ködmön</i> 13	<i>urambátyám</i> 20
<i>fejkötő</i> 20	<i>mál</i> 28	<i>vágta</i> 23
<i>findzsa</i> 21	<i>málé</i> 11	<i>várta</i> 109
<i>gondola</i> 16	<i>mente</i> 14	<i>vitézkötés</i> 13
<i>hajdankor</i> 23	<i>messzely</i> 28	

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

d) A magyar szókincs magja

Mind a két vizsgált korpuszban előfordul a teszt szólista 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 szövegtörzsben. 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.

<i>ad</i> 1,13	<i>két</i> 1,17	<i>rendel</i> 0,85
<i>csinál</i> 0,99	<i>kevés</i> 0,91	<i>rossz</i> 1,03
<i>elmege</i> 0,87	<i>könnyű</i> 0,83	<i>sok</i> 1,03
<i>esik</i> 1,15	<i>könyv</i> 1,03	<i>század</i> 0,89
<i>este</i> 0,92	<i>magas</i> 0,92	<i>széles</i> 0,89

<i>férfi</i> 1,17	<i>magyaráz</i> 1,00	<i>szeret</i> 1,10
<i>fiatal</i> 0,90	<i>mond</i> 0,83	<i>szó</i> 0,92
<i>gondol</i> 1,03	<i>nagyon</i> 1,18	<i>szolgál</i> 0,92
<i>győz</i> 0,86	<i>nap</i> 1,03	<i>tart</i> 1,18
<i>három</i> 0,99	<i>négy</i> 1,17	<i>tartozik</i> 1,04
<i>hat (szn)</i> 1,02	<i>nehéz</i> 0,93	<i>teremt</i> 1,15
<i>hisz</i> 0,85	<i>név</i> 1,20	<i>tud</i> 1,19
<i>hoz</i> 1,00	<i>nyár</i> 1,09	<i>út</i> 1,12
<i>idő</i> 1,13	<i>ok</i> 1,13	<i>város</i> 1,13
<i>ismer</i> 0,89	<i>olvas</i> 0,88	<i>vesz</i> 0,95
<i>jut</i> 1,06	<i>óra</i> 1,17	<i>víz</i> 0,87
<i>katona</i> 0,88	<i>orvos</i> 1,02	

Állandó intenzitású és gyakori szavak a magyar szókincs magjából

IV. Összegzés

Összegzésként elmondhatjuk, hogy újszerű vizsgálatunk bebizonyította, hogy két szövegtörzs szókincsének összehasonlítása eredményesen elvégezhető egy kisebb alkalmas *teszt szólista* – akár egy szótár címszójegyzékének – segítségével.

Két különböző jellegű szövegtörzsből hatékonyan gyűjthetők ki egy megfelelő teszt szólista segítségével a törzsekre külön-külön is jellemző szavak, a mi esetünkben neologizmusok és archaizmusok. A kigyűjtött neologizmusok a 20. század végének, az ezredfordulónak a korfestői, míg az archaizmusok jól jellemzik a 18. század vége, illetve a 19. század magyar szókincsét.

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

A két magyar szövegtörzs összehasonlítása során nyert eredmények rávilágítanak arra is, hogy több és különböző típusú magyar szövegtörzsről lenne szükség, mert a szövegtörzsek összevetésével jellegzetes csoportokat alkotó szavak gyűjthetők egybe többé-kevésbé automatikusan. A törzsekből

kigyűjtött szócsoportok alkalmas kiindulásai lehetnek a magyar szókészlet különböző szempontú szótári munkálatainak.

Bibliográfia

EÓRY, Vilma (főszerk.), 2007, *Értelmező szótár+. Értelmezések, példamondatok, szinonimák, ellentétek, szólások, közmondások, etimológiák, nyelvhasználati tanácsok és fogalomköri csoportok*, Budapest, TINTA Könyvkiadó.

PUSZTAI, Ferenc (főszerk.), 2003, *Magyar értelmező kéziszótár*, Budapest, Akadémiai Kiadó.

KISS GÁBOR

Tinta Könyvkiadó

E-mail: kissgabo@tintakiado.hu

KISS MÁRTON

Tinta Könyvkiadó

E-mail: kissmarci@tintakiado.hu