A Regresszív Képzeleti Szótár magyar nyelvű változatának létrehozása

Pólya Tibor[1], Szász Levente[1, 2]

[1] MTA TTK Kognitív Idegtudományi és Pszichológiai Intézet
1132 Budapest, Victor Hugó utca 18-22.
polya.tibor@mta.ttk.hu
[2] Pécsi Tudományegyetem, Pszichológiai Intézet
7624 Pécs, Ifjúság útja 6.
levente.szasz@mtapi.hu

Kivonat: A Regresszív Képzeleti Szótár az egyik legelterjedtebben használt automatikus pszichológiai szövegelemző eljárás. Az előadás a szótár magyar nyelvű változatának elkészítési folyamatát mutatja be. A magyar nyelvű szótár megbízhatóságának mérését Wilson [17] eljárása alapján végeztük el. Eredményeink azt mutatják, hogy a Regresszív Képzeleti Szótár magyar nyelvű változata megbízható mérési eszköz.

1 A Regresszív Képzeleti Szótár

A Regresszív Képzeleti Szótárt – amelynek eredeti neve Regressive Imagery Dictionary, röviden RID – Colin Martindale [5] hozta létre angol nyelven. A RID a legismertebb pszichológiai tartalomelemző eljárások közé tartozik. Népszerűségét mutatja az is, hogy az elmúlt négy évtized során öt nyelvre fordították le [9]. A RID az elsődleges és a másodlagos gondolkodási folyamatokra utaló tartalmakat azonosítja a szövegben. Az elsődleges gondolkodási folyamatra az jellemző, hogy asszociatív, konkrét és a realitáshoz kevésbé kapcsolódó [12]. A fantázia, az ábrándozás és az álmok fő gondolkodási módja [10]. Ezzel szemben a másodlagos gondolkodási folyamat absztrakt, logikus, realitás központú és problémamegoldásra fókuszáló [12].

A RID az elsődleges és másodlagos gondolkodási folyamatokhoz kapcsolódó tartalmakat a szavak szintjén azonosítja, amihez hierarchikusan szervezett szótárakat használ. A hierarchikus szerveződésnek két csúcskategóriája van: az elsődleges gondolkodási folyamatra utaló szavak szótára, amely 1828 szócsonkot tartalmaz és a másodlagos gondolkodási folyamatra utaló szavak szótára, amely 714 szócsonkot foglal magában. (A két csúcskategóriát Martindale a későbbiekben kiegészítette egy érzelmi szótárral is. Ez azonban elméletileg nem kapcsolódik a gondolkodási mód fogalmához, így ezt a szótárt nem fordítottuk le. Ugyanakkor rendelkezésre áll magyar nyelvű érzelmi szótár [4].) Az *Elsődleges gondolkodási folyamatok szótára* 5 kategóriára bomlik, ezek a kategóriák a következő szinten 29 alszótárt foglalnak magukban

(lásd 1. táblázat). A *Másodlagos gondolkodási folyamatok szótára* 7 alszótárból áll, és nem tartalmaz köztes szintet (lásd 2. táblázat).

1. táblázat: Az elsődleges gondolkodási folyamat kategóriái angol és magyar nyelvű példákkal.

Kategória		**Angol nyelvű példák**	**Magyar nyelvű példák**
Drive	**Oralitás**	Breast, drink, lip	Mell, iszik, ajak
	Analitás	Sweat, rot, dirty	Izzad, rohad, piszkos
	Szex	Lover, kiss, naked	Szerető, csók, meztelen
Érzékelés	**Általános érzékelés**	Fair, charm, beauty	Tetszetős, báj, szépség
	Érintés	Touch, thick, stroke	Érintés, sűrű, cirógat
	Íz	Sweet, taste, bitter	Édes, íz, keserű
	Szag	Breath, perfume, scent	Lehelet, parfüm, illat
	Hang	Hear, voice, sound	Hall, hang, zörej
	Látvány	See, light, look	Lát, fény, néz
	Hideg	Cold, winter, snow	Hideg, tél, hó
	Kemény	Rock, stone, hard	Szikla, kő, kemény
	Lágy	Soft, gentle, tender	Lágy, enyhe, puha
Védekezés	**Passzivitás**	Die, lie, bed	Meghal, fekszik, ágy
	Utazás	Wander, desert, pilgrim	Vándorlás, sivatag, zarándok
	Random mozgás	Wave, roll, spread	Hullám, gurul, terjed
	Diffúzió	Shadow, cloud, fog	Árnyék, felhő, köd
	Káosz	Wild, crowd, jungle	Vad, tömeg, dzsungel
Regresszió	**Ismeretlen**	Secret, mystic, unknown	Titok, misztikus, ismeretlen
	Időtlen	Eternal, forever, immortal	Örök, örökké, halhatatlan
	Tudatváltozás	Dream, sleep, wake	Álom, alszik, ébred
	Áthaladás	Road, wall, door	Út, fal, ajtó
	Nárcizmus	Eye, heart, hand	Szem, szív, kéz
	Konkrétság	Here, behind, west	Itt, mögött, nyugat
Ikaroszi képzelet	**Emelkedés**	Rise, fly, throw	Emelkedik, repül, eldob
	Magasság	Airplane, bird, tower	Repülőgép, madár, torony
	Esés	Fall, slide, sink	Zuhan, csúszda, süllyed
	Mélység	Cave, valley, submarine	Barlang, völgy, tengeralattjáró
	Tűz	Fire, flame, smoke	Tűz, láng, füst
	Víz	Sea, water, swim	Tenger, víz, úszik

2. táblázat: A másodlagos gondolkodási folyamat kategóriái angol és magyar nyelvű példákkal.

Kategória	**Angol nyelvű példák**	**Magyar nyelvű példák**
Absztrakció	Know, reason, think	Tud, ok, gondol
Társas	Tell, help, advice	Mond, segít, tanács
Instrumentális	Win, find, work	Nyer, talál, munka
Korlátozás	Arrest, forbid, stop	Letartóztat, tilt, megállít
Rend	List, simple, symmetric	Lista, egyszerű, szimmetrikus
Idő	Yesterday, year, now	Tegnap, év, most
Erkölcs	Law, virtue, responsibility	Törvény, erény, felelősség

A RID pszichológiai validitását számos empirikus vizsgálat eredménye igazolta, amelyeket – többek között – gyerekektől [16], pszichotikus betegektől [14], illetve akut droghatás alatt álló személyektől [15] nyert szövegeken végeztek el. A RID-et gyakran alkalmazzák az irodalmi szövegek alkotásához köthető pszichológiai folyamatok kutatására is [6].

2 A magyar Regresszív Képzeleti Szótár fordításának folyamata

2.1 Döntés a karakteralapú keresés alkalmazása mellett

A RID magyar nyelvű változatát – az angol eredetivel megegyező módon – a karakteres keresés elvén hoztuk létre. Választásunkat két szempont indokolta. Egyrészt a pszichológiai szövegelemzésben a karakteres keresést alkalmazó tartalomelemző szoftverek terjedtek el (például WordStat [2], LIWC [8]). Így a karakteres keresés elvét alkalmazva könnyebben tudjuk kombinálni ezt az elemzési eljárást más elemzési eszközökkel. Másrészt a munka elkezdésekor – 2010-ben – nem állt rendelkezésünkre megfelelő lefedettséget biztosító magyar nyelvű szótár.

2.2 A folyamat fontosabb lépéseinek áttekintése

Az első lépés az úgynevezett nyers fordítás elkészítésének fázisa volt. Ennek során az angol nyelvű RID-ben szereplő szócsonkok alapján összegyűjtöttük azokat a magyar nyelvű szavakat, amelyek angol megfelelőit a RID angol változata találatként azonosítja. Ebben a munkában 10 pszichológus hallgató vett részt.

A második lépésben ezen szavak listájáról a cikk két szerzője kiválogatta azokat a szavakat, amelyek jelentése kapcsolódik a RID valamelyik alszótárához.

Harmadik lépésként előállítottuk azokat a magyar nyelvű szócsonkokat, amelyek a toldalékolástól függetlenül azonosítják az előző lépésben felsorolt szavakat. Az így kapott szócsonkokat találati listákon helyeztük el, amelyeken helyet kaptak többszavas kifejezések is.

2.3 A fejlesztéshez használt program

A szótárépítést a Max Silberztein által megalkotott NooJ [11] számítógépes nyelvi fejlesztő környezete segítségével valósítottuk meg. A NooJ grafikus felületét felhasználva hoztuk létre a gráfokat vagy más néven lokális nyelvtanokat. A szavakat virtuális keretekbe, úgynevezett boxokba helyeztük el. Ezek tetszőleges módon összeköthetőek, így akár több szóból álló, szintaktikai információt is tartalmazó keresőkifejezések is létrehozhatóak.

2.4 A keresés módja

A karakteres keresésnek két módja van. Az alapértelmezett mód a kezdő karaktersor megadása. Ebben az esetben az algoritmus az összes olyan szót megtalálja, amely ezt a feltételt teljesíti. Például a "szép*" karaktersor (*Általános érzékelés kategória*) megadásával a rendszer kinyeri a szövegből a "szépet", "szépnek", "szépről" stb. alakokat is. (A „*" karakter azt jelöli, hogy a szócsonk tetszőleges karakterrel/karakterekkel folytatódhat.)

Figyelembe vettük azt is, hogy bizonyos lexémák változó tővel rendelkeznek. Emiatt például az „alma" (*Oralitás kategória*) szó esetében az „almá*" karaktersort is felvettünk a listára, hogy többek között a birtokos személyjellel ellátott „almám", valamint a tárgyas „almát" alakokat is felismerje a rendszer.

A kettős mássalhangzóra végződő szavaknál úgy kellett megadnunk a kezdő karaktersort, hogy a –val, –vel ragos hasonult alakokat is megtalálja a kereső algoritmus. Például „kalács*" (*Oralitás kategória*) helyett „kalác*" gráfba építésére volt szükség a „kaláccsal" szóalak megtalálása érdekében.

A karakteres keresés második módja a pontos karaktersor megadása, amely csak a teljes mértékben egyező karaktersorból álló szóalakra ad találatot. Ezt alkalmaztuk például az "itt" (*Konkrétság kategória*) határozószó felismeréséhez. A kettőnél több tővel rendelkező igéknél is egyszerűbbnek bizonyult az összes ragozott alak pontos bemásolása az adott gráfba ahhoz képest, mintha például az „eszik" (*Oralitás kategória*) ige „esz-", „ev-"„e-", „é-", „en-" töveit adtuk volna meg kezdő karaktersorként, mivel így nagyon sok téves találat keletkezett volna.

A kezdő karaktersorral való azonosítást jóval gyakrabban alkalmaztuk, mint a pontos karaktersorral való azonosítást.

2.4.1 A találatok és kizárások

Találati listák létrehozása mellett készítettünk olyan listákat is, amelyeket a NooJ kizár az elemzésből. Például a "menta*" (*Oralitás kategória*) kezdő karaktersor megadásával kinyerésre kerül a szövegből a "mentalevél" szó, ami beletartozik az *Oralitás kategóriába,* azonban a "mentalitás" és "mentalista" szavak is találatként jelentkeznek, holott ezek nem tartoznak bele ebbe a kategóriába. Ezért az utóbbiakat felvettük a kizárási listára, amit az '+EXCLUDE' „tag" használatával valósítottunk meg.

Minden egyes szócsonk esetén az összes lehetséges téves találatot számításba vettük. Ezt az ELRAGOZ (Elektronikus magyar ragozási szótár [3]) programnak az a funkciója tette lehetővé, amely valamennyi olyan szót kilistáz (a szoftver memóriájá-

ban tárolt 73810 címszó közül), amely a felhasználó által megadott karaktersorral kezdődik. A „nyer*” (*Instrumentális kategória*) karaktersor esetén a listába kerül például a „nyers” és a „nyereg” szó is.

2.5 Az igekötős igék kezelése

Ha egy adott ige és a belőle származtatható összes igekötős alak adekvátnak számított egy adott alszótár szempontjából, akkor felsoroltuk az összes olyan esetet, ahol az igekötő az ige előtt áll – vele egybeírva. Például „besegít”, „átsegít”, „kisegít” (*Társas kategória*). Ezeken túl csak magát az igét kellett megnevezni („segít”), amelynek megadásával egyúttal a fordított sorrendű változatok is (például: „segít be”, „segít át”) megtalálására kerülnek a gráf lefuttatásakor.

Amennyiben azonban az adott ige csak bizonyos igekötőkkel képez találatot, másokkal együtt állva pedig kategórián kívülinek minősül, akkor magának az igének (például „dönt” [*Absztrakció kategória*]), valamint az ’igekötő az ige előtt áll’ formáknak (például: „eldönt”) a gráfban történő feltüntetésén túl az is szükséges volt, hogy az adott alszótár szempontjából nem odaillő, fordított sorrendű változatokat, például a „dönt fel” kifejezést kizárjuk.

2.6 Az azonos alakú szavak esete

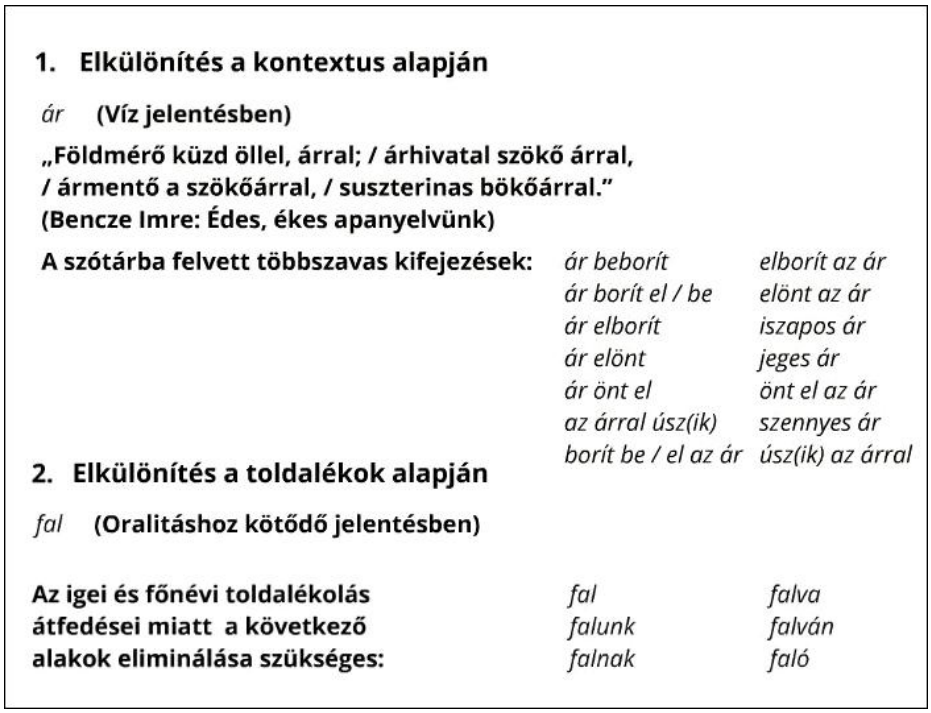
A karakteres kereső algoritmusok létrehozásakor az egyik leginkább időigényes folyamatot az azonos alakú, találati és téves találati minőségben egyaránt előforduló szavak elkülönítése jelentette. Ezekben az esetekben leggyakrabban az ige és a névszó differenciálására volt szükség.

A kiindulást minden esetben az jelentette, hogy a Magyar Nemzeti Szövegtár [13] korpusznyelvészeti adatbázis segítségével felmértük a találati és a téves találati előfordulások gyakoriságát. Ezekre az adatokra támaszkodva hoztuk meg a döntésünket arra vonatkozólag, hogy szerepeltessük-e az adott karaktersort a szótárban, és amennyiben igen, akkor milyen módon végezzük az egyértelműsítést. Erre mutatunk az alábbiakban két példát.

Az elkülönítés egyik módja a kontextus figyelembevételével történt. Ebben az esetben több szóból álló kifejezéseket használtunk fel az azonosításhoz. Például az „ár” szónak a *Víz kategória* szempontjából adekvát jelentésén kívül más használatai is ismeretesek (lásd az 1. ábrán szereplő idézetet). Emiatt magának az „ár” karaktersornak a találati listára való felvétele helyett kizárólag az 1. ábrán szereplő kifejezéseket szerepeltettük a gráfban.

Másik lehetőség a toldalékok alapján történő elkülönítés volt. A „fal” (eszik) igeként az *Oralitás kategóriába* tartozik, főnévként (épület része) azonban nem képezi részét sem ennek az alszótárnak, sem más alszótárnak. Annak érdekében, hogy az alaki egybeesés ellenére – adekvát jelentésben – szerepelhessen a kategóriában, az ELRAGOZ program segítségével kilistáztuk a „fal” szó toldalékolt alakjait mind az igei, mind a főnévi előfordulás szerint. Eliminátuk azokat a szóalakokat (lásd 1. ábra), amelyek egybeesést mutattak. Ez 3 eset törlését jelentette, a többi igealakot, ami-

ből 56 volt, feltüntethettük a szótárban. Továbbá eltávolításra került három igenév is, amelyek két másik főnév (falu és faló) meghatározott alakjaival voltak azonosak.

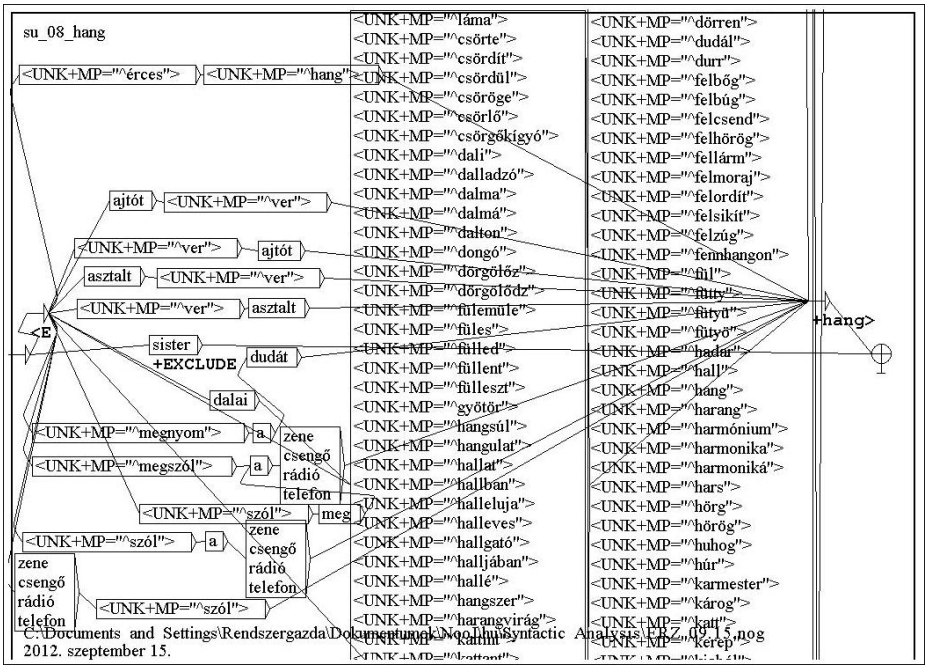


1. ábra: Példák az azonos alakú szavak elkülönítésének lehetőségeire.

A magyar nyelvű változat *Elsődleges gondolkodási folyamat szótára* 4521 karaktersort és 260 két vagy több karaktersorból álló kifejezést tartalmaz. A *Másodlagos gondolkodási folyamat szótár* 2020 karaktersorból és 1098 kifejezésből áll. A kizárási listán 1785 karaktersor, illetve kifejezés szerepel. Az egyik alszótár, a *Hang* gráfjának részlete a 2. ábrán látható.

2.7 A magyar Regresszív Képzeleti Szótár a WordStat rendszerében

A WordStat [2] kereskedelmi forgalomban megvásárolható, tartalomelemzésre és szövegbányászatra alkalmas szoftver. A Wordstat a RID összes nyelvi változatának használatát lehetővé teszi. A NooJ programmal létrehozott szótárunkat áthelyeztük erre a platformra. Ez a folyamat viszonylag kevés erőfeszítést igényelt; a 2 évig tartó fejlesztés idejének töredékét tette csak ki. A magyar RID a WordStat honlapján is elérhető, illetve használható. (Azok, akik szeretnék a magyar nyelvű RID-et elemzésre használni, szövegeiket közvetlenül a szerzőknek is elküldhetik.)



2. ábra: A Hang alszótár gráfjának részlete.

3 A reliabilitás vizsgálata

A magyar nyelvű RID reliabilitásának megállapítása az elsődleges és másodlagos gondolkodási folyamatok szintjén történt. A beméréshez Wilson [17] eljárását követtük, aki a RID portugál, latin és német nyelvű fordításainak megbízhatóságát vetette össze, gold standardként az eredeti, angol verziót használva. Elgondolása szerint a reliabilitás mértékét az mutatja meg, hogy mennyire őrzi meg az adott fordítás az elsődleges és másodlagos tartalmak egymáshoz viszonyított arányát. A Biblia 150 zsoltárán végezte el az elemzést. Elemzésében a zsoltárokat külön egységként kezelte. Az angol nyelvű RID-et a Challoner által revideált Douay-Rheims-féle bibliafordításon [1] futtatta le.

Wilson minden egyes zsoltárt az alábbi 3 csoport valamelyikébe sorolta be: 1. Az elsődleges folyamat domináns 2. A másodlagos folyamat domináns 3. Egymáshoz képest egyik szókategória sem domináns. A dominancia azt jelenti, hogy 5%-os szinten szignifikáns eltérés mutatkozik az eloszlások egyenlőségéhez képest.

Ezt követően páronként végzett összehasonlítást: mindig az angol nyelvű zsoltárhoz hasonlítva a másik nyelvű verziót. Ebben az összehasonlításban ötféle konstelláció lehetséges: 1. Helyes azonosítás: az adott zsoltár angol, illetve más nyelvű változatában azonos módon vagy az *Elsődleges* vagy a *Másodlagos gondolkodási folyamat*

kategória domináns. 2. Helyes elutasítás: az angol és a másik nyelvű szövegre is igaz, hogy egyik kategória sem domináns. 3. Helytelen azonosítás: az angol zsoltárban egyik kategória sem domináns, azonban a másik nyelven a szöveg szignifikáns eltérést mutat akár az elsődleges, akár a másodlagos kategória előfordulásának irányában. 4. Helytelen elutasítás: az angol zsoltárban domináns az elsődleges vagy a másodlagos tartalom, azonban a másik nyelvű verzióban nincs domináns kategória. 5. Fordított azonosítás: mind az angol, mind a másik változatnál jelentkezik dominancia, azonban ezek éppen ellentétes irányúak: ha az angolnál az elsődleges kategóriából van több, akkor a másiknál a másodlagosból, vagy fordítva. A fenti öt pártípus abszolút gyakoriságait fordításonként összesítve Wilson a 3. táblázatban található eredményeket kapta.

A magyar nyelvű zsoltárok elemzését a Káldi György által fordított Szentírás [7] szövegének felhasználásával készítettük el. A magyar nyelvű RID-re vonatkozó adatokat a 3. táblázat utolsó oszlopa tartalmazza. A reliabilitás méréséhez Wilson nyomán a következő mutatókat használtuk fel. 1. Pontosság (accuracy): A helyesen (vagyis az angol változattal megegyezően) kategorizált szövegek arányát adja meg az összes szöveg számához viszonyítva. 2. Érzékenység (sensitivity): A helyes azonosítások arányát mutatja azokban az esetekben, amikor az angol szövegben valamelyik kategória domináns. 3. Specifikusság (specificity): A helyes elutasítások arányát mutatja azokban az esetekben, amikor az angol szövegben egyik kategória sem domináns.

3. táblázat: A zsoltárszövegek konstellációinak gyakoriságai
(Wilson [17] adatainak felhasználásával)

A konstelláció típusa	**Portugál**	**Latin**	**Német**	**Magyar**
Helyes azonosítás	27	25	14	24
Helyes elutasítás	58	91	88	92
Helytelen azonosítás	53	20	23	19
Helytelen elutasítás	12	14	25	15
Fordított azonosítás	0	0	0	0

A magyar nyelvű Regresszív Képzeleti Szótár megbízhatóságára vonatkozó eredményeket a 4. táblázat utolsó oszlopa mutatja. Látható, hogy a magyar fordítás két mutató tekintetében ért el az összehasonlított nyelvi változatok között első helyezést (egyik ezek közül holtverseny), egy esetben pedig harmadik helyezést a négy közül. Ez alapján megállapítható, hogy a magyar fordítás megbízhatóan használható az elsődleges és másodlagos gondolkodási folyamatokhoz kapcsolódó tartalmak mérésére.

4. táblázat: A RID-fordítások reliabilitás mutatói
(Wilson [17] adatainak felhasználásával)

A megbízhatóság mutatói	**Portugál**	**Latin**	**Német**	**Magyar**
Pontosság	56,67 %	77,33 %	68 %	77,33 %
Érzékenység	69,23 %	64,1 %	35,9 %	61,54 %
Specifikusság	52,25 %	81,98 %	79,28 %	82,88 %

Hivatkozások

1. Challoner's revised Douay-Rheims Version Old Testament (1609–1610) The Whole Revised and Diligently Compared with the Latin Vulgate by Bishop Richard Challoner (1749-1752). Letöltve: http://www.gutenberg.org/cache/epub/1610/pg1610.html, Letöltés időpontja: 2012. 08. 01.
2. Davi, A., Haughton, D., Nasr, N., Shah, G., Skaletsky, M., Spack, R.: A review of two text-mining packages: SAS TextMining and WordStat. American Statistician, Vol. 59, No. 1 (2005) 89–103. A program elérhetősége: http://provalisresearch.com/products/content-analysis-software
3. ELRAGOZ (Elektronikus magyar ragozási szótár) szoftver. MorphoLogic Kft.
4. Fülöp, É., László, J.: Az elbeszélések érzelmi aspektusának vizsgálata tartalomelemző program segítségével. In: IV. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2006) 296–304
5. Martindale, C.: Romantic Progression: The Psychology of Literary History. Hemisphere, Washington (1975)
6. Martindale, C.: The Clockwork Muse: The Predictability of Artistic Change. Basic Books, New York (1990)
7. Ószövetségi Szentírás a Neovulgáta alapján. Fordította: Káldi György. Szent Jeromos Bibliatársulat, Budapest (1997). Letöltve: http://www.biblia-tarsulat.hu/bibliaszoveg.htm. Letöltés időpontja: 2012. 08. 03.
8. Pennebaker, J. W., Francis M. E., Booth, R. J.: Linguistic Inquiry and Word Count (LIWC): LIWC2001. Lawrence Erlbaum Associates, Mahwah (2001)
9. RID különböző nyelvű moduljainak frissített listája az alábbi webcímen érhető el: . http://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/ regressive-imagery-dictionary-by-colin-martindale-free/
10. Russ, S. W.: Primary process thinking and creativity: Affect and cognition. Creativity Research Journal, Vol. 13 (2001) 27–35
11. Silberztein, M.: Nooj Manual. (2003) Letöltve: http://www.nooj4nlp.net/NooJManual.pdf Letöltés időpontja: 2012. 08. 02.
12. Suler, J. R.: Primary process thinking and creativity. Psychological Bulletin, Vol. 88. (1980) 144–165
13. Váradi T.: The Hungarian National Corpus. In: Proceedings of the 3rd LREC Conference, Las Palmas, Spanyolország (2002) 385–389. Elérhetőség: http://corpus.nytud.hu/mnsz
14. West, A. N., Martindale, C.: Primary process content in paranoid schizophrenic speech. Journal of Genetic Psychology, Vol. 149 (1988) 547–553
15. West, A. N., Martindale, C., Hines, D., Roth, W.: Marijuana-induced primary process content in the TAT. Journal of Personality Assessment, Vol. 47 (1983) 466–467
16. West, A. N., Martindale, C., Sutton-Smith, B.: Age trends in the content of children's spontaneous fantasy narratives. Genetic, Social, and General Psychology Monographs, Vol. 111. (1985) 389–405
17. Wilson, A.: The Regressive Imagery Dictionary: A test of its concurrent validity in English, German, Latin, and Portuguese. Literary and Linguistic Computing, Vol. 26, No. 1 (2011) 125–135