A nyelvi kategória modell kategóriáinak automatikus elemzése angol nyelvű szövegben

Pólya Tibor[1], Kővágó Pál[2], Szász Levente[2]
[1] Magyar Tudományos Akadémia, Természettudományi Kutatóközpont,
Kognitív Idegtudományi és Pszichológiai Intézet
1117 Budapest, Magyar tudósok körútja 2.
polya.tibor@ttk.mta.hu
[2] Pécsi Tudományegyetem, Pszichológiai Intézet
7624 Pécs, Ifjúság útja 6.
kovago.pal@ttk.mta.hu
szasz.levente@ttk.mta.hu

Kivonat: A nyelvi kategória modell a hétköznapi nyelvhasználat szociálpszichológiai kutatásának egyik leggyakrabban használt elemzési eszköze és elmélete. A modell az interperszonális cselekvés leírásában megjelenő absztrakció 5 kategóriáját különbözteti meg. A tanulmányban a modell által meghatározott kategóriák automatikus azonosítására képes eszközt mutatunk be. Az elemzés első lépéseként a szöveg szófaji és szintaktikai elemzését a coreNLP végzi el. A második lépésben az absztrakciós kategóriák felismerését a NooJ szoftverben írt gráfok végzik el. Végül az elemzés harmadik lépése lehetőséget ad arra, hogy a felhasználó különböző csoportokba sorolja a találatokat.

1 A nyelvi kategória modell

A hétköznapi nyelvhasználat szociálpszichológiai kutatásainak egyik leggyakrabban használt elmélete és elemzési eszköze a Semin és Fiedler nevéhez köthető nyelvi kategória modell [8] (angolul Linguistic Category Model, rövidítve: LCM). A nyelvi kategória modell az interperszonális cselekvések leírásának konkrét-absztrakt dimenzió mentén elhelyezhető változatait ragadja meg. A modell szerint az interperszonális cselekvéseket az absztrakció öt szintjén írhatjuk le. A legkonkrétabb fogalmazásmód a leíró cselekvő igével (descriptive action verb, rövidítve: DAV) történő leírás. Például: „Józsi *megüti* Gézát”. A leíró cselekvő igék mindig egy cselekvésre vonatkoznak. A cselekvés kezdete és vége egyértelműen azonosítható. A cselekvésnek van invariáns fizikai jellemzője. Végül önmagában a leíró cselekvő igéknek nincs értékelő jelentése.

Ennél absztraktabb az értelmező cselekvő ige (interpretative action verb, rövidítve: IAV) felhasználásával történő leírás. Például: „Józsi bántja Gézát”. Az értelmező cselekvő igék több azonos cselekvésre vonatkoznak. A cselekvés kezdete és vége szintén egyértelműen azonosítható, de a cselekvésnek nincs egyértelmű invariáns fizikai jellemzője. Az értelmező cselekvő igék esetében a negatív vagy pozitív irányú értelmező mozzanat már tetten érhető.

Az állapotot kifejező cselekvő igék (state action verb, rövidítve: SAV) a IAV-ok közeli rokonai, absztraktságuk szintje az értelmező cselekvő igékkel azonosnak tekinthető. Például: „Józsi *felbőszíti* Gézát." A cselekvés állapot igék egyedi eseményekre vagy események csoportjára vonatkoznak, de a leírás a cselekvés érzelmi következményeire irányítja a figyelmet. A leírt cselekvés ebben az esetben is egyértelmű kezdettel, illetve befejezéssel rendelkezik, de a cselekvés állapot igének önmagában értékelő jelentése van.

Az állapotjelző igék (state verb, rövidítve: SV) hosszan fennálló kognitív vagy érzelmi állapotot írnak le, így kezdetük és befejezésük nem azonosítható. Például: „Józsi *utálja* Gézát".

A legabsztraktabb kategória a cselekvés melléknévvel (ADJ) történő leírása. Például: „Józsi *agresszív*." Ilyenkor a leírás azt implikálja, hogy a cselekvés a célszemély állandó, belső személyes tulajdonsága miatt jött létre.

A nyelvi kategória modellnek két kódolási útmutatója létezik. Az egyiket Klaus Fiedler és munkatársai [7] készítették, a másikat Gün Semin és munkatársai [1] hozták létre. Az automatikus elemző kidolgozása során az elsőként említett leírást követtük.

A szociálpszichológiai vizsgálatok eredményei szerint az interperszonális cselekvés leírásának absztraktsága magyarázó erővel bír például az attribúciós következtetések [8], a sztereotípiák terjedésének módjával [9] és a csoportközi elfogultsággal kapcsolatban. Utóbbit Maass és munkatársai [4,5] tették vizsgálódásuk tárgyává. Kutatásuk eredményeként jött létre a nyelvi csoportközi elfogultság (linguistic intergroup bias, rövidítve: LIB) fogalma. Univerzális emberi jelenség, hogy önértékelésünk egyik fontos összetevőjét azok a csoportok adják, amelyeknek mi is a tagjai vagyunk [11]. A pozitív önértékelésre való törekvés elvéből következően a saját csoport tagjainak viselkedését úgy próbáljuk láttatni, hogy annak pozitív cselekedetei belső okokkal legyenek magyarázhatók, míg a negatív megnyilvánulásait külső, szituációs tényezőknek lehessen tulajdonítani. Ezt nyelvi szinten úgy érjük el, hogy a pozitív cselekedeteket absztraktabban fogalmazzuk meg a negatív cselekedetekhez képest. A külső csoport esetén is hasonló „logika" mentén járunk el, csak éppen fordítva. Azt szeretnénk, hogy a külső csoport rosszabb minőségben tűnjön fel a saját csoportunkhoz képest, ezért annak negatív tetteit absztraktabban, pozitív cselekvéseit pedig konkrétabban fogalmazzuk meg.

Az interperszonális cselekvések leírásában tetten érhető absztraktság szociálpszichológiai vizsgálatainak többsége úgy jár el, hogy az ingeranyagként adott mondatok absztraktságát variálva azonosítja annak hatásait. Hosszabb szövegek absztraktságának kódolása nagy kihívást jelent az empirikus vizsgálatok számára, mivel ehhez akár több száz igét kell kategorizálni. Az általunk kidolgozott elemzési eszköz célja az, hogy megbízhatóan képes legyen nagy terjedelmű szövegben előforduló interperszonális cselekvések absztraktságának megállapítására.

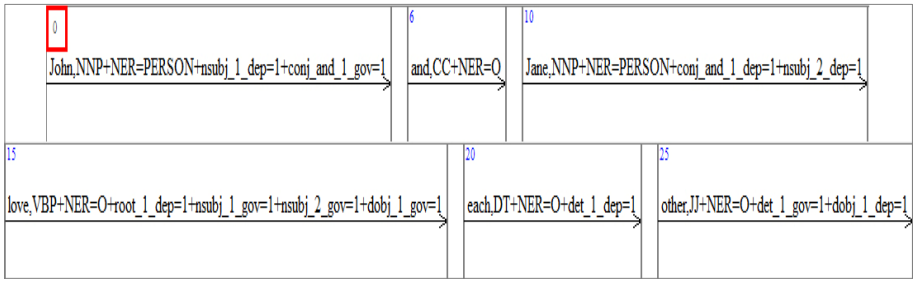
2 A nyelvi kategória modell kategóriáinak automatikus elemzése

Kézenfekvőnek tűnhet, hogy egy szófaji alapon nyugvó kategóriarendszer automatizálása egyszerűen megoldható szótár alapú keresőkkel. Ahhoz azonban, hogy az

elemzés szociálpszichológiai mondanivalóval is bírjon, nem elegendő tudnunk a nyelvi kategória modell kategóriáinak előfordulási gyakoriságát, hanem azt is tudnunk kell, hogy az adott absztrakciós szintű szóalak melyik szereplőhöz tartozik. Annak érdekében, hogy a megtalált ige vagy melléknév összeköthető legyen a cselekvő argumentumával vagy a minősített személlyel, ismernünk kell a szöveg szintaktikai szerkezeti jellemzőit. Egy ilyen elemző használata ráadásul minimalizálja a szavak azonos alakúságából fakadó hibákat is.

Az általunk elkészített angol nyelvű automatizált LCM elemző tehát nem egyszerűen szótár alapon keresi ki és kategorizálja a szövegben előforduló állítmányokat, hanem szintaktikai adatokra támaszkodva hozza összefüggésbe azokat alanyukkal. A melléknévi kategória esetén azt a tárgyat vagy személyt is képes azonosítani az elemző, amelyhez kapcsolódik az adott melléknév.

Az elemzés három lépésben történik. Az első lépésben a szöveg POS taggelését, a tulajdonnevek felismerését és a szöveg szintaktikai elemzését a coreNLP látja el [2, 13]. Az outputként kapott XML formátumú fájlt egy XSLT fordítóval[1] transzformáljuk, hogy a NooJ [10] külön tudja választani a szöveget és annak annotációit. A szöveg annotációi ebben az esetben szavanként tartalmaznak egy POS taget, egy NER értéket, illetve minden egyes függőségi kapcsolatot, amelyet az adott szó a coreNLP által megkapott. A coreNLP szintaktikai elemzője a mondat szerkezetét szópárok egymáshoz való viszonyának jelölésével képezi le. Az alany-állítmányi kapcsolatban például az állítmány ún. „nsubj governor”, az alany pedig „nsubj dependent” annotációt kap. Egy szó több ilyen kategóriát is kaphat, hiszen például egy állítmányhoz több alany is kapcsolódhat. A NooJ nyelvi elemzőben definiálható szabályok sajátosságai miatt ahhoz, hogy össze tudjuk kötni, mely szavak alkotnak egy szintaktikai szópárt, minden szintaktikai pár kap indexként egy számot. Amikor tehát két alanya van egy állítmánynak, az állítmány két nsubj governor szintaktikai kategóriát kap, melyeket 1 és 2 indexszel látunk el az XML fordítás során. Ugyanezt a két indexet fogja megkapni az első és a második alanya az állítmánynak (lásd. 1 ábra).



1. ábra: „John and Jane love each other.” mondat coreNLP általi elemzésének bemenete a NooJ szoftverben

1 Az XML fordításban közreműködött Matuszka Tamás és Rácz Gábor

A második lépésben a coreNLP-ben elemzett szöveget a NooJ-ban elkészített LCM nevű gráffal elemezzük tovább. Ahogy az az 1. ábrán látható a „John" és „love" szavak, illetve a „Jane" és „love" szóalakok nsubj dependency kategóriával kapcsolódnak össze. A példában szereplő „love" ige az állapotjelző ige (SV) LCM kategóriába kerül. Ez az információ egy háttérszótárnak köszönhetően áll rendelkezésre, melyet a fejlesztés korai szakaszában hoztunk létre. A szótár összeállításához a British National Corpus[2] adatbázisát használtuk fel. A leggyakoribb 6318 szótő listájából [3] kigyűjtöttünk az igéket. A listán 1281 ige szerepelt. A legtöbb igének több jelentése is van. A kódolás során az igék leggyakoribb jelentése alapján végeztük el a kategorizálást. Az igék leggyakoribb jelentését a The Longman Dictionary of Contemporary English Online [12] alapján választottuk ki. Az igéket két független kódoló kategorizálta be a nyelvi kategória modell 4 igekategóriájába. A kódolók közötti egyet nem értést egy harmadik kódoló bevonásával oldottuk fel. Tapasztalataink szerint a leggyakoribb igék használata önmagában magas találati arányok elérését teszi lehetővé, azonban a háttér szótárakat könnyedén bővíthetjük a vizsgálatunkban szereplő szövegekben előforduló speciális szavakkal. A melléknevek azonosítására a coreNLP POS taggerét alkalmaztuk.

Az általunk elkészített LCM NooJ gráf kategóriába sorolja a szövegben előforduló azon igéket, amelyek szerepelnek a háttérszótárban. A kategóriába soroláshoz az ige szótövén kívül felhasználjuk a POS tag-et és a szintaktikai elemzés eredményét is. A 2. ábrán látható példánál a mondat egyszerűségéből következően az állítmányi pozícióban levő igét kell megtalálnia a gráfnak, majd egy összekapcsoló gráf párba állítja az azonos indexszel szereplő állítmányokat és alanyokat, illetve jelzős szószerkezeteket egy mondaton belül.

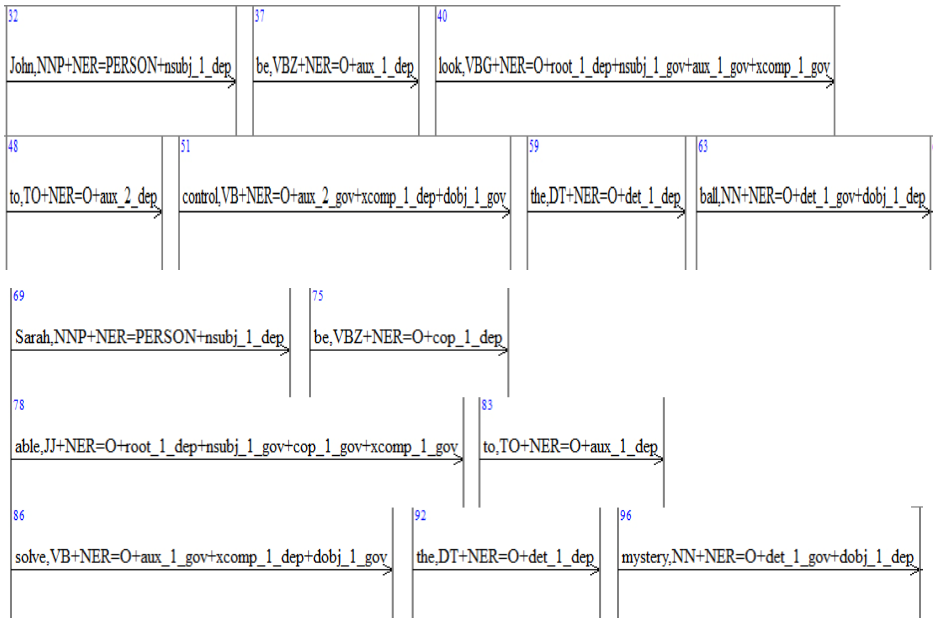
2. ábra: „John and Jane love each other." Az LCM gráf működése egy példán keresztül. A gráf először megtalálja a „love" állítmányt, majd összeköti azt a két alanyával.

Az elemzés harmadik és egyben utolsó lépése egy manuális elemzés a NooJ által megadott konkordancia lista segítségével. A konkordancia listában LCM kategóriába sorolva szerepelnek a találatok, illetve az azokhoz kapcsolódó alanyok vagy minősített entitások. A konkordancia adatok alapján manuálisan döntést hozhat az elemzés végzője arról, hogy tovább szűkíti-e a találatokat. Például elképzelhető, hogy az elemzés végzője csak azokat a találatokat veszi figyelembe, amelyek élő személyek

2 http://www.natcorp.ox.ac.uk

által végrehajtott cselekvéseket írnak le. A nyelvi kategória modellt alkalmazó szociálpszichológusok között nincs egyetértés abban, hogy általában a cselekvés vagy csak az interperszonális cselekvés az, ami elemzendő a szövegben. Szintén indokolt lehet az, hogy az elemzés végzője külön csoportba sorolva veszi figyelembe a saját és a külső csoport tagjainak cselekvésében megjelenő absztrakciót.

Az eddigiekben csak olyan esetekről szóltunk, amikor a megtalálandó ige állítmányi pozícióban van a mondatban. A következőkben két olyan példát mutatunk be, ahol a megtalálandó ige nem kap „nsubj depedency" kategóriát. Ez fakadhat a coreNLP elemzési sajátosságaiból vagy abból, hogy az adott ige valóban nem állítmányi pozíciót foglal el a mondatban. Ilyen esetekben az elemző célja összekapcsolni az igét azzal az entitással, amire vonatkozik, erre láthatunk példát a 3. ábrán.

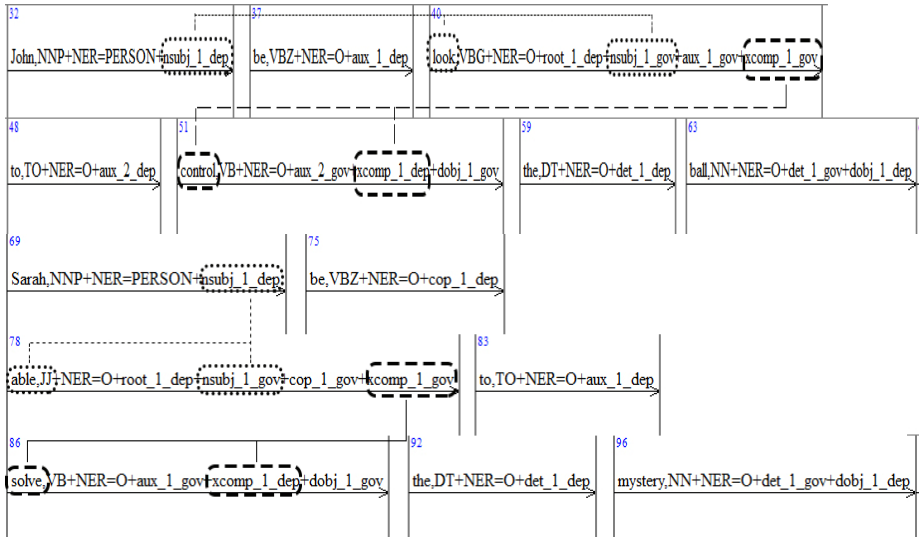


3. ábra: Két példamondat, ahol a megtalálandó ige nem közvetlenül kapcsolódik a cselekvőhöz. „John is looking to control the ball." és "Sarah is able to solve the mystery." mondatok coreNLP általi elemzésének bemenete a NooJ szoftverben

Az első mondatnál az „nsubj" kategóriát a „looking" „linking verb" [3] fogja megkapni, a másodiknál pedig nem egy ige, hanem egy melléknév: „able". Ezekben az esetekben az általunk elkészített LCM NooJ gráf megtalálja a nyelvi kategória modell szempontjából releváns igéket: az első mondat esetében a „control" IAV kategóriájú igét, a második mondat esetében a „solve" IAV kategóriájú igét. A „control" és a

[3] A „linking verb" olyan szó vagy kifejezés, amely egy mondatban az alanyt és a hozzá tartozó állítmányt kapcsolja össze.

„solve” igék az „xcomp dependency” [4] kategóriával kapcsolódnak a coreNLP által megjelölt állítmányokhoz. A gráf ebben az esetben összekapcsolja az „xcomp dependency” kategória segítségével a megtalálandó igéket a mondat állítmányával úgy, hogy az ideiglenesen megadott LCM kategória az állítmány indexét vigye tovább annak érdekében, hogy az alapesetnek vett alany-állítmányi szerkezetnek megfelelő módon összekapcsolható legyen a számunkra fontos ige azzal az entitással, amire vonatkozik (lásd 4. ábra).



4. ábra: „John is looking to control the ball.” és “Sarah is able to solve the mystery.” mondatok elemzése az LCM gráffal. A gráf elsőként a szaggatott vonallal jelölt elemeket találja meg a háttérszótárak segítségével. Második lépésben ezeket köti össze a pontozott vonallal jelölt entitásokkal

Fontos megemlíteni, hogy a gráf jelenlegi verziójában a „looking to control” szerkezet téves találatot is hozni fog, hiszen a „looking” igét meg fogja találni mint állítmányi pozícióban levő DAV kategóriájú igét. Ezt a típusú hibát az LCM elemzőnk több részletben történő futtatásával, illetve komplex kizárási szabályokkal el lehet hárítani. A téves találat elhárításával kapcsolatosan elméleti kérdések is felmerülnek, hiszen bizonyos szerzők [pl. 4,5] az elemzéseikben a „linking verb”-eket is figyelembe veszik mint találatot.

[4] Az xcomp dependency kategória az “open clausal complement” mondatszerkezetet jelöli. Magyarul ehhez a legközelebb azok az esetek állnak, amikor az állítmányt egy főnévi igenévvel rendelkező bővítmény követi.

3 A nyelvi kategória modell elemző reliabilitása

3.1 Szövegminta

Az általunk létrehozott nyelvi kategória modell elemző rendszer reliabalitásának méréséhez futballszurkolók internetes fórum bejegyzéseit használtuk fel. A választás mellett három érv is felhozható. Egyrészt azért döntöttünk sporttal kapcsolatos szövegminta alkalmazása mellett, mert a versengés könnyen kiválthatja a csoportközi elfogultság erőteljes megjelenését és annak nyelvi manifesztációját is. Másrészt az is fontos szempont volt, hogy természetes szöveget (spontán nyelvi megnyilvánulásokat) szerettünk volna elemezni, valódi kihívás elé állítva az elemző rendszerünket. Harmadrészt a fórumokra rendszerint több személy ír véleményt, ami heterogenitást biztosít az elemzett nyelvi mintának.

A Manchester City angol labdarúgó csapat internetes fórumáról [6] a 2013. szeptember és október hónapok legjelentősebb meccseiről szóló kommentárokat válogattuk be az elemzésbe. Ezek között győzelmek és vereségek egyaránt megtalálhatók. Két változó mentén csoportosítottuk a szövegminta mondatait: a saját vagy a külső csoportról (az ellenfél meccsről meccsre változik) mond véleményt, illetve pozitív vagy negatív véleményt fejez ki a kommentet író személy. A fentiek figyelembevételével négy alminta jött létre. Az elemző rendszerünket ezeken futtattuk le. Valamint a kézi kódolást is elvégeztük, melyet „gold standard”-nek tekintettük.

3.2 Eredmények

Az automatikus elemzés megbízhatóságát két módon mértük. A megbízhatóság egyik indikátora az, hogy az elemző rendszer által elvégzett és a kézi kódolás mennyire vezet hasonló kimenetekhez. Az 1. táblázat ad erre vonatkozó információkat. A magas találati és pontossági értékek azt mutatják, hogy az elemző eszközünk megbízhatóan azonosítja a nyelvi kategória modell kategóriáit.

1. Táblázat: A nyelvi kategória modell megbízhatósága: összesített értékek

Nyelvi kategória modell kategóriái	Kézi kódolás eredménye	Találat %	Pontosság %
Leíró cselekvő ige (DAV)	25	80,0	80,0
Értelmező cselekvő ige (IAV)	31	67,7	84,0
Állapotot kifejező cselekvő ige (SAV)	1	0	0
Állapotjelző ige (SV)	9	100	100
Melléknév (ADJ)	85	84,7	90,0
Összes kategória	161	81,9	80,6

A megbízhatóság másik indikátora a 4 alminta összesített absztrakciós mutatójának kiszámítása volt. A szöveg absztrakciós mutatója egy hányados segítségével adható meg [1]. A számláló kiszámításához minden LCM kategóriához egy súlyértéket rendelünk. Ez az érték a leíró cselekvő igék esetében 1, az értelmező cselekvő és az állapotot kifejező cselekvő igék esetében 2, az állapotjelző igék esetében 3, és végül a melléknevek esetében 4. A hányados számlálóját a súlyok és az egyes LCM kategóriák előfordulásának összegzett szorzatai adják. A hányados nevezőjében pedig az LCM kategóriák előfordulásának összege szerepel. A 2. táblázat tartalmazza a 4 almintában a kézi és az automatikus elemzés eredménye alapján kiszámolt absztrakciós mutatókat. Bár a mutató értékei azt jelzik, hogy az elemzett nyelvi mintában nem jelentkezik a csoportközi nyelvi elfogultság [4,5], azonban a gépi kódolás adataiból számított mutatók különbségének iránya azonos a kézi elemzés eredményéből számolt mutatókkal. Vagyis a saját csoport negatív cselekvése esetében számolt mutató értéke magasabb, mint a saját csoport pozitív cselekvése esetében számolt mutató mind a kézi, mind a gépi elemzés esetén. Hasonlóképpen a külső csoport pozitív cselekvése esetében számolt mutató értéke magasabb a külső csoport negatív cselekvésnél számolt értéknél a kézi és a gépi elemzés esetében is. A kézi és gépi elemzés alapján számított absztrakciós mutatók értéke nagyon közel van egymáshoz.

2. Táblázat: A nyelvi kategória modell megbízhatósága: absztrakciós mutatók

Szöveg almintái	**Absztrakciós mutató**	
	Kézi kódolás	**Gépi kódolás**
Saját csoport pozitív értékelése	3,095	3,064
Saját csoport negatív értékelése	3,184	3,197
Külső csoport pozitív értékelése	2,800	2,666
Külső csoport negatív értékelése	2,593	2,450

A megbízhatóság elemzésének eredményei azt mutatják, hogy az általunk létrehozott elemző eljárás megbízhatóan működik. Hangsúlyozzuk azonban, hogy ezeket a méréseket viszonylag kis terjedelmű szövegen végeztük el. A megbízhatóság megállapításához nagyobb terjedelmű szövegek elemzését is szükségesnek tartjuk.

Hivatkozások

1. Coenen, L. H. M., Hedebouw, L., Semin, G. R.: The Linguistic Category Model (LCM) Manual. (2006) Letöltve: http://www.cratylus.org/Text/1111548454250-3815/uploadedFiles/1151434300359-0007.pdf. Letöltés időpontja: 2013. 11. 03.
2. Finkel, J. R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), (2005) 363-370.
3. Kilgarriff, A.: BNC database and word frequency lists. http://www.kilgarriff.co.uk/bnc-readme.html. (2013).

4. Maass, A., Ceccarelli, R., Rudin, S.: Linguistic Intergroup Bias: Evidence for in-group-protective motivation. Journal of Personality and Social Psychology, (1996) Vol. 71(3), 512-526.
5. Maass, A., Salvi, D., Arcuri, L., Semin, G. R.: Language use in intergroup contexts: The linguistic intergroup bias. Journal of Personality and Social Psychology, Vol. 57(6). (1989) 981–993.
6. Manchester City futball csapat szurkolóinak fóruma: http://forums.bluemoon-mcfc.co.uk/
7. Schmid, J., Fiedler, K., Semin, G., Englich, B.: Measuring Implicit Causality: The Linguistic Category Model. (é.n.)
8. Semin, G. R., Fiedler, K.: The cognitive functions of linguistic categories in describing persons social cognition and language. Journal of Personality and Social Psychology, Vol. 54 (4) (1988) 558-568.
9. Semin, G.R.: Agenda 2000 – Communication: Language as an implementational device for cognition, European Journal of Social Psychology, Vol. 30(5), (2000) 595-612.
10. Silberztein, M.: Nooj Manual.: Letöltve: http://www.nooj4nlp.net/NooJManual.pdf (2003) Letöltés időpontja: 2013. 12. 02.
11. Tajfel, H.: Interindividual behaviour and intergroup behaviour. In H. Tajfel (ed), Differentiation between Social Groups. Studies in the social psychology of intergroup relations. Academic Press, London. (1978) 27-60.
12. The Longman Dictionary of Contemporary English Online http://www.ldoceonline.com/
13. Toutanova, K., Manning, C. D.: Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), (2000) 63-70.