Hibrid nyelvtechnológiák

Ács Judit, Borbély Gábor, Makrai Márton, Nemeskey Dávid,
Recski Gábor és Kornai András

Kivonat

Az elmúlt harminc év nyelvészetét a „racionalista" (szabály-alapú, szimbólumkezelő) és az „empirista" (statisztikai alapú, gépi tanulásos) nyelvészeti modellek harca jellemezte. Míg a nyolcvanas években még egyértelműen a racionalista paradigma volt az uralkodó, mára ez, különösen az utolsó néhány év mélytanulásos forradalmának köszönhetően megfordult, és egyértelműen az empirista paradigma a domináns. Az MTA SZTAKI nyelvtechnológiai csoportja elsősorban a hibridizáció kérdéseivel foglalkozik, azzal, hogy miképp találhatjuk meg a diszkrét, szimbolikus struktúrát a folytonos, zajos adatokban, illetve hogyan tudjuk a struktúráról való ismereteinket hatékonyabb algoritmusok építésében kamatoztatni.

1. Bevezetés

Az MTA SZTAKI Nyelvtechnológiai (human language technology, HLT) Kutatócsoportjának előzményei az origo.hu és a Northern Light Technologies (NLT) közti együttműködés időszakára nyúlnak vissza. Ma az Origo csupán egy a számtalan webes portál közül, de 2002-ben, amikor az együttműködés az addig használt AltaVista (AV) keresőtechnológia tarthatatlansága miatt szükségessé vált, az Origo még úgy uralkodott a magyar weben, mint a XIX. században Britannia a habok felett: látogatottsága nagyobb volt, mint az őt követő két legnagyobb portálé együttvéve. Az NLT, melynek akkoriban Kornai volt a tudományos vezetője, 1999-ben nőtt nagyobbra mint az AltaVista (Yahoo), és kettejük versenyében (melyet végül a Google nyert meg) már tetten érhető volt az a szemléletbeli különbség a racionalista és az empirista megközelítések közt, amit pár évvel korábban már igen markánsan jelzett Klavans és Resnik (1996.).

Míg az AV (web yahoo-knak nevezett) szerkesztők százait foglalkoztatta, akik szabály-alapon kézzel sorolták be a weblapokat eleinte néhány tucat, később több ezer, hierarchikusan elrendezett tartalmi kategóriába, addig az NLT statisztikai módszerekkel alakította ki az egyes kategóriák modelljeit, és mivel a besorolás teljesen automatikus volt, nem volt szükség a szerkesztői gárdának a web robbanásszerű növekedését követő bővítésére (mely végső soron a Yahoo/AV vesztét is okozta). A magyar tematikus hierarchia úttörője Ungváry Rudolf (OSZK) volt, az Origóban használt rendszert az ő munkáját továbbfej-

lesztve dolgozta ki Kárpáti András és Halácsy Péter (ma PTE ill. Prezi, akkoriban az Axelero munkatársai). Az NLT az általuk készített katalógus, mint tanulóadat alapján építette fel a saját modelljeit gépi tanulásos módszerekkel (Kornai 2003)[1]. Mint ismeretes, a gépi tanulást (machine learning) máig a címkézett adatokon alapuló ún. felügyelt tanulás (supervised learning) dominálja. A nyers adatokon alapuló felügyeletlen (unsupervised) tanulás nagy erőkkel kutatott terület, ahol komoly eredményekről csak az utóbbi 10 évben beszélhetünk (Erhan et al., 2010.), és az igazi áttörés, a felügyeletlen struktúra-tanulás, még várat magára.

Ebben a cikkben a kutatásoknak a Műegyetemen otthont adó Média Oktató- és Kutató Központtal (MOKK) nem tudunk annak jelentőségéhez mérten foglalkozni, bár kétségkívül ez volt a számítógépes társadalomtudomány első multidiszciplináris műhelye hazánkban, ahol a számítógépes nyelvészet csupán egy volt a digitális szerzői joggal, kulturális termeléssel, a digitális térrel és annak szociológiájával, formális cselekvéselmélettel, az új médiával, peer to peer hálózatokkal, stb. foglalkozó kutatások közül. Reméljük, hogy a nemrég a Preziben Babarczy Eszter, Bodó Balázs, Csigó Péter, György Péter, Halácsy Péter, Kacsuk Zoltán, Szakadát István, Varga Dániel, és Vályi Gábor részvételével megrendezett MOKKtóber találkozó anyagai megteremtik az alapot e műhely történetének és máig érezhető hatásának alaposabb feltárásához.

Az akkori nyelvtechnológiai munkák közül megemlítjük az első magyar szabadon letölthető korpuszt (WebKorpusz), az első párhuzamos magyar-angol korpuszt (hunglish.hu), és a Hun* eszközláncot, melyek az első nyílt forráskódú (open source) magyar nyelvi szoftverek közt voltak. Ebbe az eszközláncba épült be az eredetileg Németh László által külön fejlesztett HunSpell helyesírás-ellenőrző is, mely azóta is a szabad világ vezető helyesírás-ellenőrzője (több mint 100 nyelvhez, megtalálható a Thunderbird, FireFox, LibreOffice sok millió példányában), a Simon Eszter által épített HunNER névelem-felismerő (Simon 2013,Nemeskey 2012, 2013), és még sok más eszköz, melyekről az alábbiakban részletesen lesz szó. A mokk.bme.hu, és a nyelvtechnológiai vonalon ezt továbbvivő hlt.bme.hu máig a nyílt forráskódú nyelvtechnológia egyik vezető képviselője, azzal a fontos különbséggel, hogy az elmúlt másfél-két évtizedben megfordult a széljárás, és az egykor ignorált, majd kinevetett, majd ellenségnek tekintett nyílt forráskódú megközelítés mára uralkodóvá vált.

2. Kétféle szemlélet

Tudományszociológiai szemszögből a racionalista és az empirista kutatási modellek közti különbség lényege a felülről vezérelt (top down) és az alulról kiinduló (bottom up) keresési stratégia. Előbbi klasszikus példája a Manhattan Project, amely egyetlen embernek (a fizikusok elismert vezetőjének, Einsteinnek) az elnökhöz intézett levele alapján indult be: legfelül pár tucat elméleti fizikus, alattuk több száz mérnök és kísérleti fizikus, akik alatt munkások ezrei

[1] A munkacsoport azon cikkei, melyek itt nincsenek hivatkozva, elérhetőek a `hlt.bme.hu` `weblapon`

dolgoztak. A nyelvészetnek is megvolt a maga elismert vezetője, Chomsky, aki nagyon is határozott irányú kutatásokat kezdeményezett.Annak az egyszerű, de előtte kevéssé hangsúlyozott ténynek az alapján, hogy a kisgyermekek viszonylag gyorsan, néhány év alatt lényegében tökéletesen megtanulják anyanyelvüket (és bármely nyelvi környezetbe helyezzük a csecsemőt, az ottani nyelvet képes ilyen szinten megtanulni), arra a következtetésre jutott, hogy ennek a tanulási képességnek kizárólag az lehet a magyarázata, hogy a gyermek fejében a tudásanyag egy nagy része, az univerzális grammatika, már örökletesen ott van.

Bár kezdettől voltak ennek az elméletnek komoly ellenzői, pl. Piaget (Chomskyval való vitájának hiteles összefoglalóját adja Piattelli-Palmarini, Piaget, és Chomsky (1980.)), nyugodtan elmondhatjuk, hogy a fentebb idézett nagy hatású publikációktól kezdve a modern nyelvészeti kutatások fővonalát a XX. században Chomsky jelölte ki (Kornai 2010) és nem kevesek számára máig az ő felfogása szolgál iránytűként. De a Zeitgeist megváltozott, a bölcs vezetők kora lejárt, és ami a legfontosabb: a predikciók nehezen megfoghatónak bizonyultak, specifikus nyelvtani struktúrákat/géneket nem sikerült azonosítani a szótan és mondattan területén. A kudarc annál is fájóbb volt, mert a hangtanban frappáns csecsemőkísérletek sora (összefoglalásukat ld. Werker és Tees (1984.)) nyilvánvalóvá tette, hogy Chomskynak igaza van: az egyes nyelvek hangtanának kisgyermekkori elsajátítása nem magyarázható univerzális fonetika tételezése nélkül.

Ez a megváltozott Zeitgeist tette lehetővé, hogy a terméketlennek bizonyult elméleti megfontolásokból nagyrészt kiábrándult nyelvészek egyre komolyabban vegyék a lentről, a kutatás lövészárkaiból érkező empirikus anyagot. Egyre nagyobb és nagyobb egy- és többnyelvű korpuszt lehetett számítógépes elemzés alá vetni. A bevezetőben már érintettük azokat a korpusz-fejlesztési munkálatokat, melyeket a HLT csoport végzett. Ezek jelentősége nem pusztán abban áll, hogy az addigi nagyon komoly, és szakmailag jól megalapozott korpuszokat mint pl. a Magyar Nemzeti Szövegtár akkori változata (Váradi, 2002.) vagy az elemzett (és ezért természetesen jóval kisebb) Szeged Korpusz (Vincze et al., 2014.) nyíltabbá, jobban elérhetővé tette (ezt inkább a megváltozott Zeitgeistnak mint a Webkorpusz és a Hunglish megjelenésének tudjuk be), hanem abban, hogy elődeiknél lényegesen nagyobbak voltak.

A modern számítógépes elemzés legfontosabb alapanyagát a milliárd szavas (gigaword) korpuszok adják. Azok az elemzési technikák, melyek ma a kutatást uralják, kisebb anyagokon egyszerűen nem működnek jól. A legfontosabb elméleti újítás, mely az utóbbi öt-tíz évben áttörést hozott számos olyan területen, mint a képek és nyelvi leírásuk (caption) közti szemantikai kapcsolat gépi tanulása (Karpathy, Joulin, és Li, 2014.), a szóvektorok (embedding) bevezetése volt. Minden szóhoz egy véges (általában pár száz) dimenziós vektort rendelünk úgy, hogy a hasonló kontextusokban szereplő szavak vektorai egymáshoz hasonlóak (euklideszi térben közeliek) legyenek. Az első áttörést Collobert et al. (2011.) hozták meg, akik egyszerre, ugyanazon vektorok felhasználásával, tudtak javítani több olyan klasszikus feladat addigi legjobb eredményén, mint a szófaj szerinti címkézés (part of speech tagging), a névelem-felismerés (named entity recognition), a sekély mondattani elemzés (tehát a mondatok pszichológiailag releváns

darabokra, pl. főnévi csoportokra bontása, chunking), és a szemantikai szerep felismerése (semantic role labeling). A kulcsmomentum itt az, hogy Colloberték nem egy új feladatot oldottak meg az új reprezentációval, hanem már régről ismert, nehéz, kutatók százai által vizsgált feladatokra (melyek többségével csoportunk is foglalkozott, pl. a HunTag szekvenciális címkéző (Varga 2007) vagy a sekély mondattani elemzés, mely máig aktív témánk, (Recski 2014)) értek el az eddigieknél jobb eredményeket.

A szemantika területén, ahol régen a vezető kutatók, Chomsky és Montague jelölték ki a kutatás fő irányát évtizedekig előre hatóan, ma a kutatók többsége egy olyan jelenséggel foglalkozik, amit egy brünni műegyetemista, Tomas Mikolov fedezett fel: a szóvektorok lineáris struktúrát mutatnak, pl. $\vec{v}(king) - \vec{v}(man) + \vec{v}(woman) \approx \vec{v}(queen)$ (Mikolov, Yih, és Zweig, 2013.). Csoportunk a vektoros szófordítás (lineáris fordítás, (Mikolov, Le, és Sutskever, 2013.)) módszerét alkalmazta Közép-Európai nyelvekre. (Makrai 2013)-ban olyan ritkábban vizsgált lexikai relációk felé általánosítottuk az analógia vektoralgebrai megfogalmazását, mint a jó–rossz (peace–war, pleasure–pain) vagy a fönt-lent (tall–short, rise–fall), (Makrai 2014) pedig oksági párok (pl. sérül–fáj) geometriáját elemezte. Új módszereket vezettünk be többjelentésű beágyazások (multi-sense embeddings) szemantikai felbontóképességének mérésére (Borbély 2016). Ezekben a reprezentációkban egy-egy szóalakhoz több vektor is tartozhat, melyek elvileg a szó különböző jelentéseinek felelnek meg. A gyakorlatban azonban a jelentésvektorok között nem mindig figyelhető meg fogalmi különbség, egy-egy általánosabb vektor több jelentést is lefed, és fölösleges vektorok is lehetnek, melyek a modellnek egy alkalmazásban való hasznosságát ronthatják.

Utólag természetesen megtalálható a szóvektorok használatának elméleti megalapozása: a kontextus nyilvánvalóan fontos, és a gondolat, hogy egy szó jelentését a használati kontextuson keresztül érdemes megragadni kétségkívül jelen van már a nagy brit strukturalista, John Rupert Firth munkáiban is, aki azt írta, „a word is characterized by the company it keeps” (a szavakat a társaságuk jellemezi). Ugyanakkor világosan kell látni, hogy Firth (akinek a prozódiára vonatkozó felfogása is újra életre kelt a modern fonológiában, Goldsmith (1990.)) éppen ahhoz az iskolához tartozik mely ellen Chomsky egész életében harcolt. A nagy tömegű adat viszont minden területen a strukturalistákat, nem pedig az elsősorban szellemes anekdotikus példákra és nyelvi intuícióra alapozott Chomskyánus megközelítést látszik igazolni.

3. Hibrid modellek

A fentiek után talán meglepően hangzik, de korunkban az egész nyelvészet Chomsky programját követi két alapvető tekintetben is. Az egyik a már Chomsky (1965.) által középpontba állított magyarázó adekvátság (explanatory adequacy) elve, mely szerint a nyelvelmélet nem állhat meg a tények leírásánál, hanem arra is magyarázatot kell adnia, hogy a kisgyermek hogyan sajátítja el a nyelvet, a másik az univerzálék (minden nyelvre egyaránt igaz állítások) keresése, melynek Greenberg (1963.) után szintén Chomsky fentebb vázolt programja

adott új lendületet.

A legfontosabb különbség nem a generatív felfogásban, hanem az univerzális meta-elméletet konkrétan realizáló nyelvtanok technikai apparátusában van. A szintaxis területén ez azt jelenti, hogy a környezetfüggetlen mélyszerkezeten és az ezt mozgató fa-átalakításokon alapuló transzformációs grammatika helyét átvette egy másik, szintén a strukturalista korszakból átvett formalizmus, a függőségi grammatika (Tesniére, 1959.). Ebben az elméleti keretben ma már ötven nyelvhez találunk komoly, elemzett fabank (treebank) korpuszokat, jelenleg hetvenet, de számuk egyre nő[2]. Ezek egységesített (univerzális) szófaj-és függőség-tipológián alapulnak, és ezzel nagyban elősegítik a minden emberi nyelvre kiterjedő univerzálé-kutatást. Az empirikus alapok kiterjesztésére mindig is megvolt a szándék: már Greenberg is harminc nyelvvel dolgozott, de nyersanyagául nyelvtani leírások, nem pedig a direkt empirikus adatok szolgáltak. Tekintve, hogy mintegy 6-7000 emberi nyelvről tudunk (bár ezekből gigaword korpuszra és fabankra a digitális nyelvhalál miatt legfeljebb 300-nál számíthatunk, ld. Kornai 2013), az univerzális grammatika kutatása még sok évtizedre fog programot adni a nyelvészetnek.

Az új technikai apparátusra való áttérés egyébként a fonológiában is végbement, ahol a környezetfüggő, szekvenciális szabályrendszereket egy véges automatákkal megfogalmazható elmélet, az Optimalitás Elmélete váltotta fel, (Prince és Smolensky, 1993.; Karttunen, 1998.). A technika megváltozása jelentős átalakulást hozott a szemantikában is, ahol a logikai formán (első- vagy magasabb rendű predikátumkalkuluson) alapuló reprezentációkat egy egyszerűbb, a függőségi fákkal egyenértékű függvény-argumentum szerkezet váltotta fel. Ezt tekinthetjük az ún. generatív szemantikához (Huck és Goldsmith, 1995.) való visszatérésnek, de valójában sokkal régebbre, egészen az első formalizált nyelvtanig, Pāṇini Aṣṭādhyāyī-jáig (i.e. 450 körülre) megy vissza.

Ebben a szellemben dolgoztuk ki 2009 és 2012 között a `4lang` formalizmust (Kornai 2010, 2012, 2015, 2018), mely a természetes nyelvi jelentést fogalmak irányított gráfjaként reprezentálja. Megalkottuk a text_to_4lang szoftvert (Recski 2017), mely nyers angol és magyar szövegekhez automatikusan rendel ilyen reprezentációkat; ezeket sikerrel alkalmaztuk lexikális ontológiák építéséhez (Recski 2016) és a fentebb Mikolov kapcsán már említett analógiás feladatok megoldásában (Recski 2016). Megemlítjük néhány a `4lang` jelentésreprezentációs rendszerhez (Kornai 2013, 2015) kapcsolódó kutatásunkat: az igei szerepek vizsgálata (Makrai 2014), a definiáló szókincs (Makrai 2013, Kornai 2015), és az aktivációterjedés (Nemeskey 2013) kapcsán.

A magyarázó adekvátság tekintetében is ugyanez a folyamat játszódott le: az eszme győzedelmeskedett, de a technikai apparátus gyökeresen szembe megy a Chomsky és Lasnik (1993.) által javasolttal. Kicsi, néhány tucat diszkrét (bináris) paraméter beállításán alapuló döntési fák helyett nagy, sok százezer (gyakran sok millió) folytonos paraméter gradiens-módszerrel való tanulása vált uralkodóvá. Az ilyen sokparaméteres rendszerek tanulása a beszéd- és írásfelismerés terén indult be az ún. Rejtett Markov Modellek (hidden Markov

[2] universaldependencies.org

Model, HMM) felhasználásával: itt kapott először fontos szerepet a valószínűségi nyelvmodellezés (language modeling). Csoportunk mind a hagyományos (szó-n-eseken alapuló, n-gram), mind a mélytanulásban elterjedt rekurrens neurális háló alapú modelleket kutatja. Foglalkozunk a terület mind általános, mind a magyar nyelvre specifikus problémáival is (Nemeskey 2017). A természetes nyelvi mondatok hosszára valószínűségi, generatív modellt alkottunk, ami magyarázni tudja a mondatok empirikusan mérhető hossz-eloszlását.

A magyar nyelv agglutinatív voltából fakadóan a szavak sok felszíni formában lehetnek jelen, ami az angol nyelvben jól működő szóalapú módszereknek komoly kihívást jelent. Vizsgálataink egyik fókusza annak megállapítása, hogy morfológiai eszközök mennyiben tudják ezt a problémát enyhíteni. OTKA-pályázat keretében vizsgáljuk a szavak és morfémák (legkisebb önálló jelentéssel rendelkező nyelvi egységek, pl. tárgyrag) neurális hálózatokkal történő azonosítását. A morfológiai elemzés számos nyelvtechnológiai feladat elengedhetetlen része, amit hagyományosan nyelvészek hosszas munkájával összeállított szabályok segítségével végeznek, azonban ezek a szabályok csak a világ nyelveinek töredékéhez állnak rendelkezésre. Kutatásunk célja olyan módszerek kidolgozása, amik pusztán nyers szövegből képesek ezeket a szabályokat felismerni. Bár ez a rendszer még nincs kész, előmunkálatai közül említést érdemelnek az automatikus szótárépítéssel (Ács 2013, 2014) és ékezet-visszaállítással (Ács 2016) foglalkozó rendszereink.

Foglalkozunk a szóvektorok általánosításaival mátrix- és projektívtér-modellekre. A szokásos szó-vektor alapú beágyazások szisztematikus hibája (Pennington, Socher, és Manning, 2014.), hogy antonima-párok hasonló vektorokkal reprezentálódnak, például *good* $\approx$ *bad*. Ennek egy megoldását kínálja a projektív tér, ahol egy gömbön az antipodális pontok azonosítva vannak. Egy erre épülő célfüggvénnyel sikerült javítanunk a vektor beágyazások által elért eredményt a Simlex999 adaton (Hill, Reichart, és Korhonen, 2014.). A mátrix beágyazások esetében egy szóhoz nem egy vektort, hanem egy mátrixot rendelünk. Ezzel egy nem-kommutatív általánosítását adjuk a szó-vektoroknak, melyek alkalmasak nyelvmodellezésre és speciális véges automaták tanítására is. A hibrid modellek diszkrét komponensei a mátrix-modellek ill. az ezekkel szoros formai kapcsolatban álló véges automaták, melyek tanítása súlyozott nyelveken (Kornai 2013) a szimbolikus és a probabilisztikus modellezésnek az eddigieknél mélyebb hibridizációját készíti elő.

4. Összefoglalás

A racionalista és az empirista megközelítések nem kizárják, hanem támogatják egymást. A modern gépi tanulás alapvető siker-kritériumai messze túlmennek a leíró adekvátságon (descriptive adequacy). A terület egyik legsikeresebb kutatócsoportja, Bengio, Courville, és Vincent (2013.) külön kiemeli, hogy „In good high-level representations, the factors are related to each other through simple, typically linear dependencies" (a jól működő magasszintű reprezentációkban a tényezők egyszerű, tipikusan lineáris kapcsolatban állnak). Ez alól, úgy tűnik a

nyelvtan sem kivétel: a sikeres modellek mögött egyszerű lineáris nyelvtanokat (véges automatákat, véges transzducereket) illetve ezek olyan egyszerű általánosításait találjuk mint a Rejtett Markov Modellek vagy az Eilenberg Gépek (Eilenberg, 1974.). A jövő útja, úgy véljük, az ilyenek automatikus tanulása, és ehhez úgy tűnik nincs semmilyen speciális, az ember általános kognitív képességein túlmutató eszközre szükség.

Azt gondoljuk, hogy a nyelvészeti vizsgálatok a számítógépes társadalomtudományok más területei számára is szolgálhatnak ilyen általános tanulságokkal, hiszen ezekben is az egyik fő cél a mögöttes struktúra feltárása, és ezekben is egyre inkább elérhetővé válik az a hatalmas tömegű adat, aminek alapján e struktúra algoritmikus módszerekkel megragadható.

Hivatkozások

Bengio, Yoshua, Aaron Courville, és Pascal Vincent (2013). "Representation Learning: A Review and New Perspectives". *IEEE Trans. PAMI* 35.8., 1798–1828. old. URL: http://arxiv.org/abs/1206.5538.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, Noam és Howard Lasnik (1993). "Principles and Parameters Theory". *Syntax: An International Handbook of Contemporary Research*. Szerk. J. Jacobs. Évf. 1. Berlin: de Gruyter, old. 505–569.

Collobert, R. et al. (2011). "Natural Language Processing (Almost) from Scratch". *Journal of Machine Learning Research (JMLR)*.

Eilenberg, Samuel (1974). *Automata, Languages, and Machines*. Évf. A. Academic Press.

Erhan, Dumitru et al. (2010). "Why Does Unsupervised Pre-training Help Deep Learning?": *Journal of Machine Learning Research* 11., 625–660. old.

Goldsmith, John A. (1990). *Autosegmental and Metrical Phonology*. Cambridge, MA: Blackwell.

Greenberg, Joseph H. (1963). "Some universals of grammar with particular reference to the order of meaningful elements". *Universals of Human Language*. Szerk. J.H. Greenberg. MIT Press, old. 73–113.

Hill, Felix, Roi Reichart, és Anna Korhonen (2014). "Simlex-999: Evaluating semantic models with (genuine) similarity estimation". *Computational Linguistics* 41.4., 665–695. old.

Huck, Geoffrey J. és John A. Goldsmith (1995). *Ideology and Linguistics Theory: Noam Chomsky and the Deep Structure Debates*. London: Routledge.

Karpathy, Andrej, Armand Joulin, és Fei Fei Li (2014). "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping". *Advances in Neural Information Processing Systems 27*. Szerk. Z. Ghahramani et al. Curran Associates, Inc., old. 1889–1897.

Karttunen, Lauri (1998). "The proper treatment of optimality in computational phonology: plenary talk". *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing*. Association for Computational Linguistics, 1–12. old.

Klavans, Judith L. és Philip Resnik, szerk. (1996). *The Balancing Act – Combining Symbolic and Statistical Approaches to Language*. MIT Press.

Mikolov, Tomas, Quoc V Le, és Ilya Sutskever (2013). "Exploiting similarities among languages for machine translation". arXiv preprint arXiv:1309.4168.

Mikolov, Tomas, Wen-tau Yih, és Geoffrey Zweig (2013). "Linguistic Regularities in Continuous Space Word Representations". *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. Atlanta, Georgia: Association for Computational Linguistics, 746–751. old.

Pennington, Jeffrey, Richard Socher, és Christopher Manning (2014). "Glove: Global Vectors for Word Representation". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 1532–1543. old. DOI: 10.3115/v1/D14-1162. URL: http://www.aclweb.org/anthology/D14-1162.

Piattelli-Palmarini, M., J. Piaget, és N. Chomsky (1980). *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Routledge. ISBN: 0710004389.

Prince, Alan S. és Paul Smolensky (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Piscataway, NJ: Rutgers University Center for Cognitive Science Technical Report 2.

Tesniére, Lucien (1959). *Élements de syntaxe structurale*. Paris: Klincksieck.

Váradi, Tamás (2002). "The Hungarian National Corpus". *Proceedings of the Third International Conference on Language Resources and Evaluation*, 385–389. old.

Vincze, Veronika et al. (2014. máj.). "Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus". English. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Szerk. Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA). ISBN: 978-2-9517408-8-4.

Werker, Janet F. és Richard C. Tees (1984). "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life". *Infant Behavior and Development* 7., 49–63. old.