

# 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 mélytanulások 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 tanulási módszerekkel (Kornai et al., 2003). 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 erővel 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 cselekvésemeléttel, 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 MOKKtőber találkozó anyagai megteremtik az alapot e műhely történetének és máig érezhető hatásának alaposabb feltárásához.

Az akkori nyelvtechnológiai munkák közül megemlíjtjük 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 és Simon, 2012; Nemeskey és Simon, 2013), és még sok más eszköz, melyekről az alábbiakban részletesen lesz szó. A mukk.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 szempontból 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 dolgoztak. A nyelvészetnek is megvolt a maga elismert vezetője, Chomsky,

aki nagyon is határozott irányú kutatásokat kezdeményezett (Chomsky, 1957; Chomsky, 1965; Chomsky és Halle, 1968). Annak az egyszerű, de előtte kevésbé 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, 2010a), 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 frapáns csecsemőkísérletek sora (összefoglalásukat ld. Werker és Tees (1984)) nyilvánvalóvá tette, hogy Chomskynak igaza van: az egyes nyelvek hangtanának kisgyermekkorai 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 (Várad, 2002) vagy az elemzett (és ezért természetesen jóval kisebb) Szeged Korpusz (Vincze et al., 2014) nyíltabbá, jobban elérhetővé tette (ezt inkább a megváltozott *Zeitgeist*nek 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 szövektorok (embedding) bevezetése volt. Minden szóhoz egy véges (tipikusan 300) 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 (chunking), és a szemantikai szerep felismerése (se-

mantic 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 és Simon, 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 struktúráját mutatnak, pl.  $\vec{v}(\textit{king}) - \vec{v}(\textit{man}) + \vec{v}(\textit{woman}) \approx \vec{v}(\textit{queen})$  (Mikolov, Yih, és Zweig, 2013). Csoportunk a vektoros szófordítás (lineáris fordítás, (Mikolov, Le, és Sutskever, 2013)) módszerét alkalmazta Közép-Európai nyelvekre. Makrai, Nemeskey, és Kornai (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 (2014a) pedig oksági párok (pl. sérül–fáj) geometriáját elemezte. Új módszereket vezettünk be többjelentésű beágyazások (multi-sense embeddings) szemantikai felbontóképeségének mérésére (Borbély et al., 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 szavat a társaságuk jellemezi). Ugyanakkor világosan kell látni, hogy Firth (akinek a prozódiaira vonatkozó felfogása is újra életre kelt az autoszegmentális 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 intuícóra 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 (Tesnière, 1959). Ebben az elméleti keretben ma már ötven nyelvhez találunk komoly, elemzett fabank (treebank) korpuszokat, jelenleg hetvenet, de számuk egyre nő<sup>1</sup>. 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 6-700 emberi nyelvről tudunk (bár ezekből gigaword korpuszra és fabankra a digitális nyelvhalál miatt legfeljebb 300-nál számíthatunk, (Kornai, 2013)), az univerzális grammatika kutatása még sok évtizedre fog programot adni a nyelvészetnek.

Az új technikai apparátus (mely egyébként a fonológiában is végbement, ahol a környezetfüggő, szekvenciális szabályrendszereket egy véges transzdukciónal megfogalmazható elmélet, az Optimalitás Elmélete váltotta fel, (Prince és Smolensky, 1993; Karttunen, 1998)) jelentős változást hozott a szemantikában is, ahol a logikai formán (első- vagy magasabb rendű predikátumkalkulácson) 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, 2010b; Kornai, 2012; Kornai et al., 2015; Kornai, in press), mely a természetes nyelvi jelentést fogalmak irányított gráfjaként reprezentálja. Megalkottuk a `text_to_4lang` szoftvert (Recski, to appear), 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 et al., 2016). Megemlítjük néhány a **4lang** jelentésreprezentációs rendszerhez (Kornai és Makrai, 2013; Kornai et al., 2015) kapcsolódó kutatásunkat: az igei szerepek vizsgálata (Makrai, 2014b), a definiáló szókinccs (Makrai, 2013; Kornai et al., 2015), és az aktivációterjedés (Nemeskey et al., 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 írásfelis-

---

<sup>1</sup>universaldependencies.org

merés terén indult be az ún. Rejtett Markov Modellek (hidden Markov Model, HMM) felhasználásával: itt kapott először fontos szerepet a valószínűségi nyelvmodellezés (language modeling). Csoportunk mind a hagyományos (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, Pajkossy, és Kornai, 2013; Ács, 2014) és ékezetvisztaállítással (Ács és Halmi, 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 célfüggvényrel 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 általánosítást 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, Zséder, és Recski, 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

- Ács, Judit (2014.). “Pivot-based multilingual dictionary building using Wiktionary”. *The 9th edition of the Language Resources and Evaluation Conference*.
- Ács, Judit és József Halmi (2016.). “Small footprint diacritic restoration for social media”. *Proceedings of the First Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*.
- Ács, Judit, Katalin Pajkossy, és András Kornai (2013.). “Building basic vocabulary across 40 languages”. *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*. Sofia, Bulgaria: Association for Computational Linguistics, 52–58. old.
- 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>.
- Borbély, Gábor et al. (2016.). “Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation”. *RepEval*.
- Chomsky, Noam (1957.). *Syntactic Structures*. The Hague: Mouton.
- (1965.). *Aspects of the Theory of Syntax*. MIT Press.
- Chomsky, Noam és Morris Halle (1968.). *The Sound Pattern of English*. New York: Harper és Row.
- 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 F 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.
- Kornai, András (2010.a). “Rekurzívak-e a természetes nyelvek?": *Magyar Tudomány* 171.8., 994–1005. old. ISSN: 0025-0325.
- (2010.b). “The algebra of lexical semantics”. *Proceedings of the 11th Mathematics of Language Workshop*. Szerk. Christian Ebert, Gerhard Jäger, és Jens Michaelis. LNAI 6149. Springer, old. 174–199.
- (2012.). “Eliminating ditransitives”. *Revised and Selected Papers from the 15th and 16th Formal Grammar Conferences*. Szerk. Ph. de Groote és M-J Nederhof. LNCS 7395. Springer, old. 243–261.
- (2013.). “Digital language death”. *PloS ONE* 8.10., DOI 10.1371/journal.pone.0077056. old. URL: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077056>.
- (in press.). *Semantics*. Springer Verlag. URL: <http://kornai.com/Drafts/sem.pdf>.
- Kornai, András és Márton Makrai (2013.). “A 4lang fogalmi szótár”. *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Szerk. Attila Tanács és Veronika Vincze, 62–70. old.
- Kornai, András, Attila Zséder, és Gábor Recski (2013.). “Structure Learning in Weighted Languages”. *Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13)*. Sofia, Bulgaria: Association for Computational Linguistics, 72–82. old. URL: <http://www.aclweb.org/anthology/W13-3008>.
- Kornai, András et al. (2003.). “Classifying the Hungarian Web”. *Proceedings of the EACL*. Szerk. Ann Copestake és Jan Hajic, 203–210. old.
- Kornai, András et al. (2015.). “Competence in lexical semantics”. *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (\*SEM 2015)*. Denver, Colorado: Association for Computational Linguistics, 165–175. old.
- Makrai, Márton (2013.). “Fogalmak fontossága a definíciós gráf vizsgálatával [Importance of concepts based on the analysis of the definition graph]”.



- VII. *Alkalmazott Nyelvészeti Doktoranduszkonferencia*. Szerk. Tamás Váradí. MTA Nyelvtudományi Intézet Budapest. ISBN: 978-963-9074-59-0. URL: <http://www.nytud.hu/alknyelvdok13/proceedings13/ANyD7-Makrai-Marton.pdf>.
- (2014.a). “Causality in vectors space language models”. *Spring Wind*. 6. Association of Hungarian PhD és DLA Students (DOSZ), 192–200. old.
- Makrai, Márton (2014.b). “Deep cases in the `4lang` concept lexicon”. *X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014)*. Szerk. Attila Tanács, Viktor Varga, és Veronika Vincze, 50–57 (in Hungarian), 387 (English abstract). old. ISBN: 978-963-306-246-3.
- Makrai, Márton, Dávid Márk Nemeskey, és András Kornai (2013.). “Applicative structure in vector space models”. *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*. Sofia, Bulgaria: ACL, 59–63. old. URL: <http://www.aclweb.org/anthology/W13-3207>.
- Mikolov, Tomas, Quoc V Le, és Ilya Sutskever (2013.). “Exploiting similarities among languages for machine translation”. Xiv 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.
- Nemeskey, Dávid Márk (2017.). “`emMorph` a Hungarian Language Modeling baseline”. *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*. Szeged, 91–102. old. arXiv: 1701.07880 [cs.CL].
- Nemeskey, Dávid Márk és Eszter Simon (2012.). “Automatically generated NE tagged corpora for English and Hungarian”. *Proceedings of the 4th Named Entity Workshop*. Association for Computational Linguistics, 38–46. old.
- (2013.). “Automatikus korpuszpítés tulajdonnév-felismerés céljára”. *IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2013)*. Szerk. Attila Tanács és Veronika Vincze. Szeged, 106–117. old.
- Nemeskey, Dávid et al. (2013.). “Spreading activation in language understanding”. *Proceedings of the 9th International Conference on Computer Science and Information Technologies (CSIT 2013)*. Yerevan, Armenia: Springer, 140–143. old.
- Pennington, Jeffrey, Richard Socher, és Christopher Manning (2014.). “Glove: Global Vectors for Word Representation”. *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.
- 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.
- Recski, Gábor (2014.). “Hungarian Noun Phrase Extraction Using Rule-based and Hybrid Methods”. *Acta Cybernetica* 21., 461–479. old.

- (2016.). “Building Concept Graphs from Monolingual Dictionary Entries”. English. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Szerk. Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.
- (to appear.). “Building concept definitions from explanatory dictionaries”. *International Journal of Lexicography*.
- Recski, Gábor et al. (2016.). “Measuring Semantic Similarity of Words Using Concept Networks”. *Proceedings of the 1st Workshop on Representation Learning for NLP*. Berlin, Germany: Association for Computational Linguistics, 193–200. old.
- Simon, Eszter (2013.). *Approaches to Hungarian Named Entity Recognition*. Ph.D. Thesis, Budapest University of Technology and Economics.
- Tesnière, Lucien (1959.). *Éléments 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.
- Varga, Dániel és Eszter Simon (2007.). “Hungarian Named Entity Recognition with a Maximum Entropy Approach”. *Acta Cybern.* 18.2., 293–301. old.
- Vincze, Veronika et al. (2014.). “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.