Eötvös Loránd Tudományegyetem
Bölcsészettudományi Kar
Tézisfüzet

MAKRAI MÁRTON

# SZIMBOLIKUS ÉS ELOSZLÁSOS SZÓREPREZENTÁCIÓK

## FEJEZETEK LEXIKAI RELÁCIÓKRÓL ÉS NYELVKÖZI MÓDSZEREKRŐL

DOI: 10.15476/ELTE.2023.350

Nyelvtudományi Doktori Iskola
Vezetője: Tolcsvai Nagy Gábor, az MTA rendes tagja

Elméleti Nyelvészeti Doktori Program
Vezetője: Bánréti Zoltán CSc

A bizottság tagjai:
Törkenczy Miklós DSc, elnök
Novák Attila PhD, opponens
Szécsényi Tibor PhD, opponens
Gyuris Beáta PhD
Vincze Veronika PhD

Témavezető:
Kornai András DSc

Budapest, 2023. december

# 1 előzmények és célkitűzések

A disszertáció a számítógépes lexikai szemantika területére tartozik, ami meglehetősen interdiszciplináris a számítástudomány, a nyelvészet, a matematika, a pszichológia és a filozófia határán. A jelentésreprezentációk két fő csoportja a szimbolikus és az eloszlásos; ezeket a disszertáció 2. illetve 3. fejezete mutat be. Az előbbiek fő példái a különféle szemantikus hálók, melyek csomópontjai fogalmakat, a köztük futó élek pedig állandó (a szavak jelentéséhez tartozó) vagy a feldolgozandó szövegben fennálló viszonyokat ábrázolnak, a szakértők explicit tudásának közvetlen formalizálásán alapszanak. Ezek közül a régebbieket (2.2.) az alapötletek miatt tartom fontosnak, az újabbakat (2.4.) pedig potenciális vagy konkurens számítógépes nyelvészeti erőforrásként is. A disszertáció elméleti keretét a `4lang` jelentésábrázolási elmélet adja, ami a már említett paradigmákon kívül a kognitív szemantikára (2.3.) is épít.

Az eloszlásos jelentésreprezentációk (a sekély vagy mély neurális hálók, a szóbeágyazások és a mély nyelvmodellek) tanításához nem szükséges nyelvészeti tudás, pusztán annotálatlan szöveg. A gyakorlatban hasznosabbak mint a szimbolikusak, viszont nehéz őket interpretálni. Fontos előzményeik a mátrixfelbontáson alapuló szóábrázolások, ahol a felbontásra kerülő mátrixok elemei szavak közötti asszociációkat (4.1.1.) vagy együttelőfordulásokat (4.1.2.–4.1.6.) képviselnek. A neurális nyelvmodellezés történetét nagy képben, de viszonylag teljesen leírom (4.2 és 4.3) 2003 környékétől (amikor elkezdtek valóra válni a konnekcionista álmok, melyek szerint az ember implicit nyelvi tudása mesterséges neurális hálókkal hatékonyan modellezhető) a figyelemalapú mély nyelvmodellekig. A nagyon mély előtanított nyelvmodellek, mint a GPT, csak említés szintjén szerepelnek a disszertációban.

A `4lang` jelentésábrázolási keretet Kornai András intézetközi (SZTAKI, BME, Nyelvtudományi Intézet) projektjében fejlesztettük. A több könyvben és cikkben felvázolt és részben implementált rendszerre a legegyszerűbben úgy gondolhatunk, mint egy gráfra, melynek csomópontjai fogalmak, és ezeket a fogalmakat háromféle él köti össze: két aszimmetrikus szereplő (pl. alany és tárgy, bir-

tokos és birtok, figura és háttér) esetén ún. 1-es és 2-es (a fej két bővítményének megfelelően), egyetlen szereplő (pl. IS-A, egyszereplős esemény, tulajdonság) esetén pedig ún. 0-s nyilat használunk. A csomópontoknak megfelelő fogalmak meglehetősen absztraktak: nyelvfüggetlenek, nincs szófajuk (például a *haszon* és a *használ* ugyanaz a csomópont), és szóegyértelműsítés is csak nyilvánvaló homonímia esetén történik. Ez a monoszémia elv. A kutatócsoport tagjainak módszertana a lépésenkénti lexikai felbontás (Dik 1978) elvét alkalmazza digitális szótárakra (digitalizált hagyományos szótárakra vagy a *Wiktionary*-re). A teljes szókincs elemeit a definíciójukkal reprezentáljuk, majd a definíciókban szereplő szavakat is a definíciójukkal helyettesítjük, és így tovább. Amikor már nem lehet több cserét végezni (mert minden szó definíciója visszahozna egy korábban eliminált szót), akkor megkaptuk a definiáló szókincset. Nagyrészt az én feladatom volt egy ilyen definiáló szókincs elemeit manuális munkával definiálni a kognitív szemantika és a `4lang` saját elveit követve. Az így létrejött definíciók szolgáltak a disszertáció számos eredményének bemenetéül.

A disszertáció fő fejezetei szavak közötti relációk szimbolikus és eloszlásos modellekben való megjelenésével foglalkoznak. Az 5. és a 6. a régensek és bővítményeik közötti relációkat vizsgálja a `4lang`-ben illetve egy eloszlásos modellben. A 7. fejezet azt vizsgálja, hogy a szemantikus hálókból vagy pl. Jackendoff (1983) kognitív szerkezeteiből ismerős lexikai relációk illetve analógiás és nyelvek közötti kapcsolatok hogyan jelennek meg statikus szóbeágyazásokban. Az utolsó fejezet szójelentés-klaszterezésre javasol egy nyelvközi módszert: azt veszi górcső alá, hogy egy gépi tanulási modellben többértelműnek tűnő szó valóban az-e.

## 2 MÓDSZEREK

A 3.3. szakasz a klasszikus webkeresés *PageRank* nevű módszerét alkalmazza arra, hogy megmérje, hogy a `4lang`-definíciókban használt szimbólumok (lexikai relációktól az 5. fejezetben főssze-

repet játszó mélyeseteken át egyszerű fogalmakig, mint a *place*) milyen fontosak a lépésenkénti lexikai felbontásban.

A `4lang` ún. mélyeseteiről szóló 5. fejezet módszertana talán a számítógépes lexikográfiáéhoz esik a legközelebb: az volt a feladat, hogy egy adatvezérelt módon megválasztott szókincs definícióit kognitív szemantikai megfontolások mentén manuális munkával hozzuk létre. Ezt leginkább én végeztem.

Az igék, alanyok és tárgyak közötti asszociációkat vizsgáló 6. fejezet motivációja továbbra is számítógépes lexikográfiai, de ez viszonylag komoly adattudományos és valószínűségszámítási apparátust is használ, amennyiben a normalizált pozitív pontonkénti kölcsönös információt (PPMI) általánosítja a többváltozós esetre, és ennek hasznosságát egy tenzorfelbontási kísérletben igazolja.

A hipernimakinyeréssel foglalkozó szakasz (7.1.) két módszere, a (formális) fogalomhálók (FCA) és a ritka (túlteljes) szóreprezentációk két eléggé különböző területhez tartoznak (matematikai nyelvészet illetve általános gépi tanulás, reprezentációtanulás), de egymás diszkrét-kombinatorikus illetve statisztikai (relaxált) párjának tekinthetők.

Az okságot vizsgáló 7.3. szakasz egy kétdimenziós vizualizációból nyert sejtést vizsgál a statikus szóvektorok eredeti terében, és az analógiás párok paralelogrammaként való ábrázolásának – amit vektoreltolásnak is hívnak – mintájára arra jut, hogy az okozatok jelentésvektora az ok vektorából és egy közös oksági elemnek megfelelő vektoriból áll össze. Ez tehát egy nagyon absztrakt eredmény a szavak látens teréről.

A 7.4-es szakasz a 2013–2019-es időszak népszerű módszereit, a már említett vektoreltolást és a lineáris szófordítást használja magyarra és más közepes erőforrású európai nyelvekre. Módszertani újdonság volt (Makrai 2015), hogy a vektoreltolást és a lineáris fordítást is kiterjesztettem a word2vec modellről a GloVe modellre.

A 7.5-ös szakaszban a szófordítás-kinyerésnek a korai gépi tanulásban szokásos módszerét, a háromszögelést ötvöztem a disszertáció előző szakaszában (és a következő fejezetben) is alkalmazott lineáris fordítással.

Az utolsó fejezet többjelentésű szóbeágyazásokat értékeli ki a szavak többértelműségének felismerésében való pontosságukat *(precision)* mérve a lineáris fordítást eszközével.

## 3 TÉZISEK

1. Egy webkeresésből vett módszert, a *PageRank*-et javasoltam és alkalmaztam sikerrel annak mérésére, hogy a szemantikus háló egyes csomópontjai milyen fontos szerepet játszanak abban a rekurzív folyamatban, ahogy a szavakat egymás segítségével definiáljuk. Makrai (2013)-ban a `4lang` akkori változatán mutattam be a módszert. A disszertációban megismételtem a vizsgálatot a Makrai (2014b)-es definíciókkal.

2. Mélyesetekkel láttam el a `4lang` alapszókincsének kézzel írott, képletszerű definícióit, ezáltal egy mélyesetkészletet alakítottam ki. (Makrai 2014b).

3. Alany-ige-tárgy együttelőfordulások modellezéséhez a PPMI többféle súlyozott változatát általánosítottam a magasabb rendű (>2) esetre. Angol alany-ige-tárgy hármasok harmadrendű kölcsönhatásait tenzorbontással modellezve megmutattam (Makrai 2022), hogy

   a) ezek a súlyozott magasabb rendű PPMI-változatok jobbak, mint a baseline-ként használt log gyakoriság, egyszerű PPMI és log Dice.

   b) az üres tárgyaknak a kitöltöttekkel egységesen való kezelése javára válik a szóábrázolásnak.

   c) a nemnegatív felbontással (akár kanonikus poliadikus, akár Tucker) kapott látens dimenziók szemantikailag értelmesek.

   d) az egyes főnevek alanyként és tárgyként való beágyazása közötti különbség intuitíve az ágenciához köthető.

4. Lexikai relációk

a) Szerzőtársammal[1] sikerrel nyertünk ki hipernimákat ritka szóreprezentációkkal. Ezzel megnyertük a *SemEval-2018 Task 9* három részfeladatát, Berend, Makrai és Földiák (2018).

b) Szerzőtársaimmal[2] a definíciós gráfból készült szóbeágyazást validáltunk az alapján, hogy az antonímia mely lehetséges részrelációi jelennek meg benne.

c) Megmutattam, hogy a Senna nevű klasszikus szóbeágyazásban a különféle ok-okozat párokat (pl. *bánt–sérül*) összekötő egyenesek egy közös „oksági középpont" közelében futnak, vagyis a hatások (*sérül*) jelentése a megfelelő ok (*bánt*) jelentéséből és egy állandó oksági elemből áll össze (Makrai 2014a).

d) Létrehoztam a gazdag morfológiájú magyar nyelv analógiás viszonyainak benchmarkját, és ezzel teszteltem a magyar statikus szóbeágyazásokat, megmutatva, hogy a morfológiai kapcsolatok hasonlóan megjelennek, mint a jobb erőforrásokkal rendelkező nyelvekben (Makrai 2015).

5. Lineáris fordítás

a) Kiterjesztettem a lineáris fordítás módszerét közepes erőforrású európai nyelvekre és a GloVe modellre (Makrai 2015).

b) A szófordítás-kinyerésre régi módszerét, a háromszögelést (pivot-módszert) a lineáris szófordítással hibridizáltam: egy háromszögelt fordítási szópárlista elemeit a lineáris fordításból származó jóságmértékkel láttam el. Közreadtam az akkori legnagyobb német-magyar szótárat (szópárlistát) (Makrai 2016).

1 A hozzájárulások aránya Berend:Makrai = 2:1. Földiák Pétertől az FCA-s ötlet jött.

2 Én soroltam részrelációkba az antonim párokat. A statisztikai tesztek elvégzése Nemeskey Dávidddal egyenlő hozzájárulás.

6. Szerzőtársaimmal[3] javaslatot tettünk egy módszerre, ami kiértékeli az többjelentésű szóbeágyazásokat (MSE-ket) mint a homonímia detektorait (Borbély és mások 2016). A módszer a beágyazás-alapú szótárindukció kontextusába illeszkedik. Megmutattam, hogy ennek a paradigmának a módszerei közül melyek hasznosak, ha a forrásnyelv modellje többjelentésű. Magára a kiértékelésre két mérőszámot javasoltam: az egyik a fordítás minőségét általánosságban méri, a másik pedig a kétértelmű szavak felismerésének pontosságát. A két SOTA MSE modellt összehasonlítva megmutattam, hogy a két mérőszám között fennáll a várt csereviszony: minél specifikusabbak a vektorok, annál könnyebb fordítani, de ha túl specifikusak, akkor a fordítások egybeeshetnek (Makrai és Lipp 2018).

A fenti tételszerű téziseken túl az értekezés témáihoz kapcsolódó publikációm jelent meg többek közt többértelműségről (Makrai 2007), és a `4lang` szemantikus hálójában való aktivációterjedésről (Nemeskey és mások 2013). A `4lang`-et magyarul Kornai és Makrai (2013), angolul Kornai és mások (2015), a kutatócsoportnak a hibrid adatvezérelt számítógépes nyelvészetről való felfogását Ács és mások (2018) mutatták be. A hatványeloszlás hátterét egy másik kontextusban mutatta be Makrai és Sass (2018).

3 Az első cikk többnyelvű része Borbély Gáborral egyenlő hozzájárulás, Kornai András témavezetett minket.

## A JELÖLT KÖZLEMÉNYEI


Ács, J., G. Borbély, M. Makrai, D. Nemeskey, G. Recski és A. Kornai. 2018. "Hibrid nyelvtechnológiák". *Magyar Tudomány* 6. (Hivatkozási oldal 7).

Berend, Gábor, Márton Makrai és Péter Földiák. 2018. június. "300-sparsans at SemEval-2018 Task 9: Hypernymy as interaction of sparse attributes". *Proceedings of The 12th International Workshop on Semantic Evaluation,* 928–934. oldal. New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1152. (Hivatkozási oldal 6).

Borbély, Gábor, Márton Makrai, Dávid Márk Nemeskey és András Kornai. 2016. "Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation". *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP,* 83–89. oldal. Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2515. (Hivatkozási oldal 7).

Kornai, András, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy és Gábor Recski. 2015. "Competence in lexical semantics". *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics,* 165–175. oldal. Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.18653/v1/S15-1019. (Hivatkozási oldal 7).

Kornai, András, és Márton Makrai. 2013. "A 4lang fogalmi szótár". *IX. Magyar Számitógépes Nyelvészeti Konferencia,* szerkesztette Attila Tanács és Veronika Vincze, 62–70. oldal. (Hivatkozási oldal 7).

Makrai, Márton. 2007. *Többértelműségek magyar mondatok számítógépes elemzésében – a* meg *szó szófajának vizsgálata gyakoriságokkal.* https://hlt.bme.hu/en/publ/makrai07. Témalabordolgozat. (Hivatkozási oldal 7).

———. 2013. "Fogalmak fontossága a definíciós gráf vizsgálatával [Importance of concepts based on the analysis of the definition graph". *VII. Alkalmazott Nyelvészeti Doktoranduszkonferencia,* szerkesztette Tamás Váradi. MTA Nyelvtudományi Intézet Budapest. ISBN: 978-963-9074-59-0. http://www.nytud.hu/alknyelvdok13/proceedings13/ANyD7-Makrai-Marton.pdf. (Hivatkozási oldal 5).

———. 2014a. "Causality in vectors space language models". *Spring Wind,* 6:192–200. oldal. Association of Hungarian PhD / DLA Students (DOSZ). (Hivatkozási oldal 6).

———. 2014b. "Deep cases in the `4lang` concept lexicon". *X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014),* szerkesztette Attila Tanács, Viktor Varga és Veronika Vincze, 50–57 (in Hungarian)., 387 (English abstract). ISBN: 978-963-306-246-3. (Hivatkozási oldal 5).

———. 2015. "Comparison of distributed language models on medium-resourced languages". *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015),* szerkesztette Attila Tanács, Viktor Varga és Veronika Vincze, 22–33. oldal. Szegedi Tudományegyetem Informatikai Tanszékcsoport. ISBN: 978-963-306-359-0. (Hivatkozási oldalak 4, 6).

———. 2016. "Filtering Wiktionary triangles by linear mapping between distributed models". *LREC.* (Hivatkozási oldal 6).

———. 2022. "Three-order normalized PMI and other lessons in tensor analysis of verbal selectional preferences". *XVIII. Magyar Számítógépes Nyelvészeti Konferencia,* szerkesztette Gábor Berend, Gábor Gosztolya és Veronika Vincze, 105–120. oldal. Szegedi Tudományegyetem TTIK, Informatikai Intézet. ISBN: 978-963-306-848-9. (Hivatkozási oldal 5).

Makrai, Márton, és Veronika Lipp. 2018. “Do multi-sense word embeddings learn more senses?” *K + K = 120 Workshop Dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays.* (Hivatkozási oldal 7).

Makrai, Márton, és Bálint Sass. 2018. “A szöveg mint skálafüggetlen hálózat”. *XIV. Magyar Számítógépes Nyelvészeti Konferencia.* (Hivatkozási oldal 7).

Nemeskey, Dávid, Gábor Recski, Márton Makrai, Attila Zséder és András Kornai. 2013. “Spreading activation in language understanding”. *Proceedings of the 9th International Conference on Computer Science and Information Technologies (CSIT 2013),* 140–143. oldal. Yerevan, Armenia: Springer. https://hlt.bme.hu/media/pdf/nemeskey_2013.pdf. (Hivatkozási oldal 7).