FILOLÓGIA.HU • XVI. évfolyam (2025) 1–2. szám 93–97. • 10.59648/filologia.2025.1-2.7

Tóth András

Recenzió Tóth Ágoston *The Company that Words Keep: Distributional Semantics* című könyvéről

Tóth Ágoston 2014: The Company that Words Keep: Distributional Semantics. Debrecen: Debrecen University Press

Bizonyára sokunkat érdekel, hogy a ChatGPT, a Microsoft Copilot, a Google Gemini és a hasonló csevegőmegoldások hogyan működnek. A nyelvtechnológiában ezeket *nagy nyelvi modelleknek* hívják, amelyek a *disztribúciós szemantika* kutatási hátterével létrehozott nyelvtechnológiai eszközök.

A disztribúciós szemantika automatikusan osztályozza a nyelv építőegységeit jelentésük szerint. Az építőegységek lehetnek szavak, szóalakok, szótöredékek. Az osztályozásnak az alapja az, hogy ezek az egységek milyen más egységek környezetében fordulnak elő. A nyelvészetben a 20. századi strukturalista gondolatok tekinthetők szellemi előzménynek, Firth például 1957-ben már leírta, hogy „you shall know a word by the company it keeps" (Firth 1957; 'madarat tolláról, szót szomszédjáról', fordítás: Tóth 2022).

Tóth Ágoston itt ismertetett könyve (Tóth 2014a) a disztribúciós szemantika korai állapotát mutatja be nemzetközi kutatási eredményekre alapozva, azokat saját kísérletekből származó eredményekkel kiegészítve. A könyvben leírt modellek ugyan még nem neurális nyelvi modellek, viszont továbbra is érdekesek, két fő okból. Egyrészt látható, hogy honnan indult a disztribúciós szemantika, milyen célokkal, módszerekkel, kihívásokkal és kezdeti eredményekkel. Másrészt ebben a fejlődési fázisban tisztábban látjuk az összefüggéseket az eredeti nyelvi adat és a szemantikai eredmény között. A mai neurális nyelvi modellek feketedoboz-jellege miatt ilyen vizsgálatokat csak ritkán, speciális esetekben végeznek, sokkal inkább a modellek alkalmazása, tesztelése, összevetése a cél. Ilyen kutatást, modern nyelvi modellekkel, az itt recenzált szerző is publikált angol (Tóth–Abdelzaher 2023) és magyar (Furkó–Tóth 2023) nyelven, előbbit lexikográfiai, utóbbit társalgáskutatási területen. A magyar vonatkozású eredményekért érdemes figyelemmel kísérni a Magyar Számítógépes Nyelvészeti Konferencia (MSzNy) konferenciaköteteit.[1]

Tóth (2014a) bevezető fejezete az angol *life* ('élet') szót példaként felhasználva bemutatja, hogy a nyelvtechnológia hogyan tudja három különböző módon a szó jelentését rögzíteni. Látjuk a WordNet adatbázisból vett adatokat és a szó által aktivált

1 https://rgai.inf.u-szeged.hu/mszny

FrameNet esetkeretet. A WordNet és a FrameNet hagyományosnak nevezhető számítógépes nyelvészeti adatbázisok, amelyek a szavak jelentésére összpontosítanak. Ezután, harmadik megközelítésként áttérünk a disztribúciós szemantikára. Olvashatunk a disztribúciós hipotézisről (Sahlgren 2008; Lenci 2008), amely szerint a hasonló jelentésű szavak hasonló szövegkörnyezetben fordulnak elő. Ez az automatizálás szempontjából egy fontos hipotézis, és arra utal, hogy akár címkézetlen, „nyers" korpuszok feldolgozásával is jellemezhető a szavak jelentése.

A disztribúciós hipotézisen alapulnak mind a korai kutatások, mind a legmodernebb eszközök, például a ChatGPT is. A létrehozott szótulajdonság-vektorok közvetlenül a szavak környezetét, disztribúcióját írják le. E vektorok összehasonlításával, csoportosításával szemantikai információ állítható elő.

A bevezető fejezet erre konkrét példát mutat be. Egy 50 millió szavas Wikipédia-részletből kinyert tulajdonságvektorokat hasonlít össze, és felsorolja, hogy mely szavak hasonlítanak leginkább a *life* szóhoz (Tóth 2014a: 14). Ilyen például a *lives* ('életek'), a *time* ('idő'), az *experience* ('tapasztalat'), a *love* ('szerelem'), a *people* ('emberek'), a *nature* ('természet'), az *activities* ('tevékenységek'), a *community* ('közösség') és a *relationship* ('kapcsolat', 'rokonság'). Az itt felsorolt szavak relevánsnak is tűnnek, ha az *élet* szó jelentését keressük. Vannak a listán azonban olyan szavak is, amelyek disztribúciós hasonlósága nehezen magyarázható jelentésük alapján.

A könyv első fejezete röviden összefoglalja a szójelentés megragadásának lehetőségeit számítógépes nyelvész szemmel, de először is a *szó* terminus meghatározásának nehézségeiről olvashatunk. A szerző itt felhívja a figyelmet arra, hogy ennek az alapvető fogalomnak a leírása elméleti oldalról sem könnyű, ha hasznos és pontos meghatározást keresünk. Már főleg nyelvtechnológiai szempont a többszavas kifejezések (nevek, idiómák, bizonyos szóösszetételek stb.) kezelésének említése, valamint a WordNet és a FrameNet projektek ismertetése. Ez a fejezet készíti elő a disztribúciós szemantika bemutatását.

A második fejezet a disztribúciós szemantika 2014 előtti állapotának áttekintése. Ismerteti, hogy a szavak együttes előfordulásának megszámlálásával hogyan hoztak létre olyan tulajdonságvektorokat, amelyek jelentéstani információt hordoznak. Magyar példákat is bemutat, például a *magas* és *nagy* szavak hasonlóbb környezetben fordulnak elő, mint a *magas* és *teljes* szavak. Megjegyzi, hogy az ellentétek disztribúciója is hasonló, azaz hasonló környezetekben jelennek meg például a *jó* és *rossz* szavak, valamint a hiponímiát-hiperonímiát is megjeleníti a disztribúció összehasonlítása. Például az *alma és gyümölcs* szavak esetében: nagyjából ugyanazokkal az igékkel tudjuk használni ezt a két főnevet, valamint aki megeszi őket, vagy leszüreteli, feldolgozza, az is ugyanaz lehet.

A módszertani kérdések között felmerül a szükséges korpusz méretének kérdése, a célszó környékén vizsgálandó szövegkörnyezet megfelelő méretezése és egyéb paraméterek, amelyek befolyásolják a létrehozott tulajdonságvektorok hasznosságát. A könyv bemutat több erre irányuló kísérletet, amelyek a modellezési paraméterek hatását vizsgálják számítógépes nyelvészeti feladatok megoldásában. Magyar szövegeken mért saját adatokat is kapunk a szerzőtől a könyvben, amelyeket máshol magyarul is elérhetővé tett (Tóth 2013). A könyvfejezet bevezeti és vizsgálja

a tulajdonságvektorok összehasonlításával megvalósítható disztribúciós kompatibilitási relációt *(distributional compatibility relation).*

A harmadik fejezet azt elemzi, hogy a disztribúciós szemantikával nyert adatok korrelálnak-e pszicholingvisztikai adatokkal. A szerző leírja, hogy már az 1960-as években végeztek olyan kísérletet, ahol szópárok hasonlóságát mérték. A kísérlet során a résztvevőket nyilatkoztatták a hasonlóság mértékéről, más résztvevőkkel pedig olyan mondatokat írattak, amelyekben az adott szavak szerepelnek, és e mondatok szavait hasonlították össze. A kísérlet eredményeként a kétféleképpen megállapított hasonlóságok korreláltak (Rubenstein–Goodenough 1965). Ezt a korai eredményt többen felhasználták már saját kutatásukban, az itt bemutatott könyv szerzője is közli saját, magyar nyelvű kísérletének eredményét. A magyar kísérletben is azt látjuk, hogy a szubjektív hasonlósági értékek korrelálnak a korpuszból vett, disztribúciós szemantikai módszerekkel automatikusan számított hasonlósági adatokkal. A szerző kísérletének leírása magyar nyelven is hozzáférhető (Tóth 2014b). A pszicholingvisztikai kapcsolatot több szerző előfeszítéses *(priming)* kísérleteinek bemutatásával is alátámasztja a könyv. Itt saját, magyar adatokat nem közöl a szerző.

A negyedik fejezet megoldandó feladatokat, további kutatási lehetőségeket sorol fel. Az első ilyen a szó szintjéről a nagyobb egységek irányába történő továbblépés lehetőségeit vizsgálja az erre vonatkozó publikációk áttekintésével. Ezek közt van olyan, amely a lexikális szinten maradva a többszavas kifejezésekkel foglalkozik, és olyan is, amely a mondatok szintjét igyekszik elérni. A második bemutatott terület a jelentéssel kapcsolatos ismert problémák kezelésének lehetőségeivel foglalkozik; itt a többértelműség áll a középpontban. Egy alfejezet a disztribúciós szemantika és a kognitív szemantika, azon belül a FrameNet összekapcsolásának lehetőségét veti fel, meglévő szakirodalmi eredményekre támaszkodva. Ez utóbbi kutatási lehetőséget a szerző a későbbiekben meg is vizsgálta, eredményeit Tóth (2018) közli.

A könyv megjelenése óta eltelt 10 évben a disztribúciós szemantika a neurális gépi tanulásnak köszönhetően új lendületet kapott. A folyamatos fejlesztések során az új megoldások a korábbi modellekre bevezetett tesztelési eszközrendszert használhatták, és „saját pályájukon" tudták legyőzni a korábbi megoldásokat. Előnyük nem volt ugyan megdöbbentő (például Tóth 2018 minimális különbséget mutatott ki a korábbi tulajdonságvektoros és a neurális szóbeágyazások közt), viszont éppen elég volt ahhoz, hogy a számítógépes nyelvészetben elterjedjenek (Baroni–Dinu–Kruszewski 2014). Ez tette később lehetővé, hogy olyan képességekre is szert tegyenek a modellek, amelyeket gépi tanulás nélkül nem lehetett megoldani. Így lett megvalósítható a szavak kontextualizációja is, amely a – fent is említett – többértelműség problémáját segít megoldani, és ez egy általánosan kihasználható tulajdonsága lett a neurális nyelvi modelleknek. Példaként említhetjük, hogy Tóth–Abdelzaher (2023) szótári címszavak jelentéseinek elkülönítésére, Furkó–Tóth (2023) pedig szavak diskurzusjelölői használatának azonosítására használta a kontextualizált neurális szóbeágyazásokat.

Az itt bemutatott könyv első fejezetének végén szintén megjelennek a könyv kiadásával nagyjából egy időben publikált, neurális gépi tanulásra támaszkodó disztribúciós szóbeágyazások. A könyvnek ez a része azt is bemutatja röviden, hogy az új módszerrel kapott tulajdonságvektorokkal analógiás műveleteket is el lehet végezni, például az *apple* és az *apples* szavak ('alma', 'almák') tulajdonságvektorainak

különbsége hasonló vektort eredményez a *car* és a *cars* szavak ('autó', 'autók') különbségéhez. Ez már előrevetíti azt is, hogy a disztribúciós tulajdonságvektorok nem csupán a szó jelentését, hanem használatuk további tulajdonságait, az adott példában az alaktani jegyeit (többes szám) is leírják és feldolgozhatóvá teszik. Egy későbbi tanulmányában a szerző úgy fogalmaz, hogy a neurális nyelvi modellekben „olyan reprezentációk keletkeznek, amelyek a szavak disztribúcióját tükrözik, a megfigyelt mintázatok szemantikai, morfológiai és egyéb okait egyszerre megragadva" (Tóth 2022). A disztribúciós reprezentációk e tulajdonsága magyarázhatja azt is, hogy a rájuk támaszkodó fejlesztések a 2020-as évekre a mesterségesintelligencia-kutatások univerzális eszközeivé váltak.

Irodalomjegyzék

Baroni, Marco – Dinu, Georgiana – Kruszewski, Germán 2014: Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland: Association for Computational Linguistics. 238–247. Online: https://doi.org/10.3115/v1/P14-1023

Firth, John Rupert 1957: A Synopsis of Linguistic Theory 1930–1955. In: Palmer, F. R. (szerk.): *Selected Papers of J.R. Firth 1952–59*. London: Longmans. 168–205.

Furkó Péter – Tóth Ágoston 2023: Mit tud a mesterséges intelligencia a diskurzusjelölőkről: esettanulmányok a USAS és a BERT egyértelműsítési módszereiről. In: Horváth Viktória et al. (szerk.): *Empirikus társalgáskutatás Magyarországon.* Budapest: HUN-REN Nyelvtudományi Kutatóközpont. 85–107. Online: https://doi.org/10.18135/Empirikus.2023.5

Lenci, Alessandro 2008: Distributional Semantics in Linguistic and Cognitive Research. *Italian Journal of Linguistics* 20(1): 1–31.

Rubenstein, Herbert – Goodenough, John B. 1965: Contextual Correlates of Synonymy. *CACM* 8/10: 627–633. Online: https://doi.org/10.1145/365628.365657

Sahlgren, Magnus 2008: The Distributional Hypothesis. *Italian Journal of Linguistics* 20/1: 33–53.

Tóth Ágoston 2013: Az ember, a korpusz és a számítógép. Magyar nyelvű szóhasonlósági mérések humán és disztribúciós szemantikai kísérletben. *Argumentum* 9: 301–310.

Tóth, Ágoston 2014a: *The Company that Words Keep: Distributional Semantics*. Debrecen: Debrecen University Press.

Tóth Ágoston 2014b: Disztribúciós szemantikai kísérletek: szavakról, számokban. In: Ladányi Mária – Vladár Zsuzsa – Hrenek Éva (szerk.): *MANYE XXIII. Nyelv – Társadalom – Kultúra. Interkulturális és multikulturális perspektívák I.* Budapest: MANYE – Tinta Kiadó. 139–145.

Tóth Ágoston 2018: Recognizing Semantic Frames Using Neural Networks and Distributional Word Representations. *Argumentum* 14: 400–414.

Tóth Ágoston 2022: Egy szó mint száz szám: mit tanulhatunk a neurális nyelvmodellektől a nyelv működéséről? In: Navracsics Judit – Bátyi Szilvia (szerk.): *Nyelvek, nyelvváltozatok, következmények II. Fordítástudomány, terminológia, retorika, kognitív nyelvészet, kontrasztív nyelvészet, interkulturális kommunikáció, névtan*. Budapest: Akadémiai Kiadó.

Tóth, Ágoston – Abdelzaher, Esra 2023: Probing Visualizations of Neural Word Embeddings for Lexicographic Use. In: Medveď, Marek et al. (szerk.): *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Brno, 27–29 June 2023.* Brno: Lexical Computing. 545–566.

Tóth András a Debreceni Egyetem Bölcsészettudományi Karának hallgatója, a Debreceni Egyetem Tehetséggondozó Programjának (DETEP) résztvevője. Kutatási területe a nagy nyelvi modellek működése és felhasználása a szókincsfejlesztés területén. E-mail: tandrasdeb@gmail.com

Tóth András 2025: Recenzió Tóth Ágoston *The Company that Words Keep: Distributional Semantics* című könyvéről. *Filológia.hu* 2025/1–2: 93–97.