

SZAVAK JELENTÉSVÁLTOZÁSÁNAK VIZSGÁLATA A STATISZTIKUS FIZIKA ESZKÖZEIVEL

Asztalos Bogdán
ELTE, fizikus hallgató

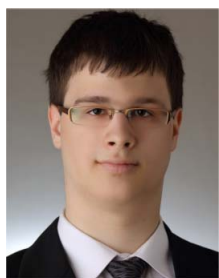
Dolgozatom címe első olvasásra talán meglepő lehet: „mit keres a TDK fizika szekciójában egy szavak jelentésével foglalkozó cikk?” – kérdezhetnénk. Ennek magyarázata az, hogy a természettudományok fejlődésével egyre több területen foglalkoznak

olyan rendszerek vizsgálatával, amelyek túl bonyolultak ahhoz, hogy egzaktul leírjuk őket, de viselkedésükről mégis tudunk törvényszerűségeket megállapítani. Ezek tanulmányozásánál – legyen szó egy összetettebb biológiai struktúráról, kémiai részecskéek összességéről, vagy egy, a fizikában gyakran emlegetett „komplex rendszerről” – fontos szerepet tölt be a statisztika, amely segítségével anélkül ismerhetjük meg a dolgok mérhető és tapasztalható tulajdonságait, hogy meg kéne értenünk annak alapvető működését.

Mivel az emberek által beszélt nyelv is egy ilyen komplex rendszer, amelynek mélyebb tanulmányozása bonyolult, de használata logikai szabályokra épül, ezért a természettudományban megismert módszerek a nyelv vizsgálatára is alkalmasak lehetnek. Kutatásunk célja éppen ez, hogy a statisztikus fizika eszköztárát alkalmazva, nyelvi folyamatokra derítsünk fényt.

Az ötlet, hogy a nyelvet statisztikai módszerekkel tanulmányozzuk nem új, mind a nyelvészet felől, mind a számítástudomány felől sokan foglalkoztak már a témával. Mint fizikusok, elsősorban fizikus szemlélettel állunk a kérdésekhez, de a téma tudományközi jellegére való tekintettel elengedhetetlen olyan nyelvészeti vagy informatikai fogalmakat megismernünk, mint például a disztribúciós hipotézis vagy a szóbeágyazás.

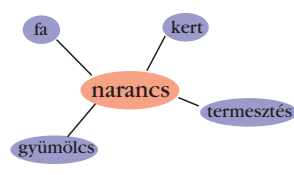
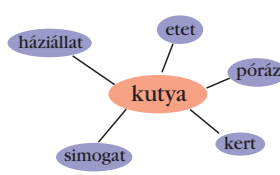
Az Országos Tudományos Diákköri versenyen – témavezetők: *Palla Gergely* és *Czégel Dániel* – 1. díjat elért tanulmány *Fizikai Szemle* részére készített összefoglalója. Az Emberi Erőforrások Minisztériuma ÚNKP-18-1 kódszámú Új Nemzeti Kiválóság Programjának támogatásával készült.



Asztalos Bogdán 2016-ban érettségizett a budapesti Baár-Madas Református Gimnáziumban, jelenleg az ELTE fizikus mesterképzés első éves hallgatója. Fizikában és matematikában is az elméleti problémák megoldása, és az ahhoz szükséges sajátos gondolkodásmód érdekli. Néhány éve részt vesz egy nyelvi folyamatokkal foglalkozó statisztikus fizikai kutatásban, amivel 2019 áprilisában OTDK 1. díjat nyert.



1. ábra. A disztribúciós hipotézis szemléltetése. Mivel a *macska* és a *kutya* szavak jelentése jobban hasonlít egymásra, a környezetükben előforduló szavak között több közös van, mint a *narancs* szó környezete között.



A továbbiakban látni fogjuk, hogy ezek segítségével miként transzformálható át egy elsősorban nyelvészeti téma fizikai rendszerré.

Disztribúciós szemantika

A szemantika a nyelvészet szavak jelentésével foglalkozó részterülete. Mivel a nyelvtudomány része, ezért elsősorban a humán tudományterületekhez tartozik, de a kognitív tudományok révén egyre több természettudományos kutatás alapjául is szolgál. Tekintve, hogy egy szó jelentésétől erősen függ, milyen szavak fordulnak elő körülötte, a szemantikának fontos eszköze, hogy a szavak egymáshoz képesti előfordulását tanulmányozzák. A disztribúciós szemantika nagy mennyiségű szöveges forrás statisztikai elemzésével próbál a szavak eloszlására vonatkozó információkat megállapítani (innen a disztribúciós jelző), és jelentésükre ebből következtetni.

A disztribúciós szemantika fontos mennyisége a szemantikai hasonlóság, amely egy adott szópár esetében megadja, hogy a két szó jelentése mennyire hasonló. Ez tulajdonképpen a nyelvészetből ismert szinonimafogalom kiterjesztése, de amíg a szinonima csak két teljesen megegyező jelentésű szót kapcsol össze, addig a szemantikai hasonlóság tetszőleges két szó közötti viszonyt képes jellemezni. Például, mondhatjuk, hogy a *kutya* és a *macska* szó ugyanúgy nem szinonimák, mint a *kutya* és a *narancs*, de tudjuk, hogy a *kutya* és *macska* szavak között nagyobb a szemantikai hasonlóság, mint a *kutya* és a *narancs* között.

A szemantikai hasonlóságot gyakran ténylegesen ki akarjuk számolni, ehhez biztosít lehetőséget a disztribúciós szemantika fontos állítása, a disztribúciós hipotézis:

Ha A és B két nyelvi kifejezés, akkor a közöttük lévő szemantikai hasonlóság mértéke azon nyelvi környezetek hasonlóságának a függvénye, amelyekben A és B elő tud fordulni [1].

Egyszerűbben megfogalmazva: hasonló szavak körül előforduló szavak szintén hasonlóak egymáshoz. Ezt szemlélteti az 1. ábra is.

A disztribúciós hipotézis jelentősége abban rejlik, hogy igen hasznos eszközt ad a kezünkbe, ugyanis eszerint a szavak szemantikai hasonlósága a környezetük statisztikai vizsgálatával tanulmányozható. Ezt alapul véve, különböző eljárások találhatók, amelyekkel a szavak közötti szemantikai hasonlóságot becsülni lehet, ilyen például az általunk használt szóbeágyazás technikája is.

Szóbeágyazás

A disztribúciós hipotézis alapján felmerülhet bennünk az igény, hogy bármely két szó közötti szemantikai hasonlóságot kiszámoljuk (vagy legalább megbecsüljük). Ha közvetlenül a szöveg alapján tennénk meg, akkor minden egyes alkalommal, amikor két szó közötti viszonyt keressük, az egész szöveget kellene statisztikailag elemezni, ami rendkívül sok munka lenne. Sokkal egyszerűbb, ha a szövegben található információkat egyszer kinyerjük, és azt egy könnyen kezelhető objektumként tároljuk.

A szóbeágyazás alapja, hogy minden szó környezetének statisztikai tulajdonságait külön-külön eltároljuk egy számvektorban, és amikor két szó szemantikai hasonlóságát keressük, elég a két szóhoz rendelt számvektorral foglalkozni. A számvektorként való reprezentáció azért praktikus, mert matematikai tulajdonságaik jól ismertek és informatikai szempontból (ezáltal az adatok feldolgozásának szempontjából) is könnyen kezelhetők.

Ezen kívül, tudjuk, hogy egy n elemű számvektor egyben egy n dimenziós tér egy pontjaként is felfogható. Emiatt, ha a szavakat vektorokkal azonosítjuk, az formálisan úgy is értelmezhető, mintha létezne egy elvont nyelvészeti tér, a jelentéstér, amiben a szavak pontoknak felelnek meg. (A szóbeágyazás név is onnan ered, hogy a szavakat egy többdimenziós térbe ágyazzák be.) Szemléletesen ez úgy magyarázható, hogy a beágyazási tér különböző dimenziói a különböző nyelvi környezeteknek felelnek meg, és a szóreprezentációk különböző komponensei azt mutatják, hogy az adott környezet mennyire jellemző a reprezentált szóra.

A klasszikus szóbeágyazási módszerek (LSA,¹ PPMI²) pontosan ilyen alapon működtek: a szót reprezentáló vektor i -ik komponense attól függött, hogy az i -ik kontextusszó hányszor szerepelt a szó környezetében. Két szó szemantikai hasonlóságát a két szót reprezentáló vektor komponenseinek korrelációjával (azaz a két vektor által bezárt szög koszinuszával) becsülték. Azt, hogy melyik szóbeágyazás jó és melyik nem, általában tapasztalati úton, azaz humán szóasszociációs vagy egyéb pszicholingvisztikai vizsgálatokkal döntötték el. Az újabb szóbeágyazások már bonyolultabb módszerrel készítik el a szavakat reprezentáló vektorokat, ezért a különböző dimenziók

ilyen szemléletes jelentése eltűnt, valamint a beágyazás mögötti elv is sokkal elvontabb lett, de a szemantikai hasonlóság kiszámítása is jóval hatékonyabbá és pontosabbá vált [2].

Ugyan a szóbeágyazás módszerének ötletét már az 1960-as években kitalálták, de igazán csak a 2000-es években lett széles körben ismert, amikor már elég adat állt rendelkezésre a reprezentáció elkészítéséhez. 2012 után, a grafikus processzorok általánossá válásával elindult a „deep learning forradalma”, amely a számítástudomány legtöbb területét gyökerestül változtatta meg, hiszen olyan területeken kezdtek el gépi tanuláson alapuló eljárásokat alkalmazni, ahol azelőtt – a nagy adatmennyiség feldolgozása miatt – reménytelen lett volna. Így történt a szóbeágyazásokkal is: megjelentek az olyan szóbeágyazó algoritmusok, amelyek a gépi tanulás segítségével, mesterséges neurális hálózatokat használva „tanulják” meg a szavak legjobb beágyazását.

A Word2vec több olyan modell csoportja, amely mesterséges neurális hálókkal végez szóbeágyazást. Kifejlesztését *Tomas Mikolov* vezetésével a Google egyik kutatócsoportja végezte, eredményeiket 2013-ban két cikkben publikálták [2, 3], és ezek annyira hatékonyak működtek, hogy az utóbbi évek legnépszerűbb szóbeágyazási modelljeivé nőttek ki magukat. Kutatásunk során az úgynevezett skip-gram modellt használtuk, amely a beágyazott vektorokat a szavak környezetének valószínűségi becslése alapján állítja elő. Ez tipikusan ezért néhány százezer szó hatékony beágyazására képes egy pár száz dimenziós térbe.

Kutatásunk alap gondolata

A beágyazott szavak – ahogy láttuk – egy (sokdimenziós) geometriai tér pontjaként képzelhetők el, így a szavak jelentésére úgy tekinthetünk, mint egy elvont nyelvészeti tér, a jelentéstér elemeire. Mindennapi tapasztalatunk, hogy a szavak jelentése az évek során időben változik, ez a jelentéstérben úgy képzelhető el, mint pontrészcskék vándorlása. Figyelve tehát, hogy az idő függvényében miként változott a beágyazott szavak halmaza, pontrészcskék bolyongását látjuk, amelynek vizsgálatára léteznek statisztikus fizikai módszerek. Ezeket használva állításokat fogalmazhatunk meg a szavak jelentésváltozására, amelyekkel akár nyelvi törvényszerűségekre is fényt deríthetünk, és mindezt a természettudomány eszközeivel.

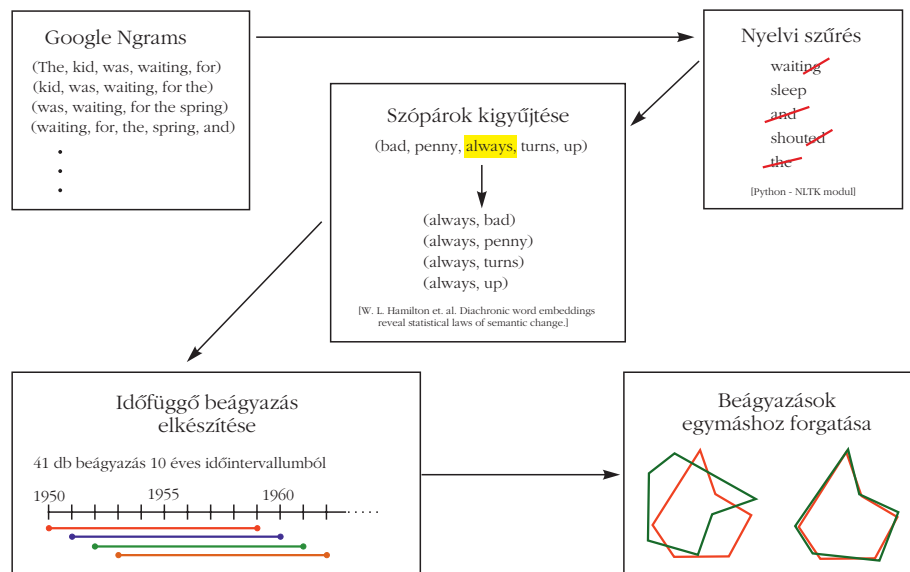
Adatok feldolgozása

Konkrét vizsgálatokat elvégzéséhez saját beágyazást (beágyazásokat) kell készítenünk, amelyeket megvizsgálva információt szerezhetünk a szavak viselkedéséről.

Az adatok feldolgozásának folyamatát vázlatosan a 2. ábra szemlélteti, a főbb munkafázisok az alábbiaként foglalhatók össze.

¹ Latent semantic analysis

² Point-wise mutual information



2. ábra. Az időfüggő beágyazás elkészítésének főbb lépései.

tött adatokból ezeket az összetartozó szó párokat gyűjtöttük ki. Az előfordulási adatokat egy olyan $N \times N$ méretű mátrixban tároltam el (N az összes szó száma), amelynek i -ik sorának j -ik oszlopa megadja, hogy az i -ik szó környezetében hányszor fordult elő a j -ik szó. Egy ilyen mátrixot szemléltet az 1. táblázat.

Időfüggő beágyazás

Következő lépésként lefutattuk a beágyazást elkészítő programot. A skip-gram modell a tanításhoz szó párokat vár, ezek az együtt-előfordulási mátrixból könnyen kigyűjtethők, hiszen minden cella az adott szó pár előfordulását je-

lenti. A programot, amely magát a beágyazást végzi, egy másik kutatócsoport publikálta a saját kutatásuk keretében [4]. A beágyazás fontos paramétere a dimenziók száma. Esetünkben, követve korábbi kutatásokat, ez 300 volt [4, 5].

Mivel a szavak jelentését időben vizsgáljuk, ezért minden évre külön elkészítettük a megfelelő együtt-előfordulási mátrixot. Ahhoz, hogy adott beágyazáshoz mégis megfelelő mennyiségű adatunk, ne csupán egy évnyi legyen, egy beágyazás során 10 év adatait vettük figyelembe, és ezeket évenként csúsztattuk. Így összességében az 1950 és 1999 közötti adatokat 41 különböző intervallumban dolgoztuk fel.

Szavak beforgatása

A beágyazás eredményeként minden szóhoz, minden évben egy öt reprezentáló 300 dimenziós vektort kaptunk. Ha egy adott időpontbeli beágyazás összes vektorát elforgatjuk a 300 dimenziós térben, akkor a szavak egymáshoz képesti távolságai nem változnak, tehát a két beágyazás ekvivalens. Ez azért okoz problémát, mert két, egymáshoz képest elforgatott pont-halmaz – bármennyire is hasonlók – hasonlósága nem vizsgálható (lásd 2. ábra, jobb alsó illusztráció).

Adatok letöltése

A disztribúciós szemantikához természetesen szorosban kapcsolódnak a nagyméretű szöveggyűjtemények, hiszen, mint a nyelv használatának írásban megmaradt nyomai, ezek a szavak statisztikai tulajdonságainak első számú információforrásai. Különös jelentőségük – amióta elektronikusan különböző formákban elérhetők, és így feldolgozásuk automatizálható – növekedett.

Kutatásunk során a Google Books Ngram Viewer ingyenesen elérhető, angol nyelvű adatbázisát³ használtuk. Ez az adatbázis a Google egyik nagyobb szabású projektjén, a Google Könyveken alapszik, amelynek keretében több millió könyv szöveges állományát gyűjtötték össze. Az Ngram Viewer keretében a szövegekből kigyűjtötték az összes szót, szó párt, szóhármast, ..., szó-n-est (más szóval n -grammot), és összeszámolták, hogy ezek a különböző éveken, 1800-tól 2008-ig milyen gyakran fordulnak elő.

Nyelvi szűrés

Miután letöltöttük az adatokat, első dolgunk, hogy a nyers szövegből kinyert szavakat annak érdekében egységesítsük, hogy rajta statisztikai vizsgálatokat lehessen végezni. E célból, egy erre készült algoritmus⁴ a toldalékoltszavakat szótőre egyszerűsítettünk, majd ezután a leggyakrabban használt, jelentéssel nem bíró szavakat – névelőket, kötőszavakat, segédigéket, amelyek csak torzítanak a statisztikai vizsgálatok eredményét – eltávolítottuk.

Együtt-előfordulási mátrix elkészítése

A beágyazó algoritmus az egymás közelében előforduló szó párokat igényli bemenetként, ezért a letöl-

1. táblázat

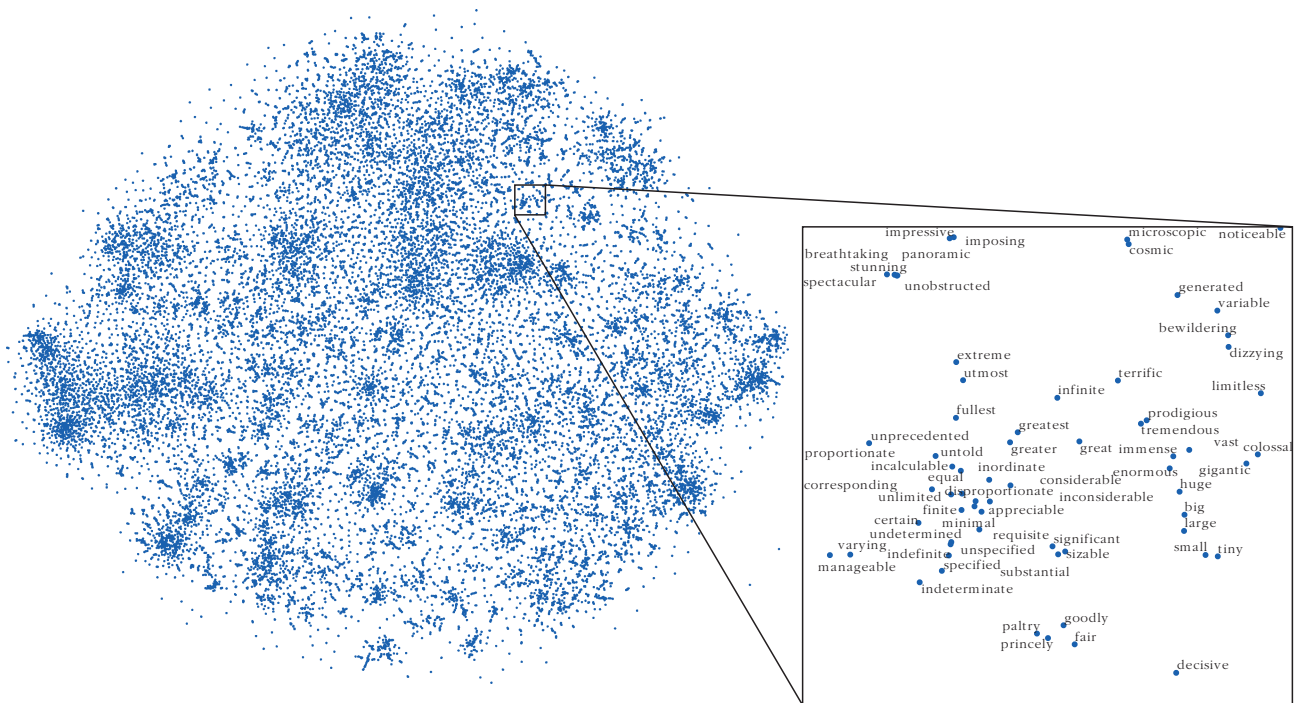
Példa egy együtt-előfordulási mátrixra

	kutya	macska	enni	póráz
kutya	2	8	23	10
macska	8	3	21	1
enni	23	21	5	0
póráz	10	1	0	2

A *kutya* szó környezetében a *póráz* 10-szer fordul elő, de a *macska* környezetében csak 1-szer, hiszen ez inkább a kutya-hoz köthető fogalom, míg az *enni* szó mindkét állat környezetében hasonló számban fordul elő.

³ Elérhető a <https://storage.googleapis.com/books/ngrams/books/datasetsv2.html> címen.

⁴ A nyelvi szűrést a Python programnyelv nltk (Natural Language Toolkit) modul megfelelő célfüggvényeivel végeztük.



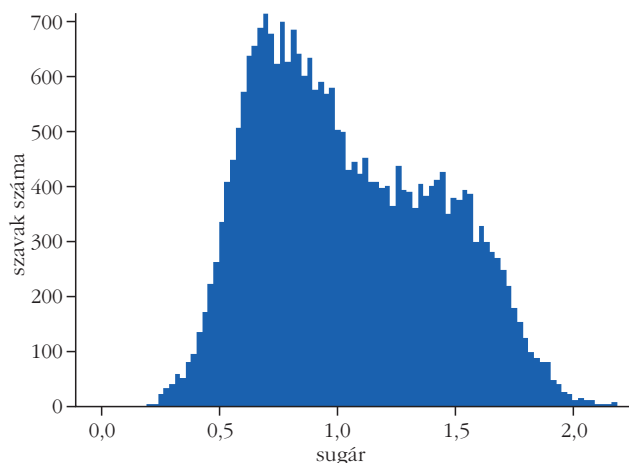
3. ábra. A 300 dimenziós beágyazás szerkezetének kétdimenziós vizualizációja. A levetítés a t-SNE (T-distributed Stochastic Neighbor Embedding) algoritmussal történt, ami egy nemlineáris sztochasztikus folyamat segítségével képez le többdimenziós adathalmazt alacsony dimenzióba úgy, hogy az adatpontok közötti távolságviszonyok minél jobban megmaradjanak [7].

Ezt a jelenséget úgy küszöböltük ki, hogy a szomszédos beágyazásokat „egymáshoz forgattuk”, azaz megkerestük mindegyik időpont beágyazásának azt az állását, ami legjobban hasonlít az előző időpontéhoz [6].

A beágyazás néhány tulajdonsága

Mielőtt a szavak időbeli viselkedését vizsgálnánk, nézzük meg egy adott beágyazás néhány érdekes tulajdonságát! (Az alábbi eredmények az 1990 és 1999 közötti évek adataiból gyártott beágyazás alapján készültek.) A beágyazás egészének a struktúráját a 3. ábra szemlélteti, ez a 300 dimenziós adathalmaz két-dimenziós levetítése.

4. ábra. A szavak origó körüli radiális sűrűsége.



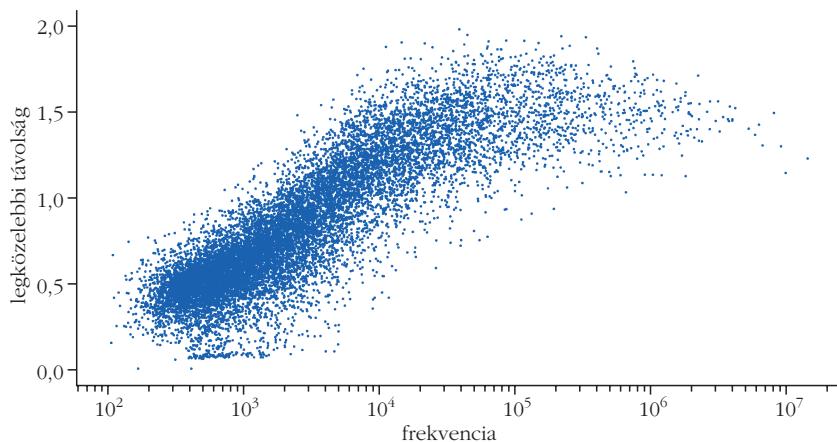
Jól látszik, hogy a szavak különböző csomópontokba – témakörökbe – gyülekeznek, és a kép egy-egy kisebb tartományára ráközelítve láthatjuk azt is, hogy egymáshoz közel a szinte teljesen megegyező jelentésű szavak helyezkednek el.

Érdekes tulajdonság, látható a 4. ábrán is, amely a szavak radiális sűrűségét ábrázolja az origó körül. Ez pontosan azt jelenti, hogy az origótól adott távolságba eső tartományban hány darab szó helyezkedik el. Ahogy látható, a szósűrűség nem egyszerűen megnő majd a végtelenben lecseng, hanem egy furcsa kettős púp alakot mutat. Ennek pontos okát nem tudjuk, de későbbi vizsgálatoknál nem árt észben tartani, hogy a szavak struktúrája nem olyan egyszerű, mint elsőre gondolnánk.

A szófelhő további különleges tulajdonságát ismerhetjük meg, ha a szavak körüli üres térrész sugarának nagyságát annak függvényében ábrázoljuk, hogy a szó milyen gyakran fordul elő a nyelvben. (Az üres térrész alatt most azt a távolságot értjük, ami az adott szóhoz legközelebb lévő szó távolsága.) Az 5. ábra alapján megállapíthatjuk, hogy a gyakoribb szavak ritkább társaiknál sokkal nagyobb térrészt foglalnak el a jelentéstérből.

Szavak időfejlődésének viszonya a Brown-mozgáshoz

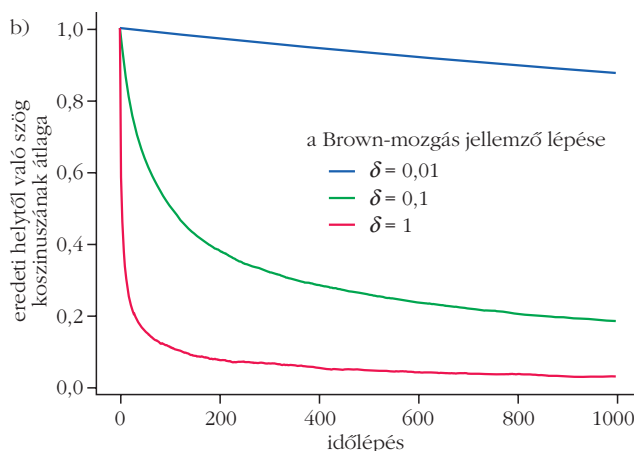
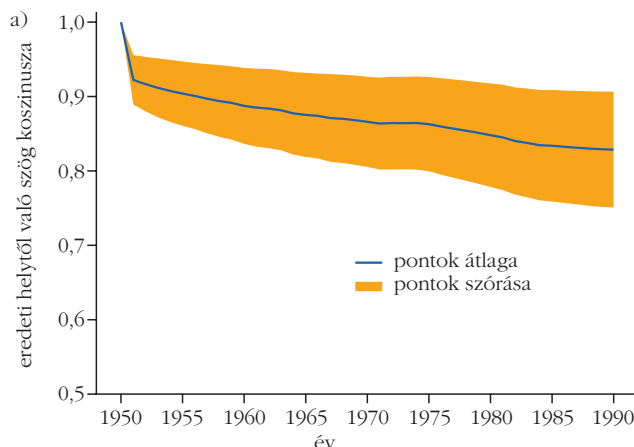
Ha a szavak mozgását a jelentéstérben úgy képzeljük el, mint a geometriai térben bolyongó részecskéket, jó lenne a mozgásukra vonatkozó törvényeket, szabályokat megállapítani. Bolyongó részecskék vizsgálata-



5. ábra. A szavakat reprezentáló pontok körüli üres térrész sugara a szavak gyakoriságának a függvényében.

kor általánosan elfogadott elemezni, hogy a kiindulási helyüktől időben miként távolodnak, és ezt ismert mozgások eredményeivel összehasonlítani. A szavak esetében ez a feladat azért bonyolultabb, mert a szemantikai hasonlóságot – konvencionálisan – a szavakat reprezentáló pontokba mutató vektorok által be-

6. ábra. (a) A beágyazási pontok első beágyazásban elfoglalt helyüktől való távolságátlaga és szórása. (b) A jelentéstérben Brown-mozgást végző részecske várható távolodása az eredeti helyéről, különböző δ lépésmagasságok esetében. A távolságot mindkét ábrán a vektorok által bezárt szög koszinuszával tüntettük fel, tehát az, hogy a görbék értékeinek 1-ről való csökkenése a kezdeti helytől való távolodást fejezi ki.



zárt szöggel mérik, emiatt viszont a távolodás mértéke nem írható fel a megszokott formában.

Egészen pontosan azt vizsgáltuk, hogy az egyszerű Brown-mozgó részecskék miként viselkednek a jelentéstérben, és ezt hasonlítottuk össze a beágyazott pontok viselkedésével. A szó jelentésváltozását és a Brown-részecskék távolodását egyaránt a különböző beágyazásokban reprezentáló vektorok által bezárt szöggel jellemezzük.

A beágyazott pontok esetében ezt a távolodást a 6.a ábrán láthatjuk. A szavak átlagosan valóban távolodnak eredeti jelentésüktől, ám ez a

távolságváltozás csak a viszonyítási pontból való első ellépés után lényeges, azt követően a viszonyítási ponttól való távolodás nagyban lelassul.

A jelentéstérbeli Brown-mozgást az alábbi módon értelmeztük. A részecske elindul a kiindulási helyéről és minden időlépésben arrébb ugrik egyet. Az ugrás irányát és nagyságát egy 0 közepű, δ szórású, 300 dimenziós normál eloszlás szerint választja meg. Mivel a távolságot a vektorok szögével mérjük, ezért a várható távolság időfüggése nem a megszokott gyökfüggvény alakú. Az összefüggés megállapításához egy mesterséges ponthalmazon, különböző δ értékek esetén végzett szimuláción mértük meg a távolodást (6.b ábra).

Brown-mozgás esetében, bár a pontok szintén távolodnak a kiindulási ponttól, de a görbék – ellentétben az eredeti szavak esetén látott első lépésben való hirtelen távolodással – elég gyorsan és egyértelműen a 0 értékhez tartanak. Ez alapján úgy tűnik, hogy a szavak dinamikája nem modellezhető jól a kölcsönhatásmentes, memóriánélküli Brown-mozgással, hanem leírásához valamilyen bonyolultabb modell szükséges.

Kitekintés

Kutatásunk jövőbeli terveihez tartozik, hogy a Brown-mozgástól való eltérést kvantitatív módon jellemezzük, és olyan modellt találjunk, amellyel jól leírható a szavak dinamikája. A kiindulási alapötletre építve elsődleges célunk, hogy törvényszerűségeket ismerjünk fel a szavak mozgásának időfüggésében, és ezekkel ismert nyelvészeti szabályok okát magyarázzuk meg, vagy újakat mondjunk ki.

Irodalom

1. Alessandro Lenci: Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics* 20/1 (2008) 1–31.
2. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. *CoRR* (2013) abs/1301.3781 <http://arxiv.org/abs/1301.3781>
3. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean: Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems* 26. (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger eds.) Curran Associates, Inc. (2013) 3111–3119.

- <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>
4. William L. Hamilton, Jure Leskovec, Dan Jurafsky: Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint* arXiv:1605.09096 (2016).
 5. Omer Levy, Yoav Goldberg, Ido Dagan: Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3 (2015) 211–225.
 6. Peter H. Schönemann: A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31/1 (1966) 1–10.
 7. Laurens van der Maaten, Geoffrey Hinton: Visualizing data using t-SNE. *Journal of machine learning research* 9 (2008. nov.) 2579–2605.