PG70

MELLÉKNEVEK DISZTRIBÚCIÓS ÉS SZEMANTIKAI MINTÁZATAI

HÉJA ENIKŐ[1], GÁBOR KATA[2], GYŐRFFY ANDRÁS[1], LIGETI-NAGY NOÉMI[1], SIMON LÁSZLÓ[1], LIPP VERONIKA[1] | [1]HUN-REN Nyelvtudományi Kutatóközpont
[2]INALCO, Equipe de recherche textes, informatique, multilinguisme
{heja.eniko, gyorffy.andras, ligeti-nagy.noemi, simon.laszlo, lipp.veronika}@nytud.hun-ren.hu
kata.gabor@inalco.fr | DOI: 10.18135/PG70.2024.7

1. BEVEZETÉS

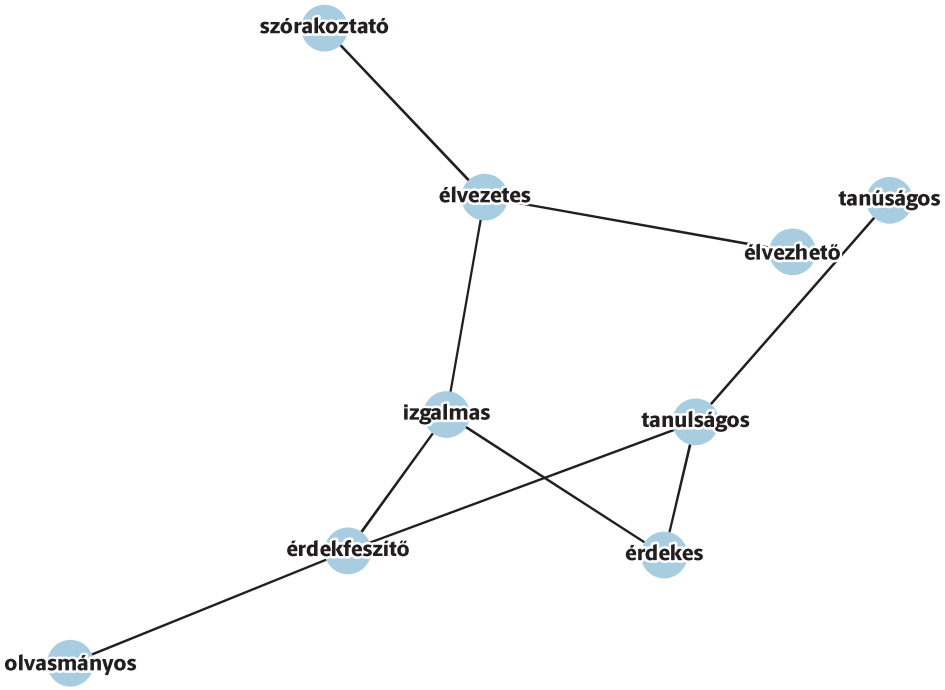
Jelen írásban egy folyamatban lévő kutatást vázolunk fel, amelynek eredményei egyaránt felhasználhatóak a lexikális szemantika, a lexikográfia és a nyelvtechnológia területén (l. Héja et al., 2023).

Amennyire tudjuk, a magyar melléknevek lexikális szemantikáját még nem írták le kimerítően. Részletesen Kiefer Ferenc foglalkozott ezzel a témakörrel a *Strukturális Magyar Nyelvtan 4* 8. fejezetében. Kiefer (2008) a szemantikai csoportok elkülönítése során a melléknevek viselkedését vette alapul, és a következő négy tulajdonság alapján a melléknéveket három osztályba sorolta: figyelembe vette, hogy a melléknevek megjelenhetnek-e attributív vagy predikatív pozícióban a módosított főnévhez képest, fokozhatóak-e, továbbá módosíthatóak-e a *nagyon* határozószóval. Ezen tulajdonságok alapján a következő csoportokat különítette el (1) abszolút melléknevek; (2) relatív melléknevek (a mértéket jelölő és értékelő melléknevek); (3) rendhagyó melléknevek. Ez utóbbi csoportot példázza a „*János volt az állítólagos betörő*” mondatban az *állítólagos* melléknév. Kutatásunk egyik célja, hogy megvizsgáljuk, hogy az általunk javasolt korpuszvezérelt módszer milyen mértékű átfedést mutat a Kiefer (2008)-ban található csoportokkal, de ambicionáljuk az osztályozás további finomítása mellett a melléknevek szélesebb körére való kiterjesztést is – amennyiben ezek lehetségesek.

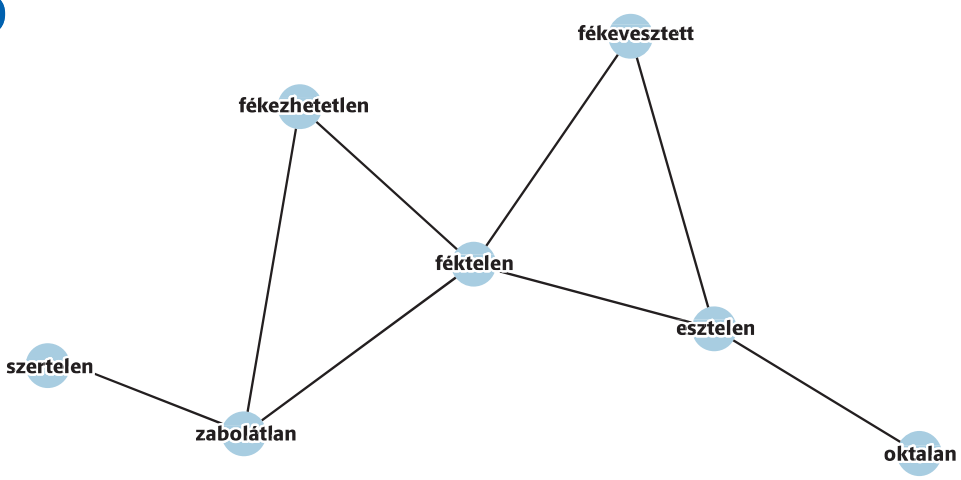
Elvárásunk szerint a melléknevek lexikai szemantikai leírása felhasználható alkalmazott nyelvészeti kutatásokban is: a lexikográfiában egynyelvű értelmező szótárak szerkesztését segíti, míg a nyelvtechnológiában a nagy nyelvi modellek (pl. ChatGPT) kiértékelését lehetővé tevő konzisztens benchmarkkorpuszok létrehozását könnyíti meg, illetve teszi konzisztensebbé. Fontos hangsúlyozni, hogy a javasolt módszer adatvezérelt, így a leírás során kiküszöböljük az érzeményszemantikát, a szemantikai intuíciónkra való támaszkodást. De vizsgáljuk annak a lehetőségét is, hogy az eredmények kiértékelése során mennyiben tudjuk kizárni az intuíció szerepét: a Kiefer (2008) által felhasznált formai kritériumok mellett más disztribúciós kritériumokra is támaszkodunk, elsősorban a koordinálhatóságra, amelyet kontextuális szóbeágyazásokra építve vizsgálunk.

2. A MÓDSZER LEÍRÁSA

Az alkalmazott módszer statikus szóbeágyazásokon alapszik (Mikolov 2013a, 2013b). Ezt a fajta szóreprezentációs technikát gyakran éri az a kritika, hogy mivel egy szóalakhoz egy vektorreprezentációt rendel, önmagában nem képes a jelentések elkülönítésére (pl. Camacho-Collados & Pilehvar, 2018). Ezt úgy orvosoltuk, hogy a szóbeágyazásokat gráfstruktúrába rendeztük. A módszer lényege, hogy a szóvektorok hasonlóságát felhasználva elkészítjük a gráfot reprezentáló szomszédsági mátrixot, amelyet a következő lépésben binarizálunk egy *K* küszöbérték mentén. A küszöbérték ebben az esetben egy vágást jelent az eredeti, teljes, súlyozott melléknévi gráfon, amelyet ezáltal összefüggő komponensekre bontunk. *K = 0.7* esetén a 10.153 melléknevet tartalmazó eredeti melléknévi gráfot 1.807 összefüggő gráfkomponensre bontjuk fel, amelyek összesen 6.417 melléknevet tartalmaznak. A komponensek figyelemre méltó tulajdonsága, hogy – egy kivétellel – jól definiált szemantikai tartományokhoz tartoznak. Az ilyen hálózatok esetében az egyik komponens mindig egy óriáskomponens, amely a bemeneti mellékneveknek kb. harmadát (3.736) tartalmazza, és amely így több szemantikai domaint is összevon. Az alábbiakban két példát mutatunk összefüggő melléknévi gráfkomponensekre, az első az *izgalmas* melléknév rokon értelmű szavait tartalmazza, míg a második a *féktelen* szemantikai domaint fedi le.



1. ábra. Összefüggő melléknévi gráfkomponens: az *izgalmas* melléknév és rokon értelmű szavai

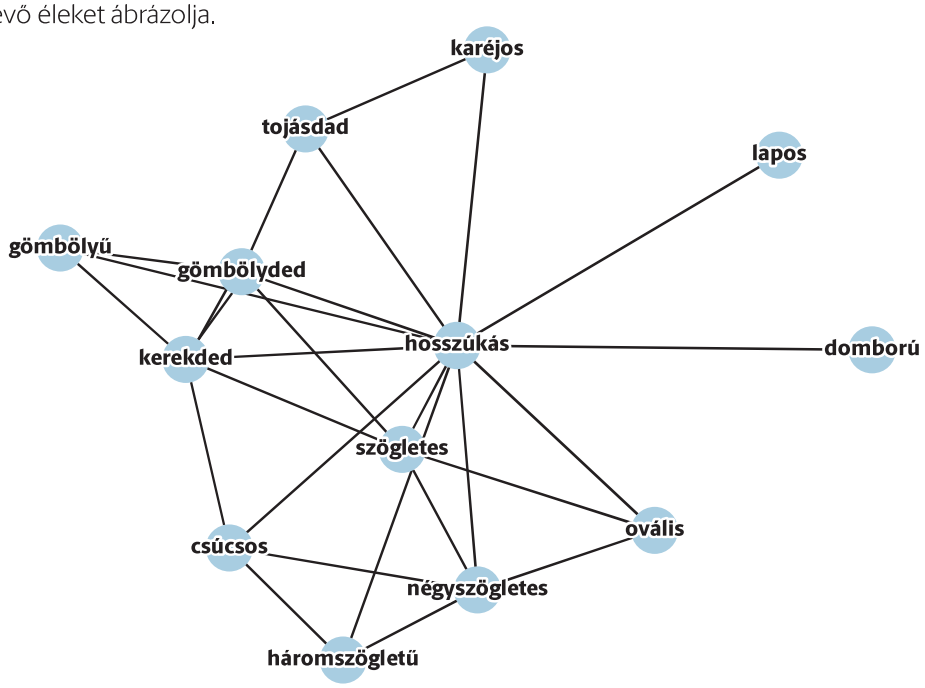


2. ábra. Összefüggő melléknévi gráfkomponens: a *féktelen* melléknév és rokon értelmű szavai

3. EREDMÉNYEK

3.1 Átfedések Kiefer melléknévi csoportjaival

Az eredmények azt mutatják, hogy az automatikusan kinyert melléknévi szemantikai domainek sok esetben megfeleltethetőek a Kiefer (2008)-ban meghatározott csoportoknak. Ez a megfigyelés elsősorban az abszolút melléknevek alcsoportjai esetében igaz: a színnevek, a nemzetiségek nevei, az eredetnevek és az alaknevek mind elkülöníthető gráfkomponenseket alkotnak az automatikusan generált melléknévi gráfban. A következő ábra a *hosszúkás* melléknév szomszédos csomópontjait és a közöttük lévő éleket ábrázolja.



3. ábra. A *hosszúkás* melléknév szomszédos csomópontjainak gráfja

Érdekes, hogy a privatív melléknevek nem alkotnak disztribúciósan koherens csoportot: a Kiefer (2008)-ban felsorolt privatív melléknevek (*vak, süket, néma, sánta*) mind magányos gráfcsomópontokat alkotnak. Ennek disztribúciós okait még vizsgálnunk kell. Az emberi tulajdonságot jelölő, poláris párral nem rendelkező mellékneveket módszerünk további alcsoportokra bontotta. Megfigyelésünk szerint a relatív melléknevek sem alkotnak koherens csoportot, ennek okait is tovább kell vizsgálnunk. A rendhagyó melléknevek a privatív melléknevekhez hasonlóan sokszor fordulnak elő izolált csomópontokként (*állítólagos, lehetséges, előző, egyedüli, fő, közeli, biztos, teljes*). Ez alól az *igazi* és a *tökéletes* melléknevek kivételek: az előbbi a *valóságos* és *valódi* melléknevekkel, az utóbbi pedig az *ideális, optimális* melléknevekkel fordul elő egy gráfkomponensben.

Összefoglalva azt mondhatjuk, hogy az eredményül kapott melléknévi osztályok Kiefer (2008) abszolút mellékneveinek alcsoportjaival mutatnak nagy fokú egyezést. A relatív és rendhagyó melléknévi osztályok esetén további vizsgálatok szükségesek az eltérések magyarázatára.

3.2 Lexikográfiai felhasználás

Láttuk, hogy az automatikusan létrehozott gráfkomponensek részben a szakirodalomban is szereplő szemantikai osztályokba sorolják a mellékneveket. A módszer egy lexikográfiai felhasználása, hogy az így kinyert melléknévcsoportokat automatikusan *jelentésstruktúrák*ba szervezzük. Ebben a lépésben vektorreprezentációik alapján klaszterezzük azokat a főneveket, amelyek előfordulnak a melléknévi szemantikai osztály elemeinek valamelyikével. Az eredményül kapott főnévi klaszterek képezik a jelentésstruktúra alapját. A következő lépésben egy egyszerű korpuszalapú asszociációs mértékkel meghatározzuk, hogy milyen erős a kapcsolat a melléknévi szemantikai osztály egy adott eleme (pl. *esztelen*) és a főnévi klaszterek egy adott eleme között (pl. [*indulat, harag, düh* stb.]). Az 1. táblázat azt mutatja, hogy az *esztelen* leginkább a [*költekezés, pazarlás*] főneveket szubkategorizálja, míg például a [*nevetés, buli, öröm* stb.] szemantikai osztályba tartozó főnevek kevésbé *esztelenek*, mint inkább *féktelenek*.

	4	**5**	**6**	**9**
esztelen	3.2	26.5	13	7.22
fékevesztett	2		1.5	6.83
fékezhetetlen	2.33	2		6.62
féktelen	21.6	12	7.5	21.64
oktalan	3.25	1.5	3.75	12.67
szertelen	6			9.88
zabolátlan	1.33			9
0	nevetés	költekezés	öldöklés	vágy
1	buli	pazarlás	lövöldözés	szenvedély
2	öröm		háborúskodás	indulat
3	száguldás		pusztítás	harag
4	zabálás		ámokfutás	düh
5	ivászat		rablás	gyűlölet
6	jókedv		rombolás	nacionalizmus
7	tombolás			önzés
8	mulatozás			

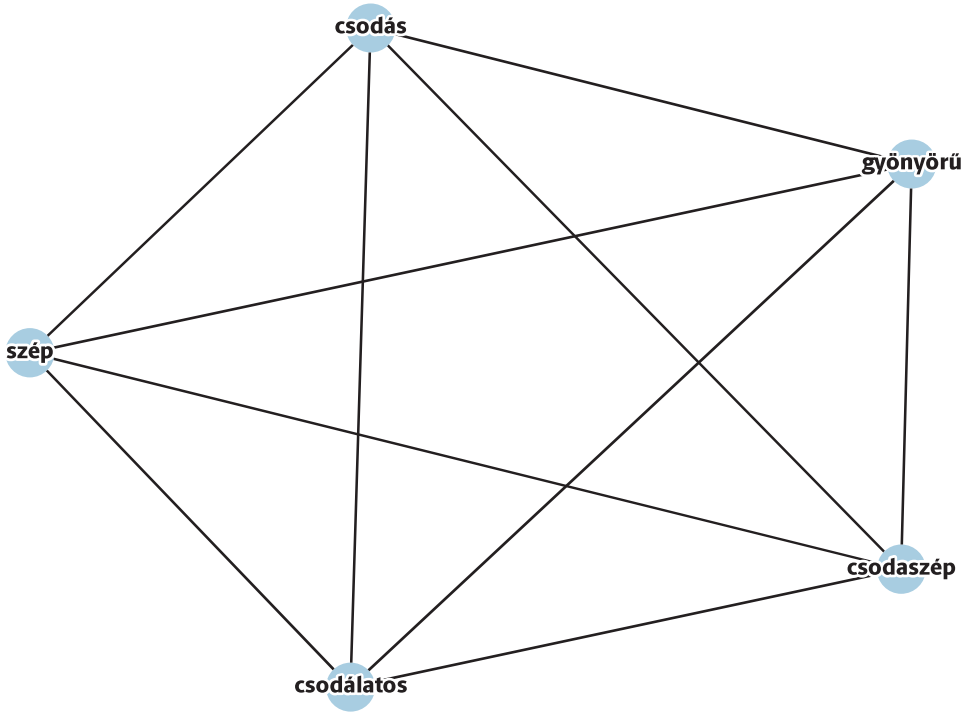
1. táblázat. A féktelen szemantikai osztályba tartozó melléknevek jelentésstruktúrája

Így tehát egy ilyen táblázatban ábrázolható jelentésstruktúra alapján képet kaphatunk arról, hogy az egy szemantikai domainbe tartozó melléknevek milyen szemantikai osztályba tartozó főneveket milyen erősséggel szubkategorizálnak.

3.3 Szemantikai relációk megragadása a melléknévi gráf lokális tulajdonságai alapján

3.3.1 Szinonima

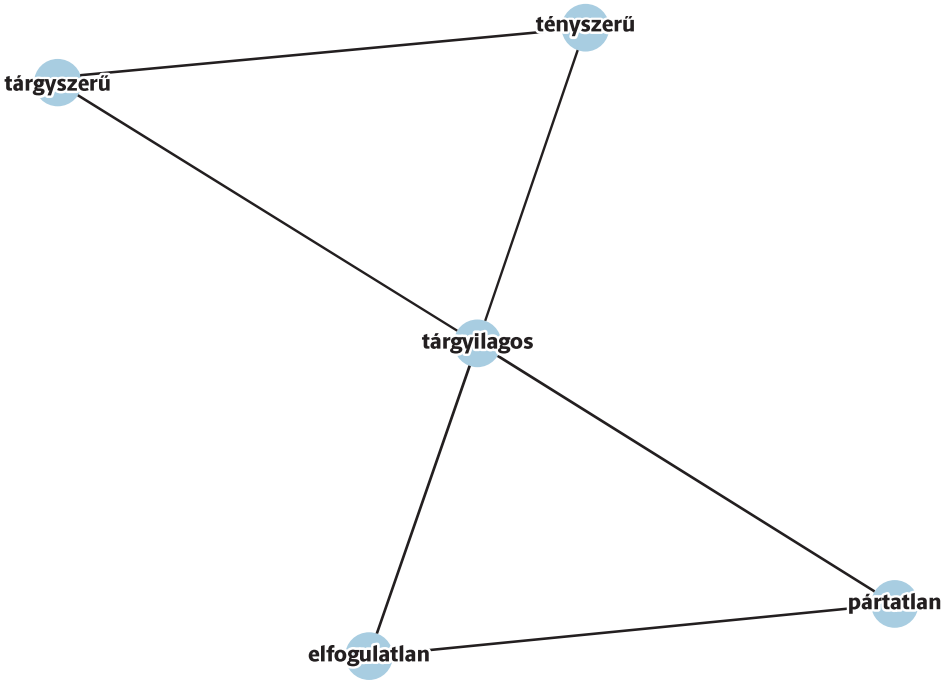
Azt már láttuk, hogy az összefüggő gráfkomponensek szemantikai domaineknek feleltethetőek meg. Ugyanakkor azt tapasztaltuk, hogy a szinonima disztribúciós fogalma is könnyedén értelmezhető a disztribúciós hasonlóság alapján létrehozott gráfokon. Mivel a gráfban két csúcs akkor szomszédos, ha a csúcs által jelölt melléknevek disztribúciós hasonlóságot mutatnak, feltételezzük, hogy a *teljes részgráf*ok (ún. klikkek) lehetővé teszik a szinonimaosztályok – vagyis a melléknévi jelentések – megragadását, hiszen egy klikkben szereplő bármely két melléknév disztribúciósan hasonló lesz egymáshoz. Az alábbi ábrán a *szép* melléknevet tartalmazó melléknévi klikk látható.



4. ábra. A szép melléknevet tartalmazó melléknévi klikk (teljes részgráf)

3.3.2 Poliszémia

Ha a klikkek lehetővé teszik a melléknévi jelentés megragadását, akkor a melléknévi poliszémiát reprezentálhatjuk úgy, hogy egy melléknév egyszerre több klikkhez is tartozik. A következő ábra a *tárgyilagos* melléknév automatikusan kinyert két aljelentését mutatja: egyfelől lehetünk tárgyilagosak úgy, hogy ragaszkodunk a valósághoz, vagyis úgy, hogy *tényszerűek* vagy *tárgyszerűek* vagyunk, másfelől pedig úgy is, hogy egyik félnek sem kedvezünk, ebben az esetben *elfogulatlanok* vagy *pártatlanok* vagyunk.



5. ábra. A tárgyilagos melléknév automatikusan kinyert két aljelentése

A poliszém jelentések automatikus kinyerésén túl az egyes jelentéseket előhívó, triggerelő főnévcsoportok is automatikusan kinyerhetőek a korpuszból. Így például a *tárgyilagos* első jelentése a *leírás, ismertetés* főnevek előtt fordul elő, míg a második jelentése a *megítélés, vélemény, eljárás* főnevek előtt.

Míg az alulspecifikált jelentésű főnevek (pl. *kutya, virág*) esetében a különböző aljelentések mellett megjelenő kontextusok könnyen koordinálhatók, a többjelentésű szavakra jellemző, hogy a különböző jelentést aktiváló kontextusok nem szerepelhetnek egyszerre egy szóelőfordulás mellett (Zwicky & Sadock, 1975). Melléknevekre alkalmazva ez azt jelenti, hogy ha két főnév ugyanazon poliszém melléknév különböző aljelentéseit reprezentáló klikkekhez tartozik, akkor várakozásunk szerint nem lesznek koordinálhatók, például: *# piszkos háború és aszfalt, # finom mozdulat és halétel, # hideg tea és ész.*

A koordinációs teszt széles körű alkalmazása lehetővé tenné az automatikusan előállított melléknévi jelentés-megkülönböztetések validálását, ám ennek emberi erőforrásigénye magas. Ezért jelenleg azt vizsgáljuk, hogy *előtanított nyelvmodellekből* kinyerhető-e hasonló információ. Ehhez olyan ADJ N1 és N2 koordinációs sémákat generálunk, melyekben a poliszém melléknév két mellérendelt főnévvel szerepel. Így összevethetjük, hogy mekkora valószínűséget társít a nyelvmodell a második főnévhez aszerint, hogy az első főnév azonos vagy eltérő melléknévi jelentést hív elő. Alacsony valószínűség esetén a megkülönböztetést érvényesnek tekinthetjük (mint a fenti példákban), ellenkező esetben pedig érdemes megvizsgálni, milyen okból tűnik elfogadhatónak a disztribúciósan elkülönülő szóhasználatok koordinálása, például: *alpári stílus és kifejezés, nagy fokú érdeklődés és nyitottság*.

A dinamikuslexikon-elméletek (Pustejovksy, 1991) megkülönböztetik a véletlenszerű és a szabályos poliszémiát; az utóbbira jellemző, hogy lexikális szemantikai tulajdonságokkal körülírható szócsoportokra többé-kevésbé produktívan, szabályszerűen illeszkedő, megjósolható jelentésváltozással jár. A szabályos poliszémia gyakran észrevehetetlen a beszélők számára, és – legalábbis egyes eseteiben – meglepő módon elfogadható a poliszém jelentéseket előhívó kontextusok koordinálása (Schumacher, 2013). A kapott eredmények validálásán túl a nyelvmodellek használata segíthet az ilyen szabályos poliszém melléknevek, melléknévcsoportok azonosításában is.

3.4 A melléknévosztályok és a nyelvtechnológia

A nagy nyelvi modellek kapcsán alapvető kérdés, hogy „értik-e" a nyelvet vagy csak a szavak bonyolult korpuszbeli valószínűségi eloszlása alapján imitálják a nyelvértést. Ennek a kérdésnek az eldöntésében segíthet a modellek lexikális szemantikai képességeinek vizsgálata. Az ilyen vizsgálatok olyan adatbázisokon – ún. benchmarkkorpuszokon – alapulnak, amelyekben a célszavak jelentése valamilyen módon jelölve van. Az egyik legelterjedtebb ilyen jellegű benchmarkkorpusz a SuperGlue benchmarkadatbázis (Wang et al., 2020) részét képező Word-in-Context (WiC) kiértékelő korpusz (Pilehvar & Camacho-Collados, 2019). Érdekes módon az egyéb feladatokkal ellentétben a WiC meglehetősen nehéznek bizonyult még a GPT-3-nak is (Brown et al., 2020): few-shot tanulás esetén a GPT-3 válaszai a véletlenszerű válaszoknak megfelelő pontosságot mutattak. Ennek feltételezhető oka, hogy a WiC benchmarkkorpusz a Princeton WordNet jelentés-megkülönböztetésein alapul, így jelentésannotációja nem következetes; még az emberek sem értenek egyet a jelentések konkrét elkülönítésében minden esetben. A melléknevek gráfalapú reprezentációjának és a szaliens főnévi kontextusok kinyerésének egyik fontos célja, hogy egy olyan lexikális szemantikai tudást mérő benchmarkkorpuszt hozzunk létre, ahol (1) a jelentések elkülönítése nem a WordNet alapján, hanem korpusz alapján történik; (2) megjelennek a többjelentésűség olyan alesetei is, amelyek hatással lehetnek a poliszémia emberi észlelésére (pl. homonímia, reguláris vagy irreguláris poliszémia). Így elvárásaink szerint kiküszöbölhetjük, hogy olyan nyelvi tudást várjunk el a nagy nyelvi modellektől, amelyekkel mi magunk sem rendelkezünk.

Irodalom

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.* Retrieved from https://arxiv.org/abs/2005.14165

Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *arXiv preprint arXiv:1805.04032.* Retrieved from https://arxiv.org/abs/1805.04032

Héja, E., Ligeti-Nagy, N., Simon, L., & Lipp, V. (2023). An unsupervised approach to characterize the adjectival microstructure in a Hungarian monolingual explanatory dictionary. In Medveď, M., Měchura, M., Tiberius, C., Kosem, I., Kallas, J., Jakubíček, M., & Krek, S. (Eds.), *Proceedings of the eLex 2023 conference: Electronic lexicography in the 21st century (eLex 2023): Invisible lexicography* (pp. 150–167). Brno, Czech Republic: Lexical Computing.

Kiefer F. (2008). A melléknevek szemantikája. In Kiefer F. (szerk.), *Strukturális magyar nyelvtan 4. A szótár szerkezete.* Budapest: Akadémiai Kiadó.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR, abs/1301.3781.*

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*(10).

Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1* (Long and Short Papers) (pp. 1267–1273).

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics, 17*(4) (pp. 409–441).

Schumacher, P. (2013). When combinatorial processing results in reconceptualization: Toward a new approach of compositionality. *Frontiers in Psychology, 4* (Article 677).

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2020). SuperGLUE: A stickier benchmark for general purpose language understanding systems. *Advances in Neural Information Processing Systems, 32* (NeurIPS 2019).

Zwicky, A. M., & Sadock, J. M. (1975). Ambiguity tests and how to fail them. In *Syntax and Semantics volume 4* (pp. 1–36). Brill.