Magyar melléknevek poliszém jelentéseinek automatikus kinyerése gráfokkal

Héja Enikő, Ligeti-Nagy Noémi

Nyelvtudományi Kutatóközpont
1068 Budapest, Benczúr u. 33.
{heja.eniko, ligeti-nagy.noemi}@nytud.hu

Kivonat A cikk egy kutatás első fázisát mutatja be, amelynek célja interpretálható poliszém melléknévi jelentések automatikus kinyerése egynyelvű korpuszból egy felügyelet nélküli tanulási keretben. Kiindulásként 4 kritériumot határoztunk meg a jelentések elkülönítésére. A melléknéveket statikus szóbeágyazásokkal reprezentáltuk, majd ezekből egy szemantikai hasonlósági gráfot állítottunk elő. A jelentések elkülönítésére szolgáló kritériumokat ezen gráf részgráfjaival modelleztük. Végül egy részletes kvalitatív kiértékelés következett. Kutatásaink hosszabb távon hozzájárulnak a lexikográfusok és a nyelvészek munkájához, de a lexikális szemantikai információt tartalmazó NLP-célú *benchmark* adatbázisok létrehozását is segítik.
Kulcsszavak: poliszémia, WSI, gráf, klikk, főnévi kontextusok

1. Bevezetés

A kutatás célja a jelentés-megkülönböztetés mögött meghúzodó nyelvi intuíció modellezése és az elkülönített jelentések lehorgonyzása megfigyelhető adatokhoz. A kutatás jelentőségét az adja, hogy számos szerző szerint (pl. Véronis, 2003; Adamska-Sałaciak, 2006; Kuti és mtsai, 2010) a beszélők intuíciója nagy eltérést mutat az egyes szavakhoz tartozó jelentések particionálásában, amely komoly következményekkel bír a lexikográfiában, a lexikális szemantikában és a természetes nyelvfedolgozásban egyaránt. Különösen igaz ez a *poliszémia* esetében, hiszen a szójelentések automatikus detekcióját (*word sense induction* – WSI) célzó kutatások is elsősorban a homoním jelentések elkülönítésére koncentrálnak.

Így a jelen kiséreletben célunk az eltérő jelentések fogalmát úgy definiálni, hogy az elégséges kritériumként szolgáljon a jelentések megkülönböztetésére. Elvárásaink szerint ezek a kritériumok nemcsak az egyes jelentések kontextusokhoz való lehorgonyozását teszik majd lehetővé, hanem az így elkülönített jelentéseket szemantikai kategóriákhoz is kötik, hogy azok az emberi intuíció számára is megragadhatóvá váljanak. A kritériumokat felügyelet nélküli tanulási módszerrel modellezzük, hogy az introspekció szerepét minimalizáljuk.

A vizsgált mellékneveket statikus szóbeágyazásokkal reprezentáljuk. Bár ezen reprezentáció egyik fő hátrányaként azt szokás kiemelni, hogy nem alkalmas egy szóalak egyes jelentéseinek az elkülönítésére (vö. 'meaning conflation deficiency'

Camacho-Collados és Pilehvar, 2018) első eredményeink azt mutatják, hogy a gráf-alapú módszerek ezt a problémát képesek kiküszöbölni – legalábbis a poliszém jelentések esetében.[1]

A 2. fejezetben részletesen kifejtjük hipotézisünket és javaslunk 4 kritériumot a jelentések megkülönböztetésére, a 3. fejezetben bemutatjuk az alkalmazott módszert, a 4. fejezetben ismertetjük, hogyan validáltuk az automatikusan kinyert eredményeket főnévi kontextusaikkal. Az 5. fejezet témája a kiértékelés, végül a cikk egy összefoglalással zárul a 6. fejezetben.

2. A jelentések elkülönítésére szolgáló kritériumok

A jelentés hagyományos definíciója a jelentések azonosságát veszi alapul: a szinonímia definíciója[2] hosszú múltra tekint vissza (pl. Frege, 1892). Ez alapján azokat a jelentéseket tekintjük különbözőknek, amelyek nem szinonimák. A jelen gondolatmenetben megfordítjuk a definíciós sorrendet: nem a szinonímia fogalmából indulunk ki, hanem abból, hogy mikor eltérő két jelentés. Ez a döntés összhangban van azzal a ténnyel, hogy a szinonímia fogalma előfeltételezi az igazságfeltételek fogalmát, amely azonban a disztribúciós szemantika számára közvetlenül nem megragadható. Ezért nehéz valódi szinonima-osztályokat kinyerni a szövegből pusztán disztribúciós alapon: az automatikusan kinyert szinonima-osztályok sok esetben egyéb szűk szemantikai osztályokat is lefednek, pl. nemzetek neveit, színneveket, sőt hasonló disztribúciós tulajdonságokat mutató antonimákat is.

2.1. A majdnem-szinonímia fogalma

Ehhez bevezetjük a 'majdnem-szinonímia' (*near-synonymy* vö. Ploux és Victorri, 1998) fogalmát, amely a szinonímia egy kiterjesztett változata: akkor nevezünk két kifejezést majdnem-szinonimáknak, ha van a kontextusoknak egy olyan *korlátozott halmaza*, ahol a két kifejezés felcserélhető egymással az eredeti mondatok jelentésének megváltozása nélkül. Például, a *finom* és a *lágy* szinonimák számos zenéhez kapcsolódó főnév előtt, mint például *dallam*, *ritmus* vagy a *zene* maga, de nem szinonimák pl. ételek nevei előtt (pl. *lágy kenyér* ≠ *finom kenyér*). A fenti definíción túl a szűk szemantikai osztályok elemeit is majdnem-szinonimáknak tekintjük. A definíció ilyen irányú kiterjesztését az indokolja, hogy az eltérő szemantikai osztályokhoz való tartozás elégséges feltétele a jelentés megkülönböztetésnek, még akkor is, ha a szűk szemantikai osztályok elemeit egymással felcserélve nem feltétlenül őrizzük meg a mondatok igazságértékét.[3]

[1] Bár számos WSI-t célzó kutatás alapul a célszavakból készített gráf valamilyen lokális tulajdonságán (pl. Dorow és mtsai, 2004; Véronis, 2004; Biemann, 2006), tudomásunk szerint csak kevesen vizsgáltak statikus sűrű szóbeágyazásokból épített gráfokat (pl. Pelevina és mtsai, 2016).

[2] Két kifejezés akkor szinonima, ha minden kontextusban felcserélhetőek egymással úgy, hogy a mondatok igazságértéke nem változik.

[3] Pl. *fekete kalap* ≠ *szürke kalap*; *Fekete István* ≠ *Német István*.

2.2. A jelentés megkülönböztetés kritérumai

A fentiek alapján egy melléknév két jelentését akkor kell megkülönböztetni, ha:

1. Mindegyik melléknévi jelentéshez találunk (min. 1) majdnem-szinonimát.
2. Vannak olyan főnevek, amelyek grammatikus konstrukciókat alkotnak az eredeti melléknévvel és ennek majdnem-szinonimáival is.
3. Az egyes jelentéseket karakterizáló főnevek halmazai nem átfedőek.
4. A nem-átfedő főnév-halmazok egy vagy több szemantikai kategóriát képeznek „amelyek tükrözik a melléknevek szubszelekciós tulajdonságait" (vö. Pustejovsky, 1995).

A kritériumokat az 1. példával illusztráljuk, amely a *napfényes* melléknév automatikusan kinyert két jelentését tartalmazza. Láthatjuk, hogy mindkét jelentéshez tartozik egy-egy majdnem-szinonima: a *napsütéses* az első jelentéshez és a *napsütötte* a második jelentéshez. Listáztuk a majdnem-szinonima halmaz elemeivel grammatikus szerkezeteket alkotó főneveket is: *napfényes/napsütéses vasárnap*, *napfényes/napsütéses nap* stb., és *napfényes/napsütötte terület*, *napfényes/napsütötte terasz* stb. A két főnévhalmaz nem átfedő: nincs olyan, hogy *napsütötte nap* vagy *napsütéses terasz*. Végezetül, a szóban forgó főnevek egy szemantikai kategóriát alkotnak: az első jelentésre sepecifikus főnevek időintervallumokat, míg a második jelentésre specifikus főnevek fizikai kiterjedéssel rendelkező dolgokat jelölnek.

(1) 1. jelentés: *napfényes*, *napsütéses*
Főnévi kontextusok: *vasárnap*, *nap*
2. jelentés: *napfényes*, *napsütötte*
Főnévi kontextusok: *terület*, *sziget*, *oldal*, *terasz*

A kutatás jelen szakaszában azt vizsgáljuk, hogy a fenti kritériumok mennyiben alkalmasak a melléknévi jelentések elkülönítésére. A következő fejezetben egy gráf-alapú megközelítést javaslunk a kritériumok modellezésére.

3. Módszer

Módszerünk alapjául Ah-Pine és Jacquet (2009) szolgált, amennyiben a jelentésmegkülönböztetéseket **teljes részgráfokkal**, vagyis **klikkekkel** reprezentáljuk. Két fő különbséget azonban érdemes kiemelni: egyfelől, míg Ah-Pine és Jacquet (2009) kísérleteik során a névelemekre fókuszáltak, mi kizárólag az attributív mellékneveket vizsgáljuk. Ennek elsődleges oka, hogy előfeltevésünk szerint az attributív pozicióban lévő melléknevek jelentése egy viszonylag egyszerű jegykészlettel – a módosított főnévi csoport feje – leírható. Másfelől, a célszavakat statikus sűrű beágyazásokkal reprezentáljuk a gyakoriság-alapú ritka vektorok helyett.

3.1. Bemeneti adatok

A vizsgálandó melléknevek körét a 180 millió szavas MNSZ (Váradi, 2002) alapján határoztuk meg, ahol csak az alanyesetű mellékneveket vettük figyelembe,

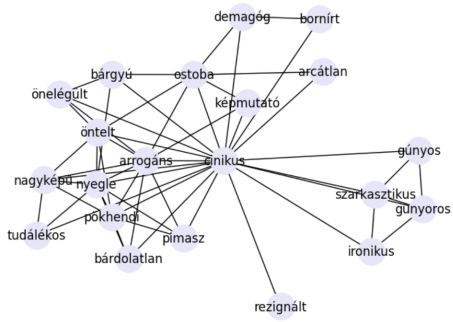
feltételezve, hogy a MN + FN szerkezetekben a MN mindig alanyesetű. Abból kiindulva, hogy a gyakoribb szavaknak több jelentése van, mint a ritka szavaknak (Zipf, 1949), csak a legalább 200-szor előforduló mellékneveket vizsgáltuk.

3.2. Reprezentációk

A melléknevek reprezentációja Bár Ah-Pine és Jacquet (2009) gyakoriság alapú vektorokkal reprezentálták a célkifejezéseket, mi statikus sűrű szóbeágyazásokat használtunk e célra.

Egy word2vec nyelvmodellt[4] (Mikolov és mtsai, 2013a,b) tanítottunk a Webcorpus 2.0 (Nemeskey, 2020) első 999 fájlján (21GB nyers szöveg). Kb. 170 M mondat,[5] a szövegek normalizált változata szolgált a tanítás bemenetéül. A 300 dimenziós CBOW vektorokat a Gensim Python csomag (Rehurek és Sojka, 2011) segítségével tanítottuk, ahol az ablak-méret 6 volt és a minimum gyakoriság 3. A tanítás során feltételeztük, hogy a szóalak által kódolt morfoszintaktikai információ hozzájárulhat a melléknevek jelentésének karakterizálásához. Ezt a hipotézisünket Novák és Novák (2018) is alátámasztja, akik különböző statikus szóbeágyazások szóhasonlósági feladatokon mért teljesítményének a vizsgálata során azt találták, hogy a melléknévi jelentéseket legjobban a szóalakokon tanított beágyazások reprezentálják. A tanítás eredményeképpen kb. 8,5M szóalakhoz rendeltünk szóbeágyazást.

A szemantikai hasonlóság reprezentációja Ahogy azt az 1. ábrán is láthatjuk, a melléknevek jelentéshasonlóságának gráf-alapú reprezentációja során olyan irányítatlan gráfokat generáltunk, amelyek csúcsai a melléknevek, élei (vagy azok hiánya) azt jelzik, hogy a szemantikai hasonlóság fennáll-e két melléknév között (vagy sem).

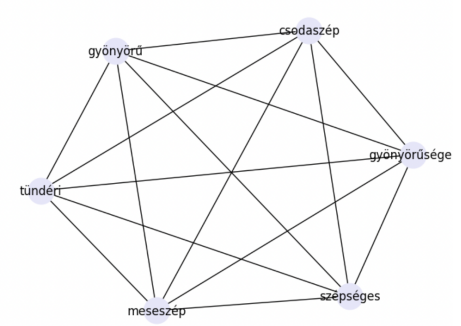


1. ábra: A *cinikus*-hoz szemantikailag hasonló melléknevek és azok közötti élek

[4] Elérhető itt: `https://nlp.nytud.hu/word2vec/cbow_3.tar.gz`

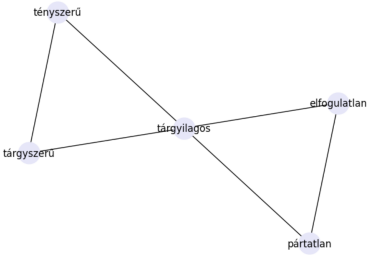
[5] Ugyanezekkel a hiperparaméterekkel tanítottunk egy word2vec modellt a teljes Webcorpus 2.0-n (kb. 591,4M mondat) is. Mindkét modellt kiértékeltük a Google Analogy Test Set magyar fordításán (Makrai, 2015). Mivel a nagyobb modell csak kisméretű javulást eredményezett a kisebb modellhez képest, a kisebb modellt használtuk a kisérleteink során.

A majdnem-szinonímák reprezentációja klikkekkel A 'különböző jelentés' fogalmát a majdnem-szinonima fogalmára támaszkodva kívánjuk megragadni. Ah-Pine és Jacquet (2009) nyomán a majdnem-szinonimákat klikkekkel reprezentáljuk, ezért a melléknévi gráfban teljes részgráfokat keresünk. Feltételezésünk szerint a klikkekkel jól modellezhetőek a majdnem-szinonima osztályok, mivel minden klikkhez tartozó melléknév disztribúciósan „nagyon hasonló" a klikk összes többi melléknevéhez (vö. 2. ábra).



2. ábra: Egy majdnem-szinonima halmaz klikk reprezentációja

A jelentés-elkülönítés reprezentációja: egy melléknév több klikk Összefoglalva az eddigieket: ha egy melléknévi lexéma több majdnem-szinonima osztályhoz tartozik, akkor több jelentéssel rendelkezik, és a majdnem-szinonima osztályokat klikkekkel modellezzük. Ebből következik, hogy egy melléknévnek akkor van több jelentése, ha több klikkhez is tartozik (vö. 3. ábra). Az így kinyert jelentéseket explicit kontextuális információhoz is le tudjuk horgonyozni, hiszen minden klikkhez kinyerhetőek azok a főnevek, amelyek a klikkben található összes melléknévvel előfordulnak.



3. ábra: A *tárgyilagos* két jelentésének reprezentációja klikkekkel

3.3. A klikkek kinyerése

***M_{sim}* ($n \times n$) hasonlósági mátrix** Az első lépésben egy M_{sim} hasonlósági mátrixot (ld. 4. ábra) hoztunk létre, amelynek soraiban és oszlopaiban is $A_{1...n}$ melléknevek word2vec reprezentációi ($a_{1...n}$) szerepelnek, ahol n a vizsgált melléknevek száma. A melléknevek szemantikai hasonlóságát a word2vec reprezentációjuk közötti koszinusz hasonlósággal számoltuk ki: $M_{sim}(a_i, a_j) = sim_{cos}(a_i, a_j)$

$$sim_{cos}(a_i, a_j) = \frac{\mathbf{a_i} \cdot \mathbf{a_j}}{||a_i|||a_j||} \tag{1}$$

	mesés	*káprázatos*	*varázslatos*
mesés	1	0,76	0,84
káprázatos	0,76	1	0,77
varázslatos	0,84	0,77	1

4. ábra: M_{sim} hasonlósági mátrix: $M_{a_i,a_j} = sim_{cos}(a_i, a_j)$.

***M_a* szomszédsági mátrix** A következő lépésben M_{sim} hasonlósági mátrixból előállítottuk M_a szomszédsági mátrixot (ld. 5. ábra), amely azt reprezentálja, hogy A_i és A_j melléknevek hasonlóak-e vagy sem: M_{sim}-ben szereplő $sim_{cos}(a_i, a_j)$ hasonlósági értékeket egy alkalmas K küszöbérték mellett "vágtuk": $sim_{cos}(a_i, a_j) \geq K$ esetén $M_a(a_i, a_j) = 1$, egyébként pedig $M_a(a_i, a_j) = 0$.

	mesés	*káprázatos*	*varázslatos*
mesés	1	0	1
káprázatos	0	1	0
varázslatos	1	0	1

5. ábra: M_a szomszédsági mátrix: $K \geq 0.8$.

Eredményül tehát M_a-t kaptuk, amely egy 0-t és 1-t tartalmazó négyzetes ($n \times n$) és szimmetrikus mátrix, amelynek a főátlójának minden eleme 1. Ezek a tulajdonságok tükrözik a hasonlósági reláció reflexív és szimmetrikus voltát.

Klikkek *M_a*-ban Ezentúl a M_a szomszédsági mátrixra úgy tekintünk, mint a melléknevek gráf reprezentációjára: a gráf csúcsai a melléknevek; $M_a(a_i, a_j) = 1$ azt jelöli, hogy a_i és a_j csúcsok éllel vannak összekötve, vagyis A_i és A_j melléknevek szemantikailag hasonlóak. Majd a gráf reprezentációban olyan mellékneveket keresünk, amelyek több klikkhez is tartoznak.

3.4. A főnévi kontextusok kinyerése – a klikkek validálása

Ebben a lépésben a melléknévi klikkek validálása történik azon klikk-specifikus főnevek halmazának előállításával, amelyekkel a klikkbe tartozó melléknevek előfordulnak. Várakozásaink szerint egy melléknév különböző jelentéseit a majdnemszinonimákon túl a vele abban a jelentésben együtt előforduló főnevek is jellemzik. Ezek a nem-átfedő főnévhalmazok explicit információval szolgálnak a jelentés-megkülönböztetésben. A klikk-specifikus főnévhalmazokat a következőképpen állítjuk elő:

1. Első lépésként összegyűjtjük az összes főnevet, amellyel egy melléknév előfordul. Ezt egy klikk minden melléknevére végrehajtjuk.[6]
2. A következő lépés a fent leírt módszerrel előállított halmazok metszetének meghatározása: ezek azok a főnevek, amelyek egy klikk minden melléknevével előfordulnak. Ha egy klikkre legalább egy ilyen főnév létezik, akkor az adott klikket jelentés-jelöltnek tekintjük.
3. Az 1. és 2. lépéseket minden klikkre elvégezzük, amelyekben az adott melléknév szerepel. Így mindegyik klikkre előáll egy főnévhalmaz.
4. Végül vesszük ezeket a halmazokat, és a metszetük tartalmát kidobjuk: csak azokat a főneveket tartjuk meg egy klikkhez, amelyek az adott klikkre kizárólagosak; nem fordulnak elő más klikkek főnévhalmazaiban.

3.5. Kiértékelés

Végül az eredményeket különböző paraméterbeállításokkal értékeltük ki. Mivel nem létezik hasonló adatbázis magyarra, kvalitatív kiértékelést végeztünk.

A kiértékelés során két célunk volt. Egyfelől, alá kívántuk támasztani, hogy az általunk javasolt jelentés-megkülönböztetést célzó kritériumok ténylegesen használhatóak erre a feladatra. Másfelől, az elkülönített melléknévi jelentéseket egy adatbázisban listáztuk a releváns kontextusaikkal együtt. Az evaluálás két lépésben történt: először az automatikusan kinyert melléknévi klikkek szemantikai tulajdonságait vizsgáltuk meg, majd a főnévi kontextusokra fókuszáltunk.

3.6. Paraméter-beállítás

A kiértékelés során azt találtuk, hogy 3 paraméternek van komoly hatása az eredményekre:

(i) a vizsgált melléknevek minimum-gyakorisága az MNSZ-ben,[7]
(ii) a K küszöbérték,
(iii) a főnevek melléknevekkel való együttes előfordulásának minimum gyakorisága a klikkvalidációs lépésben.

[6] Mindehhez a korpuszt a Magyar nemzeti szövegtár (Oravecz és mtsai, 2014) egy 91.4 millió tokenes részkorpusza jelentette, amelyet kifejezetten ennek a vizsgálatnak a céljából állítottunk össze.

[7] Ezt a paramétert jelen cikkben nem tárgyaljuk, de nyilvánvalóan nagy hatása van az eredmények fedésére.

A K küszöbérték hatása
Érdekes módon a K küszöbérték nemcsak a kinyert klikkek számára volt hatással, hanem a klikkbe tartozó melléknevek jelentésére is. Például, a legalább 200-szor előforduló melléknevek esetében a $K = 0,9$ beállítás csak egy pár melléknévi klikket eredményezett: 8 melléknév tartozott legalább 2 klikkhez és csak 2 klikket tudtunk validálni. A klikkek kizárólag számokra, hónapok és napok neveire referáltak, így nem különösebben érdekesek a kutatási célunk szempontjából. Másfelől alacsonyabb K küszöbértéket alkalmazva ($K = 0,7$) 187 különböző melléknevet kaptunk eredményül, amelyek mindegyike több validált klikkhez is tartozott. A $K = 0,7$-es küszöbérték mellett 3847 izolált csúcsot kaptunk, és 1085 olyan csúcsot, amelyekre igaz, hogy pontosan egy szomszédos csúcsa van. Ez azt jelenti, hogy a melléknevek 82%-át már eleve kizártuk a vizsgálatból. Így a $K = 0,7$-es küszöbérték valószínűleg túl magas.
A főnévi kontextus gyakoriságának a hatása
A főnévi kontextus minimum elvárt gyakoriságát ($Freq_n$) szintén figyelembe vettük a kísérleteink során. $Freq_{ADJ} = 200, K = 0,7$ rögzített paraméterek mellett három értékét vizsgáltuk. Az első esetben egy klikket akkor tekintettünk validnak, ha volt legalább 1 olyan főnév, amely a validáló korpuszban legalább 5-ször előfordult a klikkbe tartozó minden melléknévvel ($Freq_n \geq 5$). Mivel így csak kevés validált klikket kaptunk eredményül, ezt az értéket túl magasnak ítéltük. A fedés lehető legnagyobb növelése céjából a $Freq_n = 2$ értéket választottuk.[8] Ebben az esetben a 6042 input melléknévből 446 olyan melléknév maradt, amely több validált klikkhez is tartozott. Mivel ebben a kutatási szakaszban elsősorban a módszer alkalmazhatóságnak, pontosságnak a vizsgálata a cél, az alacsony fedés ellenére a további kiértékelési lépéseket ezeken a melléknévi klikkeken fogjuk végezni.

4. A releváns jelentések

A jelen fejezetben azokat a nyelvészeti megfontolásokat ismertetjük, amelyeket a kiértékelés során figyelembe vettünk.

4.1. Produktivitás

Az egyes jelentések elkülönítése során megkülönböztethetjük egy kifejezés kollokációs és produktív(abb) használatait. A jelen kutatásban a produktivitásra egy skálaként tekintünk. A skála egyik végén a szigorú értelemben vett kollokációkat találjuk, ahol az attributív melléknév és a módosított főnév egyaránt rögzítve vannak lexikailag. Ebben az esetben a kifejezés jelentése nem kompozicionális: egyik komponens sem helyettesíthető egy majdnem-szinonimával, úgy hogy az eredeti kifejezés jelentése (i) ne változzon, (ii) előre megjósolható módon változzon (pl. *fehér zaj*, *fekete doboz*).

[8] Vizsgáltuk a $Freq_n = 1$ paraméterbeállítást is: ez a nagyobb fedés mellett nehezebben interpretálható eredményeket adott – feltételezésünk szerint azért, mert ez az érték nem alkalmas a szisztematikus és véletlen hiányok szétválasztására.

Annak ellenére, hogy a kollokációk szolgálhatnak az elkülönítendő jelentések forrásául, a jelen WSI feladatban a 'félig-kompozicionális' szerkezetekre fókuszáltunk, ahol a kompozicionalitás a melléknevek és főnevek egy meghatározott körén működik (pl. *fehér/szürke/fekete gazdaság*). Ezek a kifejezések nem tekinthetőek szigorú értelemben kollokációnak, hiszen ezek a színnevek a *gazdaság* szó kontextusában egy új dimenziót vezetnek be: annak a mértékét, hogy egy gazdasági ág mennyiben kerüli ki a legális üzleti forgalmat.

4.2. Szubkategorizáció

A legérdekesebb esetek azok, ahol a főnevek egy vagy több szemantikai osztályt alkotnak, és a klikket alkotó melléknevek szinonímák a szemantikai osztályba tartozó főnevek előtt. Ezekben az esetekben a melléknév szubkategorizálja a főnévi kontextust (vö. Pustejovsky, 1995). Például a *könnyű*, *komoly*, *szép*, *éles*, *finom* melléknevek egyes jelentései elkülöníthetőek az alapján, hogy milyen szemantikai főnévcsoportokat szubkategorizálnak. A *könnyű* különböző jelentéseket hordoz a fizikai tárgyakra (*könnyű táska*), a ruhákra (*könnyű kabát*), az ételekre referáló (*könnyű vacsora*) főnevek előtt, és mást jelent az olyan főnevek előtt, mint *válasz*, *feladat*, *megoldás*.

Az egyes főnévi szemantikai osztályok mérete eltérő lehet: a produktivitási skála másik végén találjuk azokat a melléknévi majdnem-szinoníma osztályokat, amelyek sokkal több elemet tartalmazó szemantikai osztályokat szubkategorizálnak; ezek szintén fontosak a kutatás szempontjából. Például az automatikusan kinyert klikkek alapján a *szomorú* különböző jelentésekkel bír attól függően, hogy a módosított főnév emberekre vagy időtartamokra referál-e.[9] A kinyert klikkek alapján mondhatjuk azt, hogy *szomorú* [*időszak, év, nap*] és *gyászos* [*időszak, év, nap*], de nincs olyan, hogy *bánatos* [*időszak, év, nap*] és olyan sem, hogy *gyászos* [*lány, ember*].

(2) 1. klikk: *szomorú*, *gyászos*
Főnévi kontextusok: *időszak*, *év*, *nap*
2. klikk: *szomorú*, *bánatos*
Főnévi kontextusok: *lány*, *ember*

5. Kiértékelés

5.1. Az automatikusan kinyert melléknévi klikkek szemantikai osztályozása

Szűk szemantikai osztályok Az első szembetűnő probléma az volt, hogy nem minden melléknév bizonyult egyformán relevánsnak a jelentés-elkülönítés szempontjából. Például a dátumok (napok, hónapok) és mértékegységek legtöbb esetben nem mutattak semmilyen érdekes tulajdonságot. Ha több klikkhez is

[9] Érdekes, hogy a *vidám*-nak is két jelentése különült el: pontosan ugyanazokat a főnévi szemantikai osztályokat szubkategorizálta, mint a *szomorú*.

tartoztak egyszerre, akkor is általában ugyanazzal a jelentéssel szerepeltek. Hipotézisünk szerint ezen szűk szemantikai osztályok eltérő mérete és az elemeik közötti eltérő távolságok miatt a szűk szemantikai osztályokhoz tartozó melléknevek nem csoportosíthatóak következetesen egy-egy klikkbe függetlenül a paraméterek aktuális értékeitől. Egy másik indok, hogy ezeket a mellékneveket kizárjuk a további vizsgálódásból az, hogy az ilyen melléknevek jelentése általában szemantikailag egyértelmű, egyjelentésűek.[10] Néhány kivételtől eltekintve azt találtuk, hogy a számnevek, nap- és hónapnevek, színnevek, mértékegységek és különféle nemzeti valuták esetében a főnévi kontextusok nem szolgáltatnak elég alapot a jelentések elkülönítésére.

Névelemek A melléknevek egy másik szemantikai osztályát a névelemek alkották: elsősorban ország- és városnevek valamint vezetéknevek. Az elgondolkodtató csoportosítások ellenére a jelen vizsgálatban ezeket a mellékneveket is figyelmen kívül hagytuk, hiszen érdeklődésünk központjában a lexikális jelentés áll, és a névelemek esetében a klikkbe tartozást elsősorban faktuális tényezők határozzák meg (pl. az *egri* melléknév két klikkhez tartozott: [*egri*, *soproni*, *veszprémi*] illetve [*egri*, *esztergomi*], ahol az első klikk vélhetően borvidékekre, míg a második érsekségekre referáló névelemeket tartalmaz). A kézi kiértékelés egy érdekes eredménye az volt, hogy a $6k$ ablakméretű word2vec reprezentáció meglehetősen hatékony eszköznek bizonyult a szűk szemantikai osztályok és a névelemek klikkjeinek kinyerésében: 446 több klikkhez is tartozó melléknévből 99 volt névelem, 28 tartozott a mértékegységek csoportjába és 11 melléknév referált számnevekre.

Negatív emotív szemantikai tartalmú fokozó elemek Az emotív fokozó elemekkel kapcsolatban is azt találtuk, hogy a klikkek nem tükröznek eltérő jelentéseket.

(3) 1. klikk: *borzalmas iszonyatos rettenetes*
Főnévi kontextus: *szenvedés*, *kép*, *körülmény*

2. klikk: *borzalmas*, *félelmetes*, *rettenetes*, *szörnyű*
Főnévi kontextus: *látvány*, *nap*, *érzés*

3. klikk: *borzalmas*, *borzasztó*, *rettenetes*, *szörnyű*, *rémes*
Főnévi kontextus: *emlék*, *élmény*

Míg a klikkek azt támasztják alá, hogy az emotív fokozó elemek egy koherens szemantikai osztályt alkotnak a mellékneveken belül, sem a klikkek maguk, sem a főnévi kontextusok nem nyújtanak elég alapot arra, hogy az egyes klikkek jelentéseit elkülönítsük egymástól.

5.2. A főnévi kontextusok kvalitatív kiértékelése

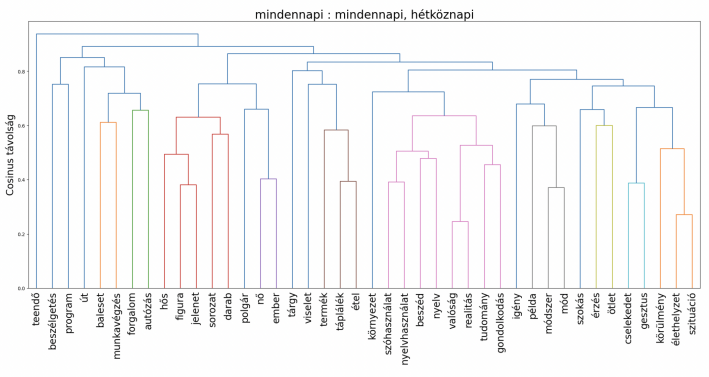
Miután kizártuk a nem-releváns eseteket (kb. 240 melléknevet), egy részletesebb kiértékelést is végeztünk, amelynek célja egy olyan melléknévi adatbázis létreho-

[10] Bár bizonyos esetekben a szűk szemantikai osztályokhoz tartozó melléknevek is lehetnek poliszémek (pl. a *fekete*, *szürke*, *fehér*).

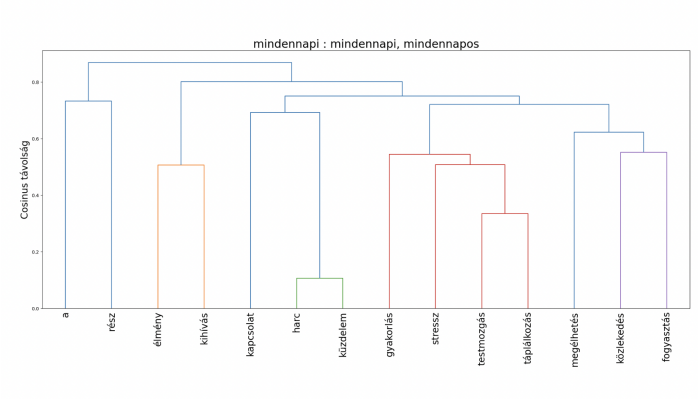
zása volt, amelyben a jelentések elkülönítése a majdnem-szinonimák által motivált, és az egyes jelentések expliciten megadott a főnévi kontextusokhoz vannak kötve. Azt is vizsgáltuk, hogy a szóban forgó főnevek segítenek-e a jelentéskülönbségek fogalmi megragadásában. Így a maradék klikkeket kézzel ellenőriztük, a készülő adatbázisba klikkenként legfeljebb 5 főnév került, úgy hogy törekedtünk arra, hogy a választott főnevek jellemzőek legyenek az adott jelentésre az emberi észlelés számára. A klikkek validálása és a releváns főnevek kiválasztása során az alábbi lépéseket követtük:

1. A főnévi kontextusok word2vec reprezentációját használtuk, amelyeket a 3.1 fejezetben leírtak alapján hoztunk létre.
2. A főnévi vektorokat klaszteresztük egy hierarchikus agglomeratív klaszterezési eljárással (*average linkage*, koszinusz távolság), melynek célja a szubkategorizációs mintázatok automatikus kinyerése volt.

Például a *mindennapi* melléknév két klikkhez is tartozott: a 6. és a 7. ábrák ábrázolják a főnévi kontextusok dendogramjait. A majdnem-szinonimák már megvilágítják a *mindennapi* melléknév két jelentését (*hétköznapi* ill. *mindennapos*), amelyek közül az egyik a *normális/bevett/megszokott* míg a másik jelentés szabályos időközönként űzött tevékenységekre utal. A dendogramok alapján levonhatjuk azt a következtetést, hogy például a nyelvhez kapcsolódó dolgok (pl. *szóhasználat*, *nyelvhasználat*) inkább normálisak vagy megszokottak, mintsem periódikusan mindennap ismétlődőek, míg a *gyakorlás* vagy a *testmozgás* szabályos időközönként végzett (végzendő) tevékenységek, és nem szükségszerűen megszokott vagy bevett dolgok. Így a dendogram ágai azt jelzik, hogy a melléknév egyes jelentései milyen szemantikai osztályba tartozó főneveket szubkategorizálnak.



6. ábra: A [*mindennapi*, *hétköznapi*] melléknévi klikk főnévi kontextusainak klaszterei



7. ábra: A [*mindennapi*, *mindennapos*] melléknévi klikk főnévi kontextusainak klaszterei

A 446 melléknévből 53 olyan melléknév maradt, amelyek validált klikkekhez tartoztak: 118 klikkhez összesen. Ez azt jelenti, hogy az 53 melléknévnek 118 jelentését különítettük el. Ezek elérhetőek a GitHub-on; `https://github.com/nytud/HuWiC`. A kvalitatív kiértékelés során azt találtuk, hogy a javasolt felügyelet-nélküli tanuláson alapuló módszer meglepően érdekes jelentés megkülönböztetéseket tesz, amelyek az emberi intuíció számára nem közvetlenül megragadhatóak. Így az alacsony fedés ellenére azt gondoljuk, hogy az ez irányú kutatásokat érdemes a jövőben folytatni.

6. Összegzés és további feladatok

A cikkben egy folyamatban lévő kutatás első lépését ismertettük, amelynek célja, hogy egy felügyelet-nélküli tanítási algoritmust alkalmazva magyar melléknevek interpretálható poliszém jelentéseit nyerje ki egynyelvű korpuszból. Ez a munka hozzájárulhat a lexikográfusok, nyelvészek munkájához, és segíti a kapcsolódó benchmarkok építését is NLP célokra. Első lépésben 4 kritériumot támasztottunk a jelentések elkülönítésére, amelyeket a következő lépésben implementáltunk. Végül egy részletes kvalitatív kiértékelés következett, amely egyfelől az automatikusan kinyert jelentés-jelöltek relevanciáját vizsgálta, másfelől a főnévi kontextusok szerepét vizsgálta a jelentések elkülönítésében. A kiértékelés azt mutatta, hogy a kinyert jelentések sok esetben nagyban segítettek a jelentések interpretálható elkülönítésében, méghozzá az alkalmazott módszertan miatt oly módon, amely véleményünk szerint sem introspektív, sem korpusz-alapú módszertan számára nem közvetlenül hozzáférhető. Az első kísérletek azt mutatták, hogy a fedés alacsony, így a jövőben elsődleges célunk ennek növelése. Ezt első körben a paraméterek változtatásával kívánjuk elérni, de vizsgálni kívánunk a klikkeknél kevésbé megszorított részgráfokat is. További feladat a poliszémiában szerepet játszó főnévi kontextusok szemantikai osztályainak feltárása is.

Hivatkozások

Adamska-Sałaciak, A.: Meaning and the Bilingual Dictionary. The Case of English and Polish. (Polish Studies in English Language and Literature 18). Peter Lang, Frankfurt am Main (2006)

Ah-Pine, J., Jacquet, G.: Clique-Based Clustering for Improving Named Entity Recognition Systems. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). pp. 51–59. Association for Computational Linguistics, Athens, Greece (Mar 2009), https://aclanthology.org/E09-1007

Biemann, C.: Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In: Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing. pp. 73–80. Association for Computational Linguistics, New York City (Jun 2006), https://aclanthology.org/W06-3812

Camacho-Collados, J., Pilehvar, M.T.: From Word to Sense Embeddings: A Survey on Vector Representations of Meaning (2018), https://arxiv.org/abs/1805.04032

Dorow, B., Widdows, D., Ling, K., Eckmann, J.P., Sergi, D., Moses, E.: Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination (2004), https://arxiv.org/abs/cond-mat/0403693

Frege, G.: Uber Sinn und Bedeutung. In: Textor, M. (szerk.) Funktion - Begriff - Bedeutung, Sammlung Philosophie, vol. 4. Vandenhoeck & Ruprecht, Göttingen (1892)

Kuti, J., Héja, E., Sass, B.: Sense disambiguation - 'Ambiguous sensation'? In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). p. 23–30. la Valetta, Malta (2010)

Makrai, M.: Comparison of distributed language models on medium-resourced languages. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015) (2015)

Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space (2013a), https://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26 (10 2013b)

Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D.-értekezés, Eötvös Loránd University (2020)

Novák, A., Novák, B.: POS, ANA and LEM: Word embeddings built from annotated corpora perform better. In: Computational Linguistics and Intelligent Text Processing: 17th International Conference, (CICLing 2018). Springer International Publishing, Cham, Hanoi, Vietnam (2018)

Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1719–1723. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf

Pelevina, M., Arefiev, N., Biemann, C., Panchenko, A.: Making Sense of Word Embeddings. In: Proceedings of the 1st Workshop on Representation Learning for NLP. pp. 174–183. Association for Computational Linguistics, Berlin, Germany (Aug 2016), https://aclanthology.org/W16-1620

Ploux, S., Victorri, B.: Construction d'espaces sémantiques a l'aide de dictionnaires de synonymes. Traitement automatique des langues 1(39), 146–162 (1998)

Pustejovsky, J.: The Generative Lexicon. MIT Press, Cambridge, MA (1995)

Rehurek, R., Sojka, P.: Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3(2) (2011)

Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas. pp. 385–389 (2002)

Véronis, J.: Sense tagging: does it make sense? In: Wilson, A., Rayson, P., McEnery, T. (szerk.) Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech. Peter Lang, Frankfurt (2003)

Véronis, J.: HyperLex: lexical cartography for information retrieval. Computer Speech & Language 18(3), 223–252 (2004), http://dblp.uni-trier.de/db/journals/csl/csl18.html#Veronis04

Zipf, G.K.: Human Behaviour and the Principle of Least-Effort. MA Thesis (1949)