Furkó Péter [1] *– Tóth Ágoston* [2]

MIT TUD A MESTERSÉGES INTELLIGENCIA A DISKURZUSJELÖLŐKRŐL: ESETTANULMÁNYOK A USAS ÉS A BERT EGYÉRTELMŰSÍTÉSI MÓDSZEREIRŐL

[1] Károli Gáspár Református Egyetem, Angol Nyelvészeti Tanszék
[2] Debreceni Egyetem, Angol Nyelvészeti Tanszék

furko.peter@kre.hu, toth.agoston@arts.unideb.hu

Bevezetés

Schiffrin (1987) óta a diskurzusjelölő-kutatások száma – mind az elméleti, mind az empirikus jellegű munkák tekintetében – ugrásszerűen megnövekedett, már a XX. század végére „virágzó üzletággá" fejlődött (l. pl. Fraser 1999). A diskurzusjelölőkkel kapcsolatos megfigyelések fényében számos nyelvészeti fogalmat újra kellett értékelni, köztük a grammatikalizáció folyamatát és velejáróit, a pragmatika és a szemantika határterületeit, és nem utolsósorban a korpusznyelvészet, közelebbről a diskurzusannotáció módszertanát, olyannyira, hogy a diskurzusjelölő-kutatást számos szerző a „pragmatika zászlóshajójának" tartja (vö. Hansen 2006, 2020, idézi pl. Fischer 2014, Crible–Cuenca 2017).

A diskurzusjelölőket számos nyelvben és műfajban, különböző elméleti és leíró keretekben vizsgálták, mégis fontos kérdésekben a mai napig bizonytalanság mutatkozik a szakirodalomban. Nincs általánosan elfogadott terminológia és osztályozás, továbbá nem létezik konszenzus a formai, szemantikai és pragmatikai tulajdonságok vonatkozásában, emellett már a definiálás fázisában következetlenségek mutatkoznak olyan meghatározó munkákban, mint például Schiffrin (1987), Brinton (1996), Fraser (1999), valamint Fischer (2006). A kortárs diskurzusjelölő-kutatások neuralgikus pontjai között említhetjük a jelölők extrém multifunkcionalitását, stigmatizáltságukat (töltelékelemként való felfogásukat), fordításuk és szemantikai egyértelműsítésük nehézségeit.

A diskurzusjelölők körül kialakult terminológiai zűrzavarban újabb neuralgikus pont mutatkozik meg, mely a kutatások fókuszából és az adott adathalmaz jellegéből adódik. A szakirodalomban gyakran ugyanazok a nyelvi elemek esetenként diskurzuscsatoló elemként (*discourse connective*), diskurzusoperátorként (*discourse operator*), pragmatikai jelölőként (*pragmatic marker*), diskurzusjelölőként (*discourse marker*), szünetjelzőként (*pause marker*), diskurzuspartikulaként (*discourse particle*), indulatszóként (*interjection*), interakciós jelzőként (*interactional signal*), célzást kifejező szóként/fráziskén (*cue word/phrase*), a

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

pragmatikai erőt megváltoztató elemként (*pragmatic force modifier*) vagy éppen pragmatikai kifejezésként (*pragmatic expression*) szerepelnek. A terminusok annak függvényében változnak, hogy a szerző célja a koherencia, az interpretációs folyamat, az interakció, a retorikai struktúra stb. vizsgálata a megfelelő elméleti keret (pl. relevanciaelmélet, beszédaktus-elmélet, grice-i pragmatika), vagy annotációs séma (RST, PDTB, CCR stb.) szemszögéből, spontán beszélt nyelvi, tervezett, félig tervezett stb. megnyilatkozásokból álló adathalmazokra vonatkozóan.

Tanulmányunkban azt vizsgáljuk, hogy a diskurzusjelölők felismerése és osztályokba sorolása számítógépes nyelvészeti eszközökkel hogyan valósítható meg, és ennek során milyen hibákra, buktatókra kell felkészülnünk. Ehhez először egy rövid áttekintést adunk a diskurzusjelölők (DJ-k) kritériumairól, különös tekintettel azokra, amelyek relevánsak az automatizált egyértelműsítés szempontjából. Ezután megvizsgáljuk a UCREL Semantic Analysis System (USAS, vö. Rayson et al. 2004) és a Google keresőoptimalizálásra is használt BERT (Devlin et al. 2019) alkalmazhatóságát. Ezek az eszközök céljaikban és működésükben jelentősen különböznek, hiszen az előbbi egy statisztikai és szabályalapú rendszer, mely előre rögzített nyelvészeti osztályokba sorolásra van felkészítve, míg a másik egy olyan neurális gépi tanulási rendszer, mely nagy mennyiségű címkézetlen nyelvi adat megfelelő feldolgozásán keresztül a számítógépes szemantika univerzális eszközévé vált, és a nyelvi mesterséges intelligencia ismert, látványos feladatainak (pl. gépi fordítás, ember-gép társalgás) megoldója. A BERT-tel kapcsolatban számunkra a kérdés – ezen a ponton – az volt, hogy a benne tárolt tudás felhasználható-e a céljainkra, tekintettel arra, hogy semmilyen explicit nyelvi kategorizálásra nincs optimalizálva, még szófaji besorolást sem kapnak a tanító- és a tesztkorpusz szavai sem a rendszer betanítása, sem annak használata során, pragmatikai jellemzők beazonosításáról nem is beszélve. Azonban a rendszer kialakítása és az abban tárolt, elemzés nélküli nyelvi tudás – a későbbiekben ismertetett módon – lehetővé teszik a szavak disztribúciós tulajdonságainak automatizált felderítését a betanítás alatt, a rendszer alkalmazása során pedig a szavak használatának kontextusfüggő jellemzését nyerhetjük ki, ami azzal kecsegtet, hogy a szavak DJ használatáról releváns adatokat kapunk. Mindkét rendszer (USAS, BERT) vizsgálatára, mindkét esettanulmányunkra igaz, hogy az egyes lexikai elemek diskurzusjelölő (DJ) és nem-DJ használata közötti egyértelműsítésnek és a marginalitásnak vizsgálatával a beszélt angol nyelvben előforduló DJ-k automatizált, illetve számítógéppel segített manuális azonosítását mutatjuk be a tapasztalt hibák, hiányosságok feltérképezése mellett.

A diskurzusjelölők jellemzői, az azonosítás problematikája

Mielőtt a számítógépes nyelvészeti megoldások használatát vizsgálnánk, tekintsük át a diskurzusjelölők jellemzőit, használatuk legfontosabb jegyeit.

A diskurzusjelölők ***formális-szintaktikai*** jellemzői között a következőket találjuk a releváns szakirodalomban (részleteket, hivatkozásokat l. 1. táblázat):

- *opcionalitás (szintaktikai leválaszthatóság)*

A legtöbb DJ-ként értelmezett elem (kifejezetten és kizárólag szintaktikai értelemben) opcionális, azaz elhagyásukkal általában nem sérül a grammatikai struktúra/grammatikalitás (pl. *Hát, nem tudom, mi a helyzet…* vs. *Nem tudom, mi a helyzet…).*

- *változatos forráskategóriák*

A diskurzusjelölők eredetük tekintetében heterogének, különböző szintaktikai osztályokhoz tartozó elemekből alakulhatnak ki, például kötőszókból (pl. *and*[1]*, but, because*), indulatszókból (pl. *oh, ah, huh*), igékből (pl. *say, look, see*), határozószókból (pl. *well*), melléknevekből (pl. *fine, right*), tagmondatokból (pl. *you see, I mean, you know*), elöljárós szószerkezetekből (pl. *in other words*) stb.

- *invariáns alak*

Kroon (1995) a diskurzusjelölők invariáns alakját definíciós kritériumnak tartja. Ez egyrészt azt jelenti, hogy a több elemből álló diskurzusjelölők is egyetlen egységként raktározódnak a beszélő mentális lexikonában, másrészt jellemzően nem léteznek alternatív alakváltozataik. A Kroon által említett *in other words vs. *in some other words* esetében ez indokolt, viszont az *I mean* vs. *you mean* vonatkozásában máris szürke területre érkezünk, hiszen az angol-magyar párhuzamos korpuszokban összefüggést találunk ezen két forrásnyelvi DJ és a célnyelvi, újrafogalmazást jelölő diskurzusjelölők (*mármint, vagyis*) között (vö. Furkó 2019).

- *szegmenskezdő pozíció*

A diskurzusjelölők leggyakrabban a (beszéd)forduló, illetve a megnyilatkozás elején fordulnak elő. Amennyiben a szintaktikai értelemben vett mondat/tagmondat végén találjuk őket, legtöbbször a korábbi nyelvállapotra jellemző, „magközeli" jelentéssel bírnak (vö. Fraser 1999: 945).

A **formális** jellemzők között említhetjük még:

- *fonológiai redukció*

A diskurzusjelölők gyakran rendelkeznek gyengén ejtett, vagy rövidített, illetve redukált fonológiai alakkal pl. *you know – y'know/ya kna, of course – 'course*.

[1] A multifunkcionalitás és a nyelvközi megfeleltethetőség hiánya miatt nem célunk magyar megfelelőket megadni az angol példák esetében.

Elsősorban a **stilisztikai** jelleggel összefüggésbe hozható jegyek:

- *csoportosulás (clustering)*

Gyakran több diskurzusjelölő fordul elő egymás közvetlen közelében, sokszor egymás funkcióját erősítve (*OK then*), máskor teljesen új funkciót létrehozva (pl. *yeah well*). Egyes esetekben azonban csupán az a funkciójuk, hogy a beszélő időt nyerhessen ahhoz, hogy a diskurzusjelölőket követő megnyilatkozást megfogalmazhassa/a megfelelő szót, kifejezést mentális lexikonából előhívhassa (lexical search function), mindemellett hezitálást is kifejezhetnek.

- *felcserélhetőség*

Bizonyos funkcióikban (főleg a fentebb említett „időnyeréssel" kapcsolatba hozhatókkal) nem érzékelhető jelentős funkcióvesztés vagy funkcióváltozás akkor sem, ha egy adott diskurzusjelölőt egy másikkal helyettesítünk (pl. *hát, tulajdonképpen* vs. *hát, hogyismondjam*).

- *oralitás – (bizonyos nyelvváltozatok esetében) gyakori előfordulás – stigmatizáció*

Mivel a szakirodalomban tárgyalt legtöbb diskurzusjelölő elsődlegesen a beszélt nyelvben fordul elő, sokan az oralitást is ezen nyelvi elemek szembetűnő jellegzetességei közé sorolják. Egyrészt azonban az írott és a beszélt (spontán/tervezett) nyelvváltozat között nem figyelhető meg éles határvonal, azok jellemzően kontinuumot alkotnak, másrészt a gyakori használat a pragmatikalizációs folyamat szükséges összetevője és katalizátora, még a lexikális keresés funkció (l. fentebb) esetében sem indokolt akár túlzott használatról beszélni, akár stigmatizációt sugalló megállapításokat tenni.

A **szemantikai-pragmatikai** jellemzők közül a következőket vehetjük számba:

- *nem propozicionális jelentés*

Általánosan elfogadott tény, hogy bizonyos diskurzusjelölők (pl. *well*, *however*) nincsenek hatással a mondat igazságfeltételeire, vagyis nem érintik a gazdaegység propozicionális jelentését. Más elemek (pl. *I think*) esetében azonban vita alakult ki a szakirodalomban azzal kapcsolatban, hogy (akár szinkrón, akár diakrón perspektívából nézve) rendelkeznek-e propozicionális jelentéssel.

- *multifunkcionalitás (szemantikai leírások szerint poliszémia)*

Köztudott, hogy a legtöbb nyelvi elem több, egymáshoz kapcsolódó vagy egymástól független jelentéssel bír, ez azonban az esetek legnagyobb részében nem vezet egy adott szövegkörnyezetben/megnyilatkozásban is többértelműséghez. A diskurzusjelölőkre azonban nem csupán az jellemző, hogy különböző szövegkörnyezetekben a (textuális, interakcionális, interperszonális, attitűdjelölő

stb.) funkciók széles körét látják el, hanem az is, hogy egy adott kontextusban egy adott gazdaegységen belül is több módon befolyásolják a megnyilatkozás értelmezését. A poliszémia-elképzelés szerint van magjelentésük, a kontextus csupán specifikusabb értelmezésüket alakítja ki. A homonímia-elképzelés szerint egy adott diskurzusjelölőnek egymással össze nem függő jelentései vannak.

- *extrém kontextusfüggőség/indexikalitás*

Sokak szerint a diskurzusjelölők a deiktikus kifejezésekhez hasonló szerepet látnak el – Levinson (2004) például diskurzusmutató szóknak nevezi őket –, mivel funkciójuk a kommunikatív szituáció egyes elemeitől függ, azokkal indexikális kapcsolatot létesítenek. A kapcsolat jellege azonban egyes esetekben összefüggésbe hozható magjelentésükkel (*core meaning*, vö. Fraser 1999:945), olyannyira, hogy az inherens jelentés együtt jelenik meg a kontextusfüggő funkcióval.

- *konnektivitás*

Mint fentebb láttuk, ezt a jellemzőt is különbözőképpen lehet értelmezni: Fraser szerint például csak azokat az elemeket nevezhetjük diskurzusjelölőknek, amelyek diskurzusegységeket kapcsolnak össze, elsődleges funkciójuk konnektív, összekapcsoló jellegű.

- *változó és funkcionális hatókör*

A diskurzusjelölők hatókörébe esetenként egyetlen lexikális elem, máskor a teljes megnyilatkozás, forduló, korábban elhangzott beszélgetés/leírt szöveg is tartozhat.

Összefoglalva tehát az egyes definíciókban az alábbi jellemzőket találjuk:

	seq.	**context**	**oral.**	**synt.**	**poly-func.**	**scope**	**non-prop.**	**inv.**
Schiffrin (1987)	x	x	x	x				
Blakemore (1987, 2002)								
Redeker (1990)	x		(x)					
Redeker (1990)	x	x						
Kroon (1995)	x	x						x
Knott and Sanders (1998)	x							

Risselada and Spooren (1998)	x			x				
Romaine and Lange (1998)	x		(x)					
Fraser (1999)	x	x		x			x	
Andersen (2001)	x		x					
González (2004)					x			
Hansen (2006)	x		(x)			x	x	
Crible (2017)				x	x	x		

1. táblázat: A diskurzusjelölők jellemzőinek megjelenése az egyes szerzők definícióiban[2]

Mint azt az 1. táblázatból láthatjuk, Crible (2017) a diskurzusjelölők jellemzői közül a szintaktikai leválaszthatóságot, a multifunkcionalitást és a funkcionális hatókört emeli ki. Véleménye szerint erre a három jellemzőre vezethető vissza a diskurzusjelölők valamennyi kritériuma, emellett a három jellemző valamelyikével lehetővé válik a DJ-előfordulások azonosítása (vö. szintaktikai leválaszthatóság), valamint adott nyelvi elem DJ és nem-DJ előfordulásai közti különbségtétel és egyértelműsítés (vö. multifunkcionalitás, funkcionális hatókör) (Crible 2017: 105).

Crible (2017) szerint „minden kategoriális meghatározás csak annyiban hasznos, amennyiben azt az azonosítás és az annotáció valamely empirikus modellje támasztja alá”[3], ezzel egyetértve a szintaktikai leválaszthatóságot, a multifunkcionalitást és a funkcionális hatókört mint meghatározó jegyeket használtuk

[2] A táblázatban használt rövidítések:
seq. – sequentiality-coherence-connectivity (szekvencialitás / a koherencia elősegítése – konnektivitás)
context – context-dependence (kontextusfüggőség)
oral. – orality (oralitás)
synt. – syntactic criteria (szintaktikai diverzitás, szintaktikai integráció hiánya, szintaktikai leválaszthatóság)
poly-funct. – poly-functionality / multifunctionality (multifunkcionalitás)
scope – variable scope (változó funkcionális hatókör)
non-prop. – non-propositional content (nem propozíciós jelentés)
inv. – invariable form (állandó alak)

[3] „any categorical definition is only useful insofar as it is endorsed by an empirical model of identification and annotation” (Crible 2017:99, saját fordítás).

a kutatás részeként végzett manuális annotációk során, melyek ezen sorrendje az egyes jegyek súlyát, meghatározó szerepét is jelzi, a funkcionális hatókör mint manuális annotációhoz szükséges jegy figyelembe vételére például az esetek kevesebb, mint 5%-ában volt szükség.

Esettanulmány: a USAS egyértelműsítési módszereinek tesztelése

A meghatározó jegyek problematikájának fenti ismertetése előrevetíti, hogy a DJ és nem-DJ előfordulások azonosítása és osztályozása a gépi feldolgozás számára is kihívásokat tartogat. Tanulmányunk első esettanulmánya a USAS automatizált annotáló rendszer egyértelműsítési módszereit és címkézési eljárását hivatott tesztelni a DJ vs. nem-DJ előfordulások fényében.

Prentice (2010) alapján az automatizált annotáló és egyértelműsítési eszközök alábbi típusait különböztethetjük meg:

- mesterséges intelligencia alapú;
- háttértudás-alapú;
- korpusz-alapú;
- előre meghatározott szemantikai taxonómián alapuló rendszerek.

A USAS rendszer Rayson et al. (2004) alapján egy statisztikai elven működő szófaji elemzőt, egy lemmatizáló komponenst, egy lexikonmodult és egy szabályalapú szemantikai címkézőt tartalmaz, így – közvetlenül vagy közvetetten – Prentice (2010) fenti eljárásai közül szinte mindegyiket használja, a mesterséges intelligencia kivételével. Ahogyan azt Furkó (2020a: 15) megállapítja, a rendszer előnye, hogy szemantikai mezőkön alapuló taxonomikus osztályozó rendszert ad a kutató kezébe, és minden egyes lexikális elem annotálásra kerül, amely a gyakran perifériális elemekként kezelt diskurzusjelölők vonatkozásában fontos tényező. Az osztályok, melyekbe az alkalmazás besorolja a szavakat, a rendszer készítői által rögzítettek.

A korpusznyelvészeti eszközök közül a USAS használata az alábbi jellemzők alapján is indokolt (vö. Furkó 2020b: 15):

- a felhasználás egyszerűsége;
- tanulhatóság;
- stabil működés, alacsony hardverigény;
- technikai támogatás;
- előzmények (korábbi kutatások, létező annotált korpuszok stb.);
- nyílt forráskód, ár/érték arány (ingyenes használat).

A USAS szemantikai elemző rendszer automatizált annotációs modullal rendelkezik, mely a 2. táblázatban látható 21 szemantikai mező mentén címkéz.

A	**B**	**C**	**E**
General and abstract terms (’általános és absztrakt kifejezések’)	The body and the individual (’az emberi test és az egyén’)	Arts and crafts (’mesterségek, szakmák, művészeti formák’)	Emotional actions, states and processes (’érzelmi cselekedetek, állapotok és folyamatok’)
F	**G**	**H**	**I**
Food and farming (’élelmiszer és gazdálkodás’)	Government and the public domain (’közigazgatás és közélet’)	Architecture, buildings, houses and the home (’építészet, épületek, az otthon’)	Money and commerce (’pénz és kereskedelem’)
K	**L**	**M**	**N**
Entertainment, sports and games (’szórakozás, sport és játék’)	Life and living things (’élet és élővilág’)	Movement, location, travel and transport (’mozg[at]ás, helyszín, utazás és szállítás’)	Entertainment, sports and games (’szórakozás, sport és játék’)
O	**P**	**Q**	**S**
Substances, materials, objects and equipment (’anyagok, tárgyak és felszerelések’)	Education (’oktatás’)	Linguistic actions, states and processes (’nyelvi műveletek, állapotok, folyamatok’)	Social actions, states and processes (’társadalmi cselekvések, állapotok és folyamatok’)
T	**W**	**X**	**Y**
Time (’idő(pont)’)	The world and our environment (’a világ és környezetünk’)	Psychological actions, states and processes (’lélektani cselekedetek, állapotok és folyamatok’)	Science and technology (’tudomány és technológia’)
Z			
Names and grammatical words (’[tulajdon]nevek és nyelvtani elemek’)			

2. táblázat: A USAS automatizált annotáló eszköz által használt szemantikai mezők

A fenti szemantikai mezők alapján a program a lexikális elemeket 232 különböző címkével látja el, az A mezőt például 47, az Y mezőt csupán két komponensre osztva. A szemantikai címkézése olyan egyértelműsítési módszerek

(disambiguation methods) összesített eredményei alapján történik, mint például a szófaji címkézés (POS tagging), az általánosvalószínűség-rangsorolás, az összetett kifejezések kibontása, a diskurzusdomén azonosítása és a kontextuális szabályok figyelembevétele (részletesen l. Rayson et al. 2004).

Az esettanulmány során az alábbi kérdésekre kerestünk választ:

(1) Alkalmas-e a USAS a gyakori DJ-típusok esetén a DJ és nem-DJ tokenek megkülönböztetésére?
(2) A Rayson et al. (2004) által mért hibahatár (9%) a DJ tokenek címkézésére is vonatkozik-e?
(3) Az egyes DJ-k tekintetében a hibahatár mennyire hasonló vagy eltérő?
(4) Amennyiben az egyes DJ-k tekintetében a hibahatár eltérő, milyen formális-funkcionális jellemzők okozhatják az eltéréseket?

A klasszikus DJ-listák (Schiffrin 1987, Redeker 1990, Fraser 1999 USAS-címkézése során igen heterogén képet kapunk:

- *oh, well, but, and, or, so, because, now, then, I mean, y'know, see, look, listen, here, there, why, gosh, boy, this is the point, what I mean is, anyway, whatever* (Schiffrin 1987) => oh_Z4[i1.2.1 well_Z4[i1.2.2 but_Z5 and_Z5 or_A13.4[i2.2.1 so_A13.4[i2.2.2 because_Z5/A2.2 now_Z4[i3.2.1 then_Z4[i3.2.2 I_Z4[i4.2.1 mean_Z4[i4.2.2 y'know_Z99 see_X3.4 look_A8 listen_X3.2 here_M6 there_M6 why_A2.2 gosh_Z4 boy_S2.2m this_Z8 is_A3+ the_Z5 point_Q2.1 what_Z8 I_Z4[i5.2.1 mean_Z4[i5.2.2 is_A3+ anyway_Z4 whatever_Z8
- *when, as, while, meanwhile; (and) then, next, now, before, after, because, (and) so, after that, all this time, well, okay, you know, I mean, mind you, anyway(s)* (Redeker 1990) => when_Z5 as_Z5 while_Z5 meanwhile_T1.1.2 then_N4 next_N4 now_T1.1.2 before_N4 after_Z5 because_Z5/A2.2 so_Z5 after_Z5 that_Z5 all_N5.1+ this_M6 time_T1 well_A5.1+ okay_A5.1+ you_Z8mf know_X2.2+ I_Z4[i6.2.1 mean_Z4[i6.2.2 mind_Z4[i7.2.1 you_Z4[i7.2.2 anyway_Z4
- *consequently, also, above all, again, anyway, alright, alternatively, besides, conversely, in other words, in any event, meanwhile, more precisely, nevertheless, next, otherwise, similarly, or, and, equally, finally, in that case, in the meantime, incidentally, OK, listen, look, on the one hand, that said, to conclude, to return to my point, while I have you* (Fraser 1990) => consequently_A2.2 also_N5++ above_A13.2[i8.2.1 all_A13.2[i8.2.2 again_N6+ anyway_Z4 alright_A5.1+ alternatively_A6.1- besides_Z5 conversely_A6.1-in_Z4[i9.3.1 other_Z4[i9.3.2 words_Z4[i9.3.3 in_Z5 any_N5.1+ event_A3+/A11.1+ meanwhile_T1.1.2 more_A13.3 precisely_A4.2+ nevertheless_Z4 next_N4 otherwise_A6.1- similarly_A6.1+ or_Z5 and_Z5 equally_A6.1+ finally_N4 in_Z4[i10.3.1 that_Z4[i10.3.2

> case_Z4[i10.3.3 in_T1.3[i12.3.1 the_T1.3[i12.3.2 meantime_T1.3[i12.3.3 incidentally_Z4 OK_A5.1+ listen_X3.2 look_A8 on_Z4[i13.4.1 the_Z4[i13.4.2 one_Z4[i13.4.3 hand_Z4[i13.4.4 that_Z8 said_Q2.1 to_Z5 conclude_X6+ to_Z5 return_M1 to_Z5 my_Z8 point_Q2.1 while_Z5 I_Z8mf have_A9+ you_Z8mf

A fenti listák és egy előzetes annotáció során megfigyelhető volt, hogy a diskurzusjelölők vonatkozásában a USAS címkézését két típusra oszthatjuk (vö. Furkó 2020a: 82):

(1) Diskurzusjelölő-releváns címkék:

- Z4 diskurzuselemek (discourse bin, pl. *oh, I mean, you know, basically, obviously, right, yeah, yes*)
- A5.x értékelő[4] kifejezések (evaluative terms depicting quality, pl. *as well, OK, okay, good, right, alright*)

(2) DJ-irreleváns címkék:

- A4.1 Általános/absztrakt kifejezések (General/abstract terms denoting types, groups, examples)
- N5 Mennyiséget kifejező terminusok (Terms depicting quantities)
- Q1.1 Kommunikációra vonatkozó általános kifejezések (Terms relating to communication in general)
- T1.1.2 Temporalitást kifejező elemek (General terms relating to a present time)
- Z5 Grammatikai elemek (Grammatical bin, Prepositions/adverbs/ conjunctions, etc.)
- X2.1 Az érveléshez, gondolkodáshoz kapcsolódó kifejezések (Terms relating to reasoning/thinking)[5]

A kutatás két 100.000 szavas alkorpuszon alapult: az első alkorpusz spontán beszélt nyelvi amerikai angol beszédfordulókat foglal magában (Santa Barbara Corpus of Spoken American English-ből random mintavételezéssel kinyert), a második (saját gyűjtés) az Egyesült Államokban sugárzott politikai interjúkból vett nyelvi adatokon alapszik. A diskurzusjelölők USAS-címkéinek azonosítása és összehasonlítása érdekében a két alkorpuszban a gyakori diskurzusjelölőkhöz (pl. *I mean, you know, so, well*) rendelt szemantikai címkékre került a hangsúly,

[4] Egyik anonim lektorunk megjegyzi, hogy a példaként megjelölt diskurzusjelölők már nem (mindig) fejeznek ki minőséget, hiszen DJ-vé váltak, akként pedig más a funkciójuk, sokszor szekvenciális szerepük van csak. Mindez számunkra azt jelzi, hogy az USAS a DJ-k eredeti vagy magjelentése alapján címkéz, ennek részletes vizsgálata azonban túlmutat a jelen tanulmány keretein.

[5] pl. *I think, you know* szintaktikailag integrált, nem-DJ előfordulásai, melyeket vélhetően az USAS a POS tagging alapján azonosított be.

majd ezen szemantikai címkék segítségével a korpuszban szaliens, potenciálisan diskurzusjelölő-funkciókat betöltő lexikai elemek kerültek azonosításra (pl. *in other words, right, now, actually*). A kutatás következő szakaszában a két korpuszban található 400 előfordulásból álló mintát állítottunk össze: a mintához kiválasztott tokeneket gyakoriságuk alapján súlyoztuk, emellett az automatizált annotáció alapján DJ és nem-DJ előfordulások egyenlő arányban szerepeltek a mintában. Egy példa a mintavételezésre: a *well* diskurzusjelölő 429 előfordulása 19,6%-át képezi az összes DJ-releváns címkével ellátott tokenek számának, így egész számra kerekítve 78 (400x0,196=78,4), azaz 39 A5.1 címkével és 39 nem A5.1 címkével ellátott *well* került a mintába.

A mintában található, a fent megjelölt három meghatározó jegy alapján diskurzusjelölőnek tekinthető elemek 1-es számkóddal, míg a nem diskurzusjelölő előfordulások 2-es számkóddal kerültek manuális annotálásra, míg az automatizált annotálás során DJ-releváns címkéket 1-es számkódra, a nem-DJ releváns címkéket 2-es számkódra cseréltük. A manuális és automatizált címkézések összehasonlítását a szakirodalomban elterjedt annotátorközi egyezések (Scott-féle Pi és a Cohen-féle Kappa értékek) meghatározásával végeztük. Az eredményt a 3. táblázat összegzi.

A Scott-féle Pi és a Cohen-féle Kappa értékek alapján négyes szintű, tehát teljes mértékű annotátorközi egyezőséget találunk a manuális és az automatizált annotáció között (vö. Crible–Degand 2017: 82), a globális értékek mögött azonban jelentős különbségeket találunk az egyes DJ-előfordulások vonatkozásában.

	Százalékos egyezőség	**Scott-féle Pi**	**Cohen-féle Kapp**	**egyezőségek száma**	**eltérések száma**	**tokenek száma**	**címkék száma**
Variáns (DJ/nem-DJ)	92,75	0,85	0,85	371	29	400	800

3. táblázat: Annotátorközi egyezés automatizált és manuális annotáció révén címkézett DJ és nem-DJ tokenek vonatkozásában

A több elemből álló DJ-k, például az *I mean* és a *you know* esetében a minta vonatkozásában is magas annotátorközi egyezést találunk (Cohen-féle kappa >0,98), tehát a USAS magas precizitással tesz különbséget a DJ (1a, 2a) és a nem-DJ (1b, 2b) tokenek között:

(1a) I_Z4[i1.2.1 ***mean***_Z4[i1.2.2 the_Z5 long-term_T1.3+ plans_X7+ for _Z5 Britain_Z2 are_A3+ for_Z5 a_Z5 second_N4 West_M7[i3.2.1 Coast_M7[i3.2.2 mainline_M3 railway_M3c ._PUNC

(1b) Well_A5.1+ ,_PUNC yes_Z4 ,_PUNC ***I***_Z8mf do_Z5 ***mean***_Q1.1 that_Z8 ._PUNC

(2a) Getting_A9+ crime_G2.1- down_Z5 below_Z5 what_Z8 it_Z8 used_A1.5.1 ,_PUNC ***you***_Z4[i1.2.1 ***know***_Z4[i1.2.2 ,_PUNC otherwise_A6.1- would_A7+ be_A3+ ._PUNC

(2b) ***You***_Z8mf ***know***_X2.2+ the_Z5 answer_Q2.2 to_Z5 that_Z5 question_Q2.2 ._PUNC

Az (1b) és (2b) *I mean*, illetve *you know* előfordulásai szintaktikai beágyazottságuk miatt nem tekinthetők diskurzusjelölőknek, míg az (1a) és (1b) esetében a beágyazottság hiánya mellett a funkcionális hatókör is indokolja a DJ-releváns címkézést. Érdekes megfigyelni, hogy a (2a) esetében ráadásul a USAS a téves kezdés (ang. *false start*) és az újrafogalmazás jelenléte ellenére helyesen elemzi a szekvenciát.

Az egyszavas DJ-tokenek esetében alacsonyabb precizitást, tehát magasabb hibaarányt találunk, az *actually* esetében például DJ-releváns címkézés figyelhető meg függetlenül az adott token megnyilatkozásban betöltött szerepétől és szintaktikai beágyazottságától:

(3a) No_Z4 ,_PUNC that_Z8 was_A3+ n't_Z6 exactly_A4.2+ the_Z5 reason_A2.2 ._PUNC ***Actually_A5.4+*** **,**_PUNC what_Z8 it_Z8 was_A3+ ,_PUNC is_Z5 I_Z8mf felt_X2.1 that_Z5 films_Q4.3 were_Z5 getting_A9+ they_Z8mfn started_T2+ to_Z5 be_Z5 repeating_N6+ ._PUNC

(3b) They_Z8mfn 're_A3+ one_T3 of_Z5 the_Z5 few_N5- cats_L2mfn in_Z5 the_Z5 world_W1 that_Z8 can_A7+ ***actually_A5.4***+ swim_M4 under_M4[i619.2.1 water_M4[i619.2.2

A (3b) *Cats can actually swim* (’A macskák valóban tudnak úszni’) megnyilatkozásában az *actually* szintaktikailag integrált, módosítószó szerepét ellátó elem, tehát nem indokolt a DJ-releváns címkézés.

A *now* címkézésekor ezzel szemben DJ-irreleváns USAS-annotációt találunk, függetlenül attól, hogy szintaktikailag független, diskurzusszervező funkcióval bíró (4a), vagy pedig ténylegesen temporalitást kifejező, propozicionális jelentéssel bíró elemről van szó (4b):

(4a) Good_Z4[i297.2.1 heavens_Z4[i297.2.2 ,_PUNC such_Z5 an_Z5 intelligent_X9.1+ man_S2.2m is_Z5 excited_X5.2+ about_Z5 a_Z5 movie_Q4.3 star_W1 ?_PUNC ***Now_T1.1.2***[6] what_Z8 about_Z5 her_Z8f and_Z5 the_Z5 Kennedy_Z1mf 's_Z5 ?

(4b) Somebody_Z8mfc explain_Q2.2/A7+ to_Z5 Paris_Z2 and_Z5 Nicole_Z1f ,_PUNC live_L1+ means_X4.2 we_Z8 're_A3+ on_Z5 television_Q4.3 right_T1.1.2[i7.2.1 ***now_T1.1.2***[i7.2.2 ._PUNC

[6] T1.1.2 USAS címke–Time: General: Present; simultaneous

A *like* szintén problematikus az automatizált annotálás szempontjából: a USAS a DJ-előfordulásokat (5a, b, c) is gyakran helytelenül nyelvtani elemként címkézi (Z5=‘grammatical bin‘), míg DJ-releváns címkét (Z4=discourse bin) kizárólag olyan esetekben találunk, ahol a *like* előtt és utána is vessző van az átiratban (5d):

(5a) This guy is so cool. I mean, he’s ***like*** the coolest person you could meet. (I_Z4[i1.2.1 mean_Z4[i1.2.2 ,_PUNC he_Z8m s_T1.3 ***like***_Z5 the_Z5 coolest_O4.6-person_S2mfc you_Z8mf could_A7+ meet_S3.1 ._PUNC)

(5b) I went to the clerk to ask him where the beer was, and he’s ***like***, ‘I don’t know, I’m new here’, so I’m ***like***, yeah, sure, ***like***, you should know this, man! (so_Z5 Im_Z99 ***like***_Z5 ,_PUNC yeah_Z4 ,_PUNC sure_A7+ ,_PUNC like_Z4 ,_PUNC you_Z8mf should_S6+ know_X2.2+ this_Z8 ,_PUNC man_S2.2m)

(5c) I missed ***like*** 40 questions on the exam. (I_Z8mf missed_A5.3- ***like***_Z5 40_N1 questions_Q2.2 on_Z5 the_Z5 exam_P1 ._PUNC)

(5d) Could you, ***like***, loan me $100? (Could_A7+ you_Z8mf ,_PUNC ***like***_Z4 ,_PUNC loan_A9- me_Z8mf $100_Z99 ?_PUNC)

A USAS címkézés és a manuális annotáció összevetése azt sugallja, hogy a rendszer a többi nyelvi elem címkézéséhez hasonlóan a Rayson és munkatársai (2004) által megállapított 9%-os hibahatáron belül képes különbséget tenni a 400 vizsgált elem diskurzusjelölői és nem diskurzusjelölői előfordulásai között, mely révén hasznos funkcionális annotációt előkészítő eszközként szolgál. A többszavas kifejezések esetében a humán és USAS annotáció közti egyezőség még magasabb, feltehetően a USAS általános valószínűségi rangsoroló („general likelihood ranking”) és többszavas kifejezéseket kivonatoló („multi-word-expression extraction”) moduljainak köszönhetően.

Esettanulmány: a BERT és az angol *right* DJ és nem-DJ előfordulásai

A USAS után rátérünk a BERT-tel végzett kutatásaink bemutatására, mely reményeink szerint további távlatokat nyithat meg a diskurzusjelölők kutatásában. A BERT nyílt forráskódú gépi tanulási rendszer természetesnyelvi feldolgozáshoz (NLP), ami kontextualizált neurális szóbeágyazásokat állít elő címkézetlen korpuszadatok feldolgozása során. A kapott szóbeágyazások a célszavak (jelen esetben a *right* szó) előfordulásainak disztribúciós tulajdonságait jellemzik adott környezetben.

A szavak korpuszokban megfigyelt disztribúciós jellemzőinek gépi tanulással történő feldolgozása a számítógépes szemantika dinamikusan fejlődő területe, az NLP kulcsfontosságú fejlesztése, mely szerepet kap a webes keresések megvalósításában, az ismert gépi fordítórendszerekben, valamint a mesterséges intelligencia megoldásoknak, társalgóeszközöknek is központi eleme. Alapja a szavak

együttes előfordulásának megfigyelése és tárolása. Például az *itta* szó – hasonlóan a *kortyolta*, *nyakalta*, *szürcsölte*, stb. szavakhoz – a *teát*, *vizet*, *tejet* szavakkal gyakrabban fordul elő, mint az *autót*, *csengőt* szavakkal, ugyanakkor a *hallotta* szó környezetében az *autót* és a *csengőt* szavak gyakoribbak az élelmiszerek nevénél. Az, hogy melyik szót, szóalakot milyen környezetben, milyen disztribúcióban figyelhetjük meg, köszönhető egyrészt a nyelv struktúrájának, beleértve az alaktan és a mondattan által vizsgált jelenségeket, másrészt jelentésbeli, használatbeli hasonlóságokból és különbségekből adódik, adott példában elsősorban abból, hogy a *tea*, *víz*, *tej* szavak iható élelmiszereket fejeznek ki, a *csengő* és az *autó* pedig hallhatók. Az, hogy a szavak disztribúciójának vizsgálatával szemantikai hasonlóságok és különbségek is megragadhatók, a *disztribúciós hipotézis*ig vezethető vissza (Harris 1954), azonban csak a 20. század végére vált elérhetővé olyan adathalmaz és számítási kapacitás, mellyel automatizált módon képesek vagyunk ezt a jelenséget megragadni és a természetesnyelv-feldolgozásban felhasználni.

Amennyiben a feladatot gépi tanulás nélkül végezzük el, akkor a vizsgált célszavak (a fenti példában az *itta* és a *hallotta* szóalakok) kiválasztott méretű környezetében megkeressük és megszámláljuk a vizsgálatba bevont, előre kijelölt kontextusszavakat (itt pl. *teát*, *autót*), az együttes előfordulások számát az adott célszót jellemző tulajdonságvektorokban tároljuk, az így kapott adatokat pedig súlyozzuk, statisztikailag feldolgozzuk. A hasonlóbb szavak tulajdonságvektorai egymáshoz közelebbi pontokra, a kevésbé hasonlók pedig távolabbi pontokra mutatnak a vektortérben, ezt pedig a számítógépes nyelvészetben ki tudjuk használni (ld. Tóth 2014).

A szódisztribúciót újabban mesterséges neurális hálózatokat alkalmazó gépi tanulással tárják fel, mi is gépi tanuláson alapuló rendszert használunk a munkánk során. Mikolov et al. (2013) – úttörő munkát végezve – két ilyen eljárást vezettek be. A „skip-gram" eljárásban a neurális hálózat egy szóalakot kap a bemenetén, feladata az adott szó környezetében előforduló szavaknak a megjóslása nagy tanítókorpusz előzetes feldolgozásának segítségével. A feldolgozás idegsejtalapú, a csatlakozó neuronok egymást hozzák működésbe súlyozott kapcsolatokon terjedő aktivációkkal. A neurális hálózat a neuronok közti súlyok megfelelő beállításával tanulja meg a feladat optimális megoldását, amely a hálózat kimenetét alkotó idegsejtek megfelelő mintázatú aktiválódásában mutatkozik meg. A bemutatott eljárásokban a hálózat a kimenetén több szót is aktiválhat: egy-egy szóalakot egy-egy kimeneti idegsejt reprezentál, a hálózat pedig a tanulás során a jósolt szavak és a korpusz adott pontján ténylegesen előforduló szó összehasonlításával a hálózat működésének hibáját méri, és ezt a hibát minimalizálja az idegsejtek közti kapcsolatok súlyának beállításával, finomhangolásával. A másik eljárás a „continuous bag-of-words" (CBOW), mely a bemenetén az aktuális mondatban a célszó környezetében talált szavakat összegyűjtő szólistát (a számítógépes nyelvészet terminológiájában „szózsákot", angolul „bag of words") kapja meg, a kimenetén pedig azt a szót várjuk aktiválódni, melynek környezetét a

bemeneten bemutattuk. Természetesen egyik feladat megoldása során sem kapunk nagy pontosságú jóslatokat, hiszen nem tudjuk pontosan megjósolni sem az adott szó aktuális környezetét („skip-gram" esetén), sem a környezetből hiányzó konkrét szót („CBOW"), de nem is erre szeretnénk használni ezeket a hálózatokat, hanem azt várjuk tőlük, hogy a szódisztribúció megtanulásával hasznos belső reprezentációt alakítsanak ki a fókuszba helyezett (a „skip-gram" eljárásban a bemeneten elhelyezett, a „CBOW" esetén a kimeneten elvárt) szóalakhoz. A kialakuló reprezentáció (neurális szóbeágyazás) az adott szó korpuszbeli disztribúcióját, ennek részeként az adott szó jelentését jellemzi, miközben a tanítókorpusz ezekben az esetekben sem tartalmaz annotációt.

A gépi tanulásra támaszkodó neurális szóbeágyazások a gyakorlatban előnyösebbnek, a szavak jelentését jobban tükrözőnek bizonyultak a korábbi eljárásoknál (összehasonlításért lásd pl. Baroni et al. 2014), ezért gyorsan elterjedtek. Az idegsejthálózatokat összeállító számítógépes nyelvészek, nyelvtechnológusok egyre komplexebb eljárásokat dolgoznak ki a mesterséges idegsejtek komplexitása, száma, topológiája és a felhasznált adatok mennyisége tekintetében. Az általunk használt eljárás is egy ilyen, továbbfejlesztett rendszer.

Az, hogy ezek a hálózatok valójában milyen mintázatokat tanulnak meg, milyen kategóriák feldolgozására allokálnak belső erőforrásokat a kijelölt feladat megoldása során, és mindez hogyan egyeztethető össze a nyelvről rendelkezésre álló nyelvészeti tudásunkkal, jelenleg is vizsgálat tárgya.

Ebben az esettanulmányban azt vizsgáljuk, hogy a diskurzusjelölők felismerésének, jellemzésének feladatát hogyan tudjuk neurális szóbeágyazások segítségével elvégezni. A használt tanulási eljárás a BERT (Devlin et al. 2019), mely kontextualizált szóbeágyazásokat állít elő, azaz a célszó aktuális, adott mondatbeli használata generál egy, az adott esetre szabott neurális szóbeágyazást a hálózat működése során. A kontextualizált szóbeágyazások ezen tulajdonsága számunkra meghatározó jelentőségű, hiszen a kapott beágyazás nem az adott szóalak általános jellemzésére szolgál, melyet többértelműség esetén egy későbbi lépésben kellene egyértelműsíteni, hanem az adott használatra szabott, egyértelműsítést nem igénylő tulajdonságvektort kapunk, melyet közvetlenül felhasználhatunk a célszó egyes előfordulásainak összehasonlítására. A BERT komplex hálózati topológiát használ, melyet itt nem tárgyalunk, azonban a hálózat számára betanított egyik feladat nagyon hasonlít ahhoz, amit fentebb a CBOW eljárással kapcsolatban leírtunk (célszóra vonatkozó predikció adott környezetben). A másik feladat a mondatok egymás szomszédságában, megfelelő sorrendben előfordulásának felismerése – ezzel kitekintve a mondathatárokon túlra, meghaladva a szópredikció szintjét, mely hasznos tulajdonságnak ígérkezik pragmatikai vizsgálatok esetén. A BERT az ismeretlen szavak ábrázolására is rendelkezik megfelelő eszközzel: azokat szótöredékekre, szükség esetén betűkre bontja. Elemzési célokra a BERT továbbra is kiváló megoldásnak tűnik úgy is, hogy közben sorra

jelennek meg neurális szóbeágyazások előállítására használható egyéb eljárások (pl. Albert, RoBERTa, GPT).

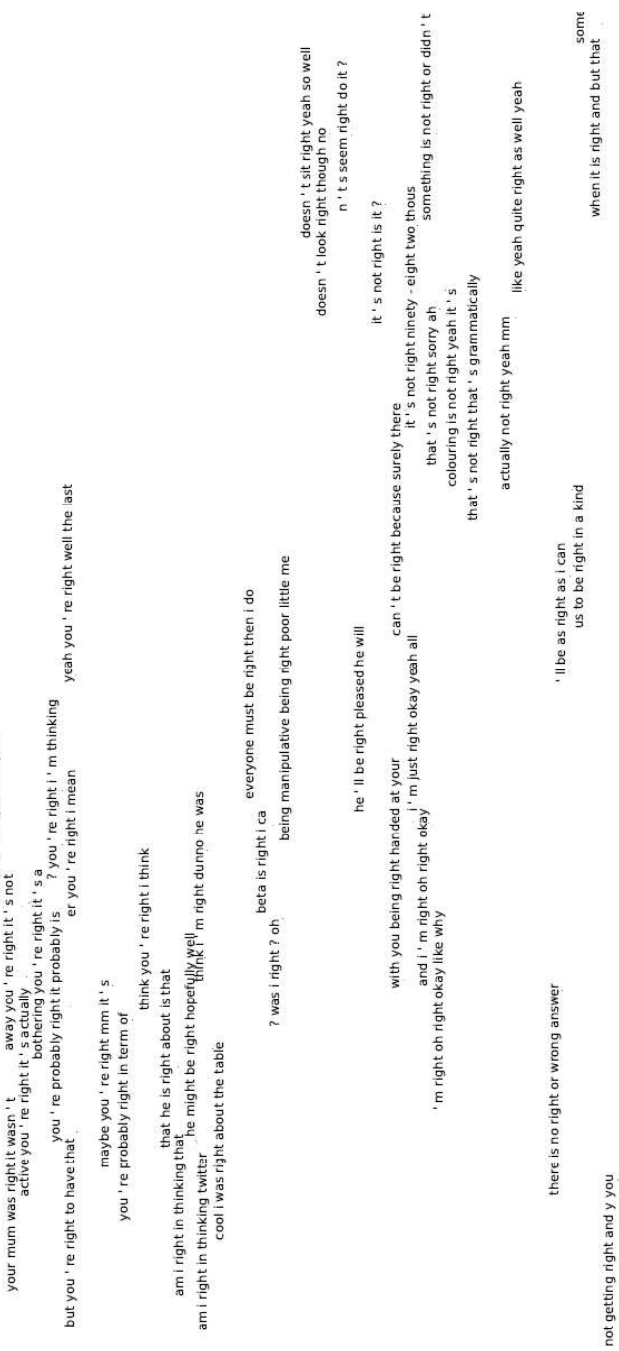
Saját megfigyeléseink szerint (Furkó–Tóth 2021) és a szakirodalmi előzmények alapján (lásd pl. Rogers et al. 2021: 843–844) a BERT-szóbeágyazások információt kódolnak a következőkről:

- szófaj,
- mondattani egységek,
- függőségek, részleges szintaktikai fák,
- alany-állítmány egyeztetés,
- negatív polaritású szavak,
- entitások típusa és relációi,
- szemantikai szerepek,
- jelentéstani protoszerepek,
- egyszerűbb, „gap-filling" feladatokhoz tudásindukció (pl. Cats like to chase __.).

Az angol *right* diskurzusjelölő vizsgálatára a következő eljárást alkalmaztuk (vö. Furkó–Tóth 2021: 20). Első lépésként korpuszt szerkesztettünk, melyben a *right* szó 10.000 előfordulását tartalmazó mondatok szerepeltek a Sketch Engine 'BNC 2014 Spoken' alkorpuszából (https://www.sketchengine.eu) véletlenszerűen kiválasztva. Ezután a *right* szó mind a 10.000 előfordulásához BERT-szóbeágyazásokat készítettünk a „Hugging face" MI erőforrás-gyűjteményben (https://huggingface.co) szereplő 1024-dimenziós, felhasználásra előkészített 'BERT-large' adathalmaz segítségével („bert-large-uncased", https://huggingface.co/bert-large-uncased). Mindez a BERT esetében nem egyszerűen a szóhoz tartozó tulajdonságvektor kikeresését jelenti egy adatbázisból, hanem a neurális hálózat futtatását igényli a korpusz mondataira. Mivel az elemzés a továbbiakban manuális klaszterezéssel folytatódott, így az 1024 dimenziós térben megjelenő mondatokat 2 dimenziós térben ábrázoltuk, melyhez az adatokon t-SNE transzformációt (van der Maaten–Hinton 2008) hajtottuk végre. A kapott eredményre jellemző, hogy az eredeti, 1024 dimenziós térben egymáshoz közeli pontok a t-SNE diagramon is egymás közelében jelennek meg. A könnyebb feldolgozás érdekében a megjelenített mondatok számát az utolsó lépésben 1000-re csökkentettük, valamint csak a mondatok egy részét (legfeljebb 3 szót a *right* mindkét oldalán) tüntettük fel a diagramon. Ezúttal automatizált osztályozást nem hajtottunk végre, hanem mi magunk kerestük a *right* használatainak csoportosulását, és annak lehetséges magyarázatát az elkészített diagramon.

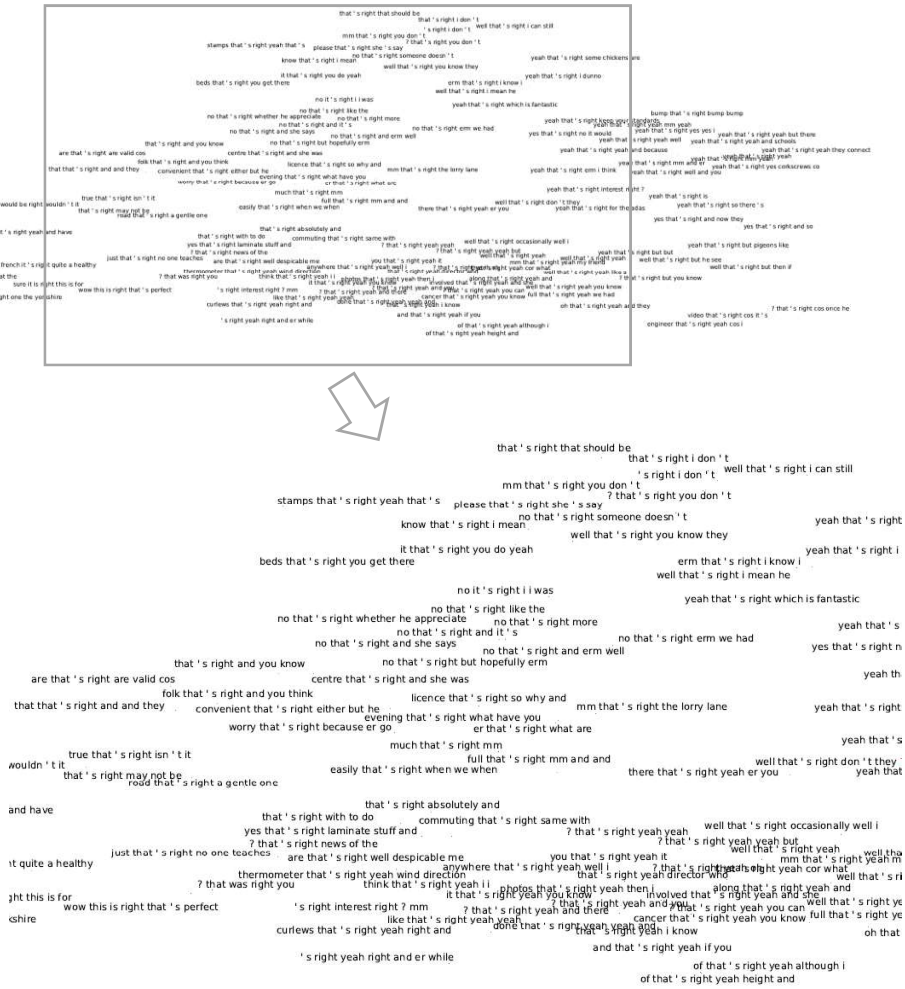
Az ily módon nyert kétdimenziós klaszterek vonatkozásában az alábbi megfigyeléseket tehetjük:

- Az 1. ábrán látható, hogy a BERT jellemzően egy klaszterben szerepelteti a *right* nem-DJ előfordulásait, függetlenül attól, hogy azok melléknévi vagy módhatározói szerepet látnak el (vö. *you're right* vs. *doesn't sit right*).



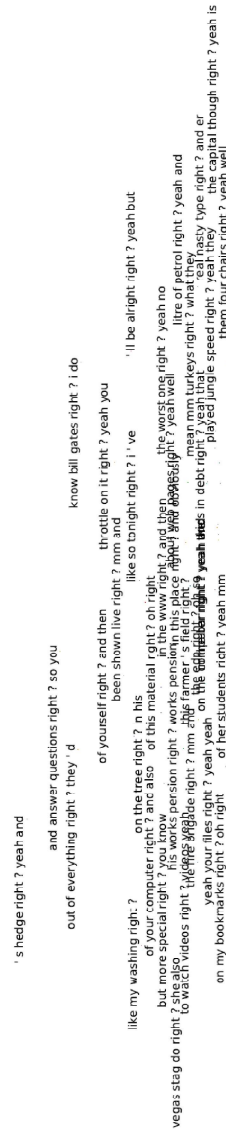
1. ábra: A right nem-DJ előfordulásai a BERT-klaszterek kétdimenziós megjelenítéseiben

- A 2. ábrán láthatjuk, hogy a DJ-előfordulások közül a válaszjelölő, megerősítő funkciót betöltő *right* tokenek szintén közel jelennek meg egymáshoz. A többszavas *that's right* egység felismerése miatt ez nem tekinthető átütő eredménynek.



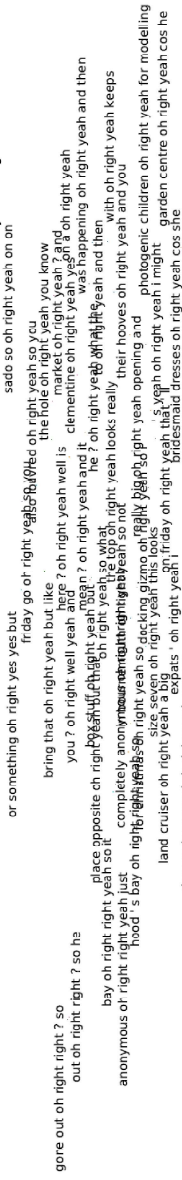
2. ábra: A right válaszjelölő/megerősítő funkciót betöltő DJ előfordulásai a BERT-klaszterek kétdimenziós megjelenítéseiben, a téglalapban lévő részlet felnagyított változatával

A 3. ábrán azt látjuk, hogy a beszélőpartner reakcióját elicitáló használat szintén felismerésre került, feltehetően az átiratokban szereplő kérdőjel közeli megjelenése révén, erre utal az a példa is (*did they? right, Chinese*), ahol helytelenül kerül azonosításra ez a funkció, mivel a kérdőjel egy korábbi szekvenciával, nem a *right* elicitáló használatával hozható összefüggésbe.



3. ábra: A beszélőpartner reakcióját elicitáló right DJ-előfordulások a BERT-klaszterek kétdimenziós megjelenítéseiben

A *right* kognitív feldolgozást (hirtelen felismerést) jelölő DJ-előfordulásait a BERT szintén külön klaszterben jeleníti meg, ahogyan azt a 4. ábrán láthatjuk. Ezekben az esetekben is valószínűsíthető, hogy a BERT az *oh right* többszavas kifejezés felismerése révén azonosítja ezt a funkciót.



4. ábra: A right kognitív feldolgozást (hirtelen felismerést) jelölő előfordulásai a BERT-klaszterek kétdimenziós megjelenítéseiben

Esettanulmányunk tanulsága, hogy – a *right* szó esetében – a kontextualizált neurális szóbeágyazások t-SNE vizualizációja részben képes volt elkülöníteni a DJ és nem-DJ használatokat, valamint megfigyelhettünk további klaszterekbe szerveződést is, melyre pragmatikai relevanciával bíró magyarázatot találtunk. A klaszterek jellemzésére azonban saját intuícióinkat, ismereteinket kellett használnunk, hiszen az általunk összeállított kísérleti rendszer címkézést jelenleg nem végez.

Összegzés, kitekintés

Írásunk két automatikus elemzőeszköz működését, használhatóságát méri fel egy-egy esettanulmány segítségével: az elsőben a statisztikai és szabályalapú UCREL Semantic Analysis System (USAS, vö. Rayson et al. 2004), a másodikban a neurális gépi tanulást megvalósító BERT (Devlin et al. 2019) tesztelése révén kíván adalékot szolgáltatni a diskurzusjelölők beazonosításához és egyértelműsítéséhez.

A USAS automatizált szemantikai annotációs eszköz egyértelműsítő módszereinek és precizitásának tesztelése kapcsán megállapítható, hogy a rendszer a többi lexikális elem címkézéséhez hasonlóan a Rayson és munkatársai (2004) által megállapított 9%-os hibahatáron belül képes különbséget tenni a 400 vizsgált elem diskurzusjelölői és nem diskurzusjelölői előfordulásai között (92,75% humán-USAS egyezőség, 0,85 Scott-féle Pi és Cohen-féle Kappa érték), mely révén hasznos funkcionális annotációt előkészítő eszközként szolgál. A többszavas kifejezések esetében a humán és USAS annotáció közti egyezőség még magasabb a USAS általános valószínűségi rangsoroló („general likelihood ranking") és többszavas kifejezéseket kivonatoló („multi-word-expression extraction") moduljainak köszönhetően.

A BERT-alapú vizsgálat során a manuális klaszterezési vizsgálathoz előállított t-SNE vizualizációk segítségével a DJ és nem-DJ előfordulások azonosításán túl további különbségtételek, funkcionális térképek nyerhetők ki, mindkét eszköz esetében megfigyelhető azonban, hogy nagymértékű eltéréseket tapasztalhatunk a több és az egyetlen elemből álló diskurzusjelölők címkézési és klaszterezési pontosságában, hibahatárában. A változó precizitás a diskurzusjelölőknek a tanulmány első részében ismertetett jellemzőivel magyarázható: a forráskategóriák változatosságával és diakrón folyamataival (rétegződésével), a szintaktikai függetlenedés skaláris jellegével, valamint a változó/funkcionális hatókörrel. Ezek a jellemzők jelenleg még kihívást jelentenek mind a USAS, mind a BERT által alkalmazott egyértelműsítési módszerek, tanulási modulok számára.

Irodalom

Andersen, G. 2001. *Pragmatic Markers and Sociolinguistic Variation: A Relevance-Theoretic Approach to the Language of Adolescents.* Amsterdam and Philadelphia: John Benjamins.

Baroni, M. – Dinu, G. – Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014.* 238–247.

Blakemore, D. 1987. *Semantic constraints on relevance*. Oxford: Blackwell.

Blakemore, D. 2002. *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers.* Cambridge: Cambridge University Press.

Brinton, L. J. 1996. *Pragmatic markers in English: Grammaticalization and discourse functions.* Berlin: Mouton de Gruyter.

Crible, L. – Cuenca, M-J. 2017: Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse* 8: 149–166.

Crible, L. – Degand, L. 2017. Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory* 15/1: 71–99.

Crible L. 2017. Towards an operational category of discourse markers - A definition and its model. In Fedriani, C. – Sansó, A. (eds.): *Pragmatic Markers, Discourse Markers and Modal Particles - New perspectives.* Amsterdam, Philadelphia: John Benjamins Publishing Company. 99–124.

Devlin, J. – Chang, M. – Lee, K. – Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics. 4171–4186.

Fischer, K. 2014. Discourse Markers. In Schneider, K. P. – Barron, A. (eds.) *Pragmatics of Discourse*. Berlin, München: De Gruyter Mouton. 271–294.

Fischer, K. (ed.). 2006. *Approaches to Discourse Particles*. Elsevier: Oxford.

Fraser, B. 1999. What are discourse markers? *Journal of Pragmatics* 31: 931–952.

Furkó P. 2019. Az újrafogalmazást jelölő diskurzusjelölők fordításának kérdései feliratok szövegében. In Dróth J. (szerk.): *Korpusz és kontrasztivitás a szakfordítás oktatásában és gyakorlatában*. Budapest: L'Harmattan. 45–60.

Furkó P. 2020a. *Discourse Markers and Beyond – Descriptive and Critical Perspectives on Discourse-Pragmatic Devices across Genres and Languages.* Cham/London: Springer / Palgrave Macmillan.

Furkó P. 2020b. Néhány hasznos korpusznyelvészeti eszközről a diskurzusjelölő-kutatás szemszögéből. In Kovács T. – Adorján M. (szerk.): *Korpusznyelvészet és nyelvi közvetítés*. Budapest: L'Harmattan. 45–56.

Furkó P. – Tóth Á. 2021. Manual and Automated Annotation of English Discourse Markers in Natural Conversations and Mediatized Political Discourse. Elhangzott: *Discourse-Pragmatic Variation and Change 5 (DiPVaC 5)*, The University of Melbourne, Arts West, 2021. december 14–16.

González, M. 2004. *Pragmatic markers in oral narrative – the case of English and Catalan.* Amsterdam/Philadelphia: John Benjamins.

Hansen, M-B. M. 2006. A dynamic polysemy approach to the lexical semantics of discourse markers. In Fischer, K. (ed.): *Approaches to Discourse Particles*. Oxford: Elsevier. 21–41.

Harris, Z. 1954. Distributional structure. *Word* 10/23: 146–162.

Knott, A. – Sanders, T. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics 30*/2: 135–175.

Kroon, C. 1995. *Discourse Particles in Latin*. Amsterdam: Gieben.

Levinson, S. C. 2004. Deixis and pragmatics. In Horn, L. – Ward, G. (eds.) *The handbook of pragmatics.* Oxford: Blackwell. 97–121.

van der Maaten, L. – Hinton, G. (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research* 1: 1–48.

Mikolov, T. – Chen, K. – Corrado, G. – Dean, J. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, Scottsdale, AZ, 2–4 May 2013. 1–12.

Prentice, S. 2010. Using automated semantic tagging in Critical Discourse Analysis: A case study on Scottish independence from a Scottish nationalist perspective. *Discourse & Society* 21/4: 405–437.

Rayson, P. – Archer, D. E. – Piato, S. – McEnery, T. 2004. The UCREL Semantic Analysis System. In *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks in Association with the 4th International Conference on Language Resources and Evaluation (LREC).* 7–12.

Redeker, G. 1990. Ideational and Pragmatic Markers of Discourse Structure. *Journal of Pragmatics* 143: 367–381.

Risselada, R. – Spooren, W. 1998. Discourse markers and coherence relations. *Journal of Pragmatics* 30/2: 131–133.

Rogers, A. – Kovaleva, O. – Rumshisky, A. (2021). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics* 8: 842–866.

Romaine, S. – Lange, D. (1998). The use of like as a marker of reported speech and thought: a case of grammaticalization in progress. Chesire, J. – Trudgill, P. (eds.): *The Sociolinguistics Reader Volume 2: Gender and Discourse*. Bristol: J W Arrowsmith. 227–279.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Stenström, A.-B. 1990. Lexical items peculiar to spoken discourse. In Svartvik, J. (ed.) *The London-Lund Corpus of Spoken English: description and research*. Lund: Lund University Press. 137–175.

Tóth Á. 2014. *The Company that Words Keep: Distributional Semantics*. Debrecen: Debreceni Egyetemi Kiadó.