A diskurzusjelölők automatizált annotálása szemantikai mezők mentén

Lexikai pragmatika (2000-2020)
Furkó Péter
Károli Gáspár Református Egyetem

•A DJ-k és szerepköreik (Dér 2010; Furkó 2011)

- változatos forráskategóriák (indulatszók, mondatszók, kötőszók, határozók, igeneves szerkezetek stb.);
- extrém kontextusfüggőség / indexikalitás;
- extrém multifunkcionalitás (diszkurzív és pragmatikai funkciók);
- (szintaktikai értelemben vett) opcionalitás;
- szemantikai alulspecifikáltság;
- változó hatókör;
- csoportosulás (*na hát, nos hát, well then, oh well*);
- beszélt nyelvben kevés típus / sok token, írott nyelvben fordítva (=>stigmatizáció);
- attitűdjelölés – konnektivitás – interperszonális funkciók.

Magyar diskurzusjelölők

- **1.**
  *aha, **akár**, azért, aztán, bár, bizony, csak, csakhogy, -e, egyáltalán, egyébiránt, egyébként, egyik, elvégre, **éppenséggel**, és, de, hát, **hiszen**, hm, így, illetve, is, izé, jaj, lám, **legalább**, már, még, na, nemde, nemhogy, netalán, no, pedig, persze, pláne, **sőt**, szóval, talán, tehát, tényleg, **tudniillik**, **tulajdonképpen**, úgy, ugyan, ugye, ugyebár, úgymond, vagyis, vajon, valóban, viszont, **voltaképpen**. (Schirm 2011)*
- **2.**
  *á; a másik meg; aha; akkor; azért; aztán; bár; **bocsánat, hogy beleszólok**; de; de hát; de igen; egyébként; egyik; érted; értem; és; ha; hát; hm; hm?; hogy; hogyha; igen; így; illetve; **inkább azt mondom**; itt; izé; ja; ja igen; jaj; jó; látod; mármint; mármint hogy; meg; **még azért visszakérdezek, hogy**; **még így visszatérve**; mert; **mit is akartam**; mmm; mmm (’nem’ jelentésű* hümmögés); *mondjuk; most; na; nem; nézd; oké; pedig; például; pláne; s; sőt; **szeretnék szólni, hogy**; szóval; tehát; tényleg; tessék; tudniillik; tudod; úgy; úgy értem, hogy; ugye; úgyhogy; ühüm; üüm (’nem’ jelentésű hümmögés); vagy; **vagy hogy mondjam**; vagyis; vagyis hogy; várj/várjál; viszont; visszatérve (Dér – Markó 2007, 4 fős társalgás)*

Angol diskurzusjelölők

- 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)

- 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)

- 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)

Az USAS (UCREL Semantic Analysis System) által használt szemantikai mezők

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

USAS címkékkel ellátott angol diskurzusjelölők

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_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_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

DJ típusok címkézése

- DJ-releváns címkék:

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

- DJ-irreleváns címkék:

- **A4.1** General/abstract terms denoting types, groups, examples
- **N5** Terms depicting quantities
- **Q1.1** Terms relating to communication in general
- **T1.1.2** General terms relating to a present (period/point in) time
- **Z5** Grammatical bin (Prepositions/adverbs/conjunctions, etc.)
- **X2.1** Terms relating to reasoning/thinking

Kutatási kérdések

- (1) A gyakori DJ típusok esetén a DJ és nem DJ tokenek megkülönböztetésére alkalmas-e az USAS?
- (2) A címkézés során tapasztalt 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 korpusz

Két azonos méretű (100.000 szavas) alkorpusz:

- MPI alkorpusz: 37db, egyenként 60 perces médiainterjú átirata a BBC *Hard Talk* és *Newsnight* programjai alapján (http://bbc.co.uk);
- CI alkorpusz: 50db egyenként 45 perces médiainterjú a CNN Larry King Live programjai alapján (http://www.cnn.com).

A leggyakoribb DJ típusok USAS által címkézett DJ és nem DJ előfordulásai

típus	DJ token normalizált gyakorisága az MPI-ben	DJ token normalizált gyakorisága a CI-ben	nem-DJ token normalizált gyakorisága az MPI-ben	Nem-DJ token normalizált gyakorisága a CI-ben
well (429)	360xA5.1	312xA5.1	14xI1.1, 55xN5	1xA7, 2xB2, 24xN5
sort (38)	14xZ4	25xZ4	21xA4.1, 3xA1.1.1	10xA4.1
now (299)	4xZ4	1xZ4	288xT1.1.2, 7xZ5	229xT1.1.2, 6xZ5
(you) know (346)	205xZ4	455xZ4	140xX2.2, 1xZ6	307xX2.2
like (97)	6xZ4	17xZ4	51xZ5, 40xE2+	238xZ5, 139xE2+
(I) mean (141)	114xZ4	201xZ4	27xQ1.1	30xQ1.1, 5xS2.2.2
(in other) words (11)	4xZ4	13xZ4	7xQ.3	7xQ.3
actually (165)	165xA5.4	72xA5.4	0	0
(I) think (549)	126xZ4	121xZ4	423xX2.1	319xX2.1
right (114)	55xZ4, 53xA5.3	211xZ4, 98xA5.3	6xT1.1.2	12xN3.8, 16xS7.4, 15xT1.1.2

Precíz címkézés (high precision tagging)

(10a) 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

(10b) Well_A5.1+ ,_PUNC yes_Z4 ,_PUNC ***I***_Z8mf do_Z5 ***mean***_Q1.1 that_Z8 ._PUNC

(11a) 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

(11b) ***You***_Z8mf ***know***_X2.2+ the_Z5 answer_Q2.2 to_Z5 that_Z5 question_Q2.2 ._PUNC

Precíz címkézés (high precision tagging)

(12) I_Z8mf want_X7+ to_Z5 assure_A7+ you_Z8mf that_Z8 ***I***_Z4[i1.2.1 ***mean***_Z4[i1.2.2

what_Z8 ***I***_Z4[i2.2.1 ***say***_Z4[i2.2.2 when_Z5 I_Z8mf tell_Q2.2 you_Z8mf I_Z8mf

appreciate_E2+ your_Z8 contributions_A9- ._PUNC

(13) As_Z5 ***you***_Z4[i1.2.1 ***know***_Z4[i1.2.2 ,_PUNC the_Z5 Government_G1.1c says_Q2.1

it_Z8 's_A3+ too_N5.2+ early_T4+ to_Z5 tell_Q2.2 about_Z5 that_Z8 ._PUNC

DJ-releváns invariáns címkézés

(1) 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

(2) 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

DJ-releváns invariáns címkézés

(3) Good_Z4[i297.2.1 heavens_Z4[i297.2.2 ,_PUNC such_Z5 an_Z5 intelli-gent_X9.1+ man_S2.2m is_Z5 excited_X5.2+ about_Z5 a_Z5 movie_Q4.3 star_W1 ?_PUNC ***Now***_T1.1.2 what_Z8 about_Z5 her_Z8f and_Z5 the_Z5 Kennedy_Z1mf 's_Z5 ?

(4) 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

Marginális DJ-előfordulások

(a) My roommate never cleans when I ask him to. ***Like***, I asked him yesterday to clean, and he never did it. (***Like***_E2+,_PUNC I_Z8mf asked_Q2.2 him_Z8m yesterday_T1.1.1 to_Z5 clean_B4 ,_PUNC and_Z5 he_Z8m never_T1/Z6 did_A1.1.1 it_Z8 ._PUNC)

(b) 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)

(c) 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)

(d) 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)

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

Az automatizált és a manuális annotáció közti (annotátorközi) különbségek

	egyezés	Scott-féle Pi	Cohen-féle Kappa	Egyezőség száma	Eltérések száma	Token szám	Annotáció szám
Variable 1 (cols 1 & 2)	92.75	0.854519	0.854527	371	29	400	800

Konklúziók

1. Az USAS által alkalmazott egyértelműsítési módszerek hatékonyak a leggyakoribb DJ-típusok DJ és nem DJ előfordulásainak, azok arányának kiszámításához: Az USAS használata lehetővé teszi a kutató számára, hogy globális képet kapjon a vizsgált lexikális elemek D-értékeiről.
2. Az észlelt hibahatár általában a DJ-k globális azonosítására is vonatkozik, többszavas egységek esetében (pl. you know, I mean) a hibahatár még alacsonyabb.
3. Nagyfokú eltérést tapasztalunk a nem többszavas DM-ek címkézési pontosságában / hibahatárában. A változó pontosság a DJ-k jellemzőivel magyarázható: a forráskategória rétegződésével, a szintaktikai függetlenséggel, a változó / funkcionális hatókörrel. Ezek a jellemzők kihívást jelentenek az USAS által alkalmazott egyértelműsítési módszerek szempontjából (is).

Hivatkozott irodalom

- Beeching, K (2016) Pragmatic Markers in British English: Meaning in Social Interaction. Cambridge University Press, Cambridge. doi: 10.1017/CBO9781139507110
- Crible, L (2017) Towards an operational category of discourse markers: A definition and its model. In: Pragmatic Markers, Discourse Markers and Modal Particles: New perspectives. John Benjamins, Amsterdam, pp 99–124. doi: 10.1075/slcs.186
- Fraser B (1999) What are discourse markers? Journal of Pragmatics 31: 931–952. doi: 10.1016/S0378-2166(98)00101-5
- Furkó BP (2014) Cooptation over grammaticalization: The characteristics of discourse markers reconsidered. Argumentum 10: 289–300.
- Furkó BP (2017) Manipulative uses of pragmatic markers in political discourse. Palgrave Communications 3/ 17054. doi:10.1057/palcomms.2017.54
- Furkó BP, Abuczki Á. (2014) English Discourse Markers in Mediatised Political Interviews. Brno Studies in English 40: 45-64. doi: 10.5817/BSE2014-1-3
- Furkó BP, Kertész A, Abuczki Á (forthcoming) Discourse Markers in Different Types of Reporting. In: Capone A (ed) Indirect Reports - Perspectives in Pragmatics, Philosophy and Psychology. Springer, Heidelberg.
- 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. doi: 10.1177/0957926510366198
- Rayson P, Archer D, Piao S, McEnery T (2004) The UCREL Semantic Analysis System. Paper given at Beyond Named Entity Recognition Semantic Labeling for NLP Tasks in LREC'04, Lisbon.
- Schourup L (1999) Discourse markers: tutorial overview. Lingua 107: 227–265. doi:10.1016/S0024-3841(96)90026-1
- Spooren W, Degand L (2010) Coding coherence relations: reliability and validity. Corpus Linguistics and Linguistic Theory 6(2): 241–266. doi:10.1515/cllt:2010.009
- Stenström AB (1990) Lexical items peculiar to spoken discourse. In: Svartvik J (ed) The London-Lund Corpus of Spoken English: description and research. Lund University Press, Lund, pp 137–175.

Köszönöm szépen a figyelmet!