Acta Universitatis Sapientiae, Philologica, **10**, 2 (2018) 155–170

DOI: 110.2478/ausp-2018-0020

The Boundaries of Discourse Markers – Drawing Lines through Manual and Automatic Annotation

Péter FURKÓ
Károli Gáspár University of the Reformed Church in Hungary (Budapest, Hungary)
Department of English Linguistics
furko.peter@kre.hu

Abstract. Discourse markers are non-propositional linguistic items that are notoriously difficult to identify as well as to categorize. We can observe several borderline phenomena and overlaps with other formal and functional categories, e.g. inserts, adverbials, contextualization cues, pragmatic force modifiers, etc. By way of addressing such overlaps as well as the disambiguation between DM uses and their source categories, the paper presents a comparison of automated and manual annotation of oral discourse markers (DMs). Firstly, an overview of the criterial features of DMs that are relevant to disambiguation are presented. Secondly, the UCREL Semantic Analysis System (USAS) and its disambiguation methods are briefly discussed. In the third part of the paper, manual and automatic decisions about categorization are compared with a view to addressing the margin of error reported to apply in general semantic annotation as well as the question of what formal-functional properties of the relevant DMs might explain possible differences between manual and automatic annotation.

Keywords: discourse markers, automated semantic annotation, manual annotation, D-function ratio, inter-annotator agreement.

1. Introduction, the problem

Despite the rapidly growing body of research on discourse markers (DMs), experts in the field observe over and over again that there are still a number of fundamental questions that need to be answered (cf. e.g. Schourup 1999, Fraser 1999, Dér 2010, Heine 2013). Some of the issues include the lack of generally accepted terminologies and classifications, uncertainty regarding essential formal, semantic, and pragmatic characteristics, as well as the absence of a model in which DMs can be related to general linguistic categories in an integrated way.

In the present paper, I am going to address the issue of categorization and category membership, i.e. demarcating boundaries between lexical items that are DMs and distinguishing them from non-DM uses of the source categories. Describing the characteristics of the functional class of DMs and developing criteria for deciding for every given instance whether it is a DM or not have been major preoccupations in recent DM research. Authors usually provide exhaustive lists of the formal, functional, and stylistic criteria that are associated with DMs as a functional class (cf. e.g. Schourup 1999, Fraser 1999, Beeching 2016); still, few provide (and many claim it is impossible to provide) an exhaustive list of criterial features that can be used to identify all instances of DMs in a given corpus. An even more challenging task is to develop annotation software that can automatically identify DMs in oral discourse and filter out non-DM tokens of lexical items that are frequently used as DM types (e.g. adverbial uses of *well* or *now*, prepositional uses of *like*, etc.). Moreover, to date, no previous attempt has been made to use automatic means of identification involving semantic criteria and semantic fields.

Accordingly, the present paper will explore the utility of using an automated semantic tagging software, USAS, as a pre-annotation tool for the identification of oral DMs, including interpersonal as well as textual markers. After an overview of the formal and functional features that can be used for manual annotation and after comparing the results of manual and automatic annotation of selected DMs, the paper will argue that automatic semantic annotation (ASA) can be an effective tool for the disambiguation between DM and non-DM uses with regard to certain items but needs to be complemented by extensive manual error correction and filtering.

2. Characteristics of DMs, criteria for DM status[1]

2.1. Non-propositionality and optionality

Many scholars (cf. Schourup 1999) consider non-propositionality (non-truth-conditionality) as a sine qua non for DM status; yet, others include propositional items such as *then* and *after that*. While it is generally agreed that certain DMs (e.g. *well*, *however*, etc.) contribute nothing to the truth-conditions of the proposition expressed by an utterance, the non-truth conditionality of others (*frankly, I think*) have generated a great deal of controversy (cf. Infantidou-Trouki 1992).

Blakemore (1987: 106) argues that a distinction has to be made between *truth-conditional* and *non-truth-conditional* meaning, on the one hand, and *conceptual* vs *procedural* meaning, on the other. Thus, many of the controversies

[1] Parts of Section 2 of the paper were previously published in Furkó 2014.

stem from the fact that certain scholars confuse the two distinctions and use them interchangeably. Schourup (1999), for example, uses the *compositionality* test to argue in favour of the *truth-conditionality* of *in addition*:

(1a) Owens is a respected drama critic. I tell you *in addition* that she has written …
(1b) Owens is a respected drama critic. *In addition*, she has written …

While *in addition* is indeed truth-conditional, the above test would predict that *frankly* is also truth-conditional, whereas Blakemore (2002) would argue that DM uses of *frankly* are non-truth-conditional but *conceptual*. It is, therefore, important to point out that the compositionality test will be a useful tool in deciding whether individual DMs have *conceptual* or *procedural* meaning; the *truth-functionality* of DMs is tested more efficiently in terms of whether they retain their original meaning when embedded in *if-clauses* or under the scope of factive connectives such as *because*:

(2a) *Allegedly / Obviously / Frankly*, the cook has poisoned the soup.
(2b) If the cook has *allegedly* / ?*obviously* / **frankly* poisoned the soup, we can eat the meal without worrying.
(2c) We shouldn't eat the soup, because the cook has *allegedly/?obviously/ *frankly* poisoned it.

The uncertainty with regard to whether or not *obviously* retains its original meaning in (2c) once again suggests that the truth-functionality–non-truth functionality distinction should be viewed as a *continuum*, rather than a dichotomy, which is consistent with the finding in grammaticalization theory that due to the diachronic grammaticalization processes that are synchronically manifested in the use of pragmatic markers there is a fuzzy boundary between uses that are non-truth-conditional and (omissible) and those that are not (for a detailed discussion, cf. Blakemore 2002 and Andersen 2001).

Optionality as a distinguishing feature is in many respects derivative of the previously discussed criterion of non-propositionality; DMs are considered optional from the perspective of sentence meaning because their absence does not change the conditions under which the sentence is true.

There are, however, two further senses in which DMs are claimed to be optional. Firstly, they may be seen as *syntactically optional* in the sense that the removal of a DM does not alter the grammaticality of its host sentence. Secondly, they are optional in the sense that if a DM is omitted, the relationship it signals is still available to the hearer though no longer explicitly cued (cf. Schourup 1999: 231).

The above statement does not entail that DMs are useless; rather, it reflects the view according to which DMs guide the hearer toward a particular interpretation of the connection between a sequence of utterances and at the same time rule out unintended interpretations.

2.2. Context-dependence

DMs' extreme context-dependence is frequently identified with their inherent indexicality. Aijmer, for example, considers indexicality as the most important property of DMs, a property whereby DMs are linked to attitudes, evaluation, types of speakers, and other features of the communicative situation (cf. Aijmer 2002: 5). In this respect, DMs can be compared to deictics, i.e. another borderline phenomenon can be observed if we look at some of the definitions of deictic expressions, which often overlap with those of DMs. Both categories are usually defined in terms of context-dependence, i.e. in terms of having meaning only by virtue of an indexical connection to some aspect of the speech event (cf. e.g. Sidnell: 1998). Levinson (2004), in fact, considers DMs as discourse deictics, other subgroups including spatial, temporal, and social deictics.

Similarities between indexicals and DMs are also recognized by proponents of Relevance Theory. Carston, for example, notes that the two seemingly disparate phenomena are brought together by the fact that both encode a *procedure* rather than a *concept* and both play a role in guiding the hearer in the pragmatic inferential phase of understanding an utterance (Carston 1998: 24). The difference between the two sets of phenomena, according to Carston, is that indexicals constrain the inferential construction of *explicatures* and DMs (discourse connectives in RT terms) constrain the derivation of *implicatures* (in other words, intended contextual assumptions and contextual effects).

2.3. Multifunctionality

In addition to playing a role in pragmatic inferencing, individual DMs are also associated with a plethora of functions, including hedging and politeness functions. What is more, they can also be salient in conversational exchanges as openers, turn-taking devices, hesitation devices, backchannels, markers of topic shift and of receipt of information, and so on (cf. e.g. Beeching 2016: 4ff). DMs are multifunctional and ambiguous by design since there is a lot of *interpersonal* and *discourse* burden on their signalling capacity. DMs signal interpersonal and discourse functions simultaneously, and thus they are ambiguous between the two levels; on the other hand, they are vague with regard to signalling particular relations on a given level as well (ibid.).

The multifunctionality of DMs also brings up the question of whether different uses of a given marker are to be considered incidental and unrelated (maximalist approach) or motivated and related (minimalist approach) and whether there is an invariant "core meaning" of DMs that is context-independent and preserves some components of the lexeme's original semantic meaning. Since the focus of the present paper is on finding the boundaries between DM and non-DM uses of a given item, further discussion will not ensue on the multifunctionality of DM uses. However, multifunctionality can be used as an important criterial feature in the course of manual annotation, as we will see in the following sections.

2.4. Weak clause association, phonological reduction, variable scope

It is frequently observed in the literature that DMs usually occur either outside the syntactic structure or loosely attached to it (cf. e.g. Crible 2017). Quirk et al. classify many linguistic items that are elsewhere included among DMs as *conjuncts* (e.g. *nonetheless*), which are considered to be clause elements but to have a detached role relative to other, more closely interrelated clause elements such as subject, complement, and object: "Conjuncts are more like disjuncts than adjuncts in having a relatively detached and 'superordinate' role as compared with other clause elements" (Quirk et al. 1985: 631). In addition, some of the items that Quirk et al. refer to as "disjuncts" (e.g. *obviously*, sentence-initial *surprisingly* and *frankly*) also display a whole range of properties associated with the functional class of DMs.

It is important to note that the property of weak clause association is relative to elements *external* to the DM's form since several DMs clearly have their own internal syntactic structure (e.g. *on the other hand*) and others (e.g. *y'know, I mean*) are clausal from a syntactic point of view despite the fact that they are no longer considered to be compositional but procedural (cf. e.g. Furkó 2014).

Weak clause association is frequently discussed in relation to phonological independence: DMs often constitute independent tone units or are set off from the main clause by 'comma intonation' (cf. Hansen 1997: 156).

Adding *weak clause association* and a corresponding *lack of intonational integration* to our list of criteria is also justified from the perspective of grammaticalization theory. An important clause of the definition of grammaticalization states that it takes place in special *morpho-syntactic environments*. In the case of DMs, this environment can be associated with sentence-initial position, hence many scholars regard quasi-initiality as yet another distinguishing feature of DMs (cf. e.g. Schourup 1999). However, once DMs enter an advanced stage of grammaticalization, they become syntactically independent and can appear at various parts of the sentence, with an accompanying 'comma intonation'; thus, this criterion is not always helpful in the course of manual annotation.

DMs' position in an utterance also influences their *scope*, which is *variable*, as is illustrated by (3a) and (3b):

> (3a) Interviewer: I know how close you are to your mom. How old is she?
> Interviewee: *Well*, she probably doesn't want me to say…
> (3b) You're not going to have quality if you can't sleep and you itch and you bitch and you weep and you cry and you bloat and you can't remember anything and you don't have a, *well*, sex drive. (examples taken from Furkó: 2014)

As the examples above show, the size of the linguistic unit *well* can take in its scope ranges from a whole sentence to a single word. Waltereit (2006) observes that this variability is a remarkable property, but it is not an exclusive feature of DMs since conjunctions as a word-class (and even some individual conjunctions as a lexical item) can also have variable scope, giving the following sentences as examples:

> (4a) Ed and Doris loved each other.
> (4b) Ed worked at the barber's, and Doris worked in a department store.

In (4a), *and* has scope over two NPs; in (4b), it has scope over two clauses. However, the difference between *and* used as a conjunction and its DM use lies in the fact that the scope of the conjunction *and* can always be determined in *grammatical terms*. It could be defined as ranging over two constituents of the same type adjacent to *and*, which, in turn, make up a constituent of again the same type. The scope of DMs, in contrast, cannot be determined in grammatical terms, as is clear from (5) below:

> (5) My husband got a notice t'go into the service
> ***and*** we moved it up.
> ***And*** my father died the week … after we got married.
> ***And*** I just felt, that move was meant to be. (Schiffrin 1987: 53, emphasis in the original)

Schiffrin (1987) concludes that *and* has "freedom of scope", rather than "variable scope", since "we can no more use *and* to identify the interactional unit that is being continued than we can use *and* to identify the idea that is being coordinated" (Schiffrin 1987: 150).

Traugott (1995) relates the feature of variable scope to grammaticalization and argues that in addition to Nominal clines (nominal adposition > case) and verbal clines (main verb > tense, aspect, mood marker), which are "staples of grammaticalization theory", a further cline: Clause internal Adverbial > Sentence Adverbial > Discourse Particle should be added to the inventory (Traugott 1995: 1). According to Traugott, this cline involves increased *syntactic freedom* and *scope*.

2.5. Procedural meaning/non-compositionality

Although most scholars (for an overview, cf. Schourup 1999) treat non-compositionality as a property of DMs per se, Blakemore (2002) associates DMs with procedural meaning and uses non-compositionality as a *test* to decide whether individual items are conceptual or procedural.

Blakemore also claims that if DMs are synonymous with their non-DM counterparts, they encode conceptual meaning. Thus, *seriously* and *in other* words in (6a) and (7a) encode a concept parallel to (6b) and (7b), respectively. On the other hand, *well* (as in 8a) encodes a procedure since it is not synonymous with *well* in (8b):

(6a) *Seriously*, you will have to leave.
(6b) He looked at me very *seriously*.
(7a) *In other words*, you're banned.
(7b) She asked me to try and put it *in other words*.
(8a) A: What time should we leave?
B: *Well*, the train leaves at 11.23.
(8b) You haven't ironed this very *well*.

A second test Blakemore uses is to see if a given item can combine with linguistic items encoding conceptual meaning to produce complex expressions.

As far as the question of synonymity is concerned, it is important to note that the fact that, on the basis of native intuitions, no correspondence can be found between the adverbial *well* and its DM counterpart does not mean that such a relationship is absent (cf. e.g. Furkó 2013). Native intuitions, naturally, disregard diachronic aspects of individual lexical and grammatical items, and it is exactly these aspects that account for the fuzziness of the category of DMs.

2.6. High frequency, orality, stigmatization

In this section, some of the stylistic features, core members of the functional class of DM display are considered. While semantic-functional properties are more important in determining class membership than formal and stylistic ones, stylistic criteria can also be helpful in determining DM status and differentiating between DM and non-DM tokens.

It is important to note that the high frequency of use is the *backbone* of various processes of grammaticalization as well as pragmaticalization (cf. e.g. Furkó 2014). In other words, the more frequently an item is used, the more likely it is that its formal-functional properties are going to change, and once it has entered the process of grammaticalization, the faster it is going to go through the substages of that process.

A number of studies on DMs observe that the frequency of DMs can be primarily observed in speech (e.g. Beeching 2016); what is more, one of the most salient features of oral style is the use of items such as *well*, *right*, *ok*, *you know*, etc. For example, in their classical study, Brown and Yule (1983: 17) label *well*, *erm*, *I think*, *you know*, *if you see what I mean*, *I mean*, and *of course* "prefabricated fillers", when drawing up a list of contrasting characteristics of spoken and written language. They also point out that these items' overuse is often stigmatized by prescriptivists (ibid.).

Despite the stigmatization of many oral DMs, it is easy to illustrate the meaningfulness and the distinctive (as opposed to random) use of even the two most used DMs, *you know* and *I mean*. As Fox Tree and Schrock (2002: 731) illustrate, it matters where *you know* or *I mean* appear in an utterance, and they are not interchangeable:

> (9a) Original: me and the Edinburgh girl got together after dinner late in the evening and decided they'd really got us along to make it look right, *you know* they had after all had candidates from other universities.
> Alternative: me and the Edinburgh girl got together after dinner *you know* late in the evening and decided they'd really got us along to make it look right, they had after all had candidates from other universities.
> (9b) Original: but I don't think it's feasible. *I mean* I know this is the first time I've done it, and I'm not in a main line paper, but I'm sure it'll take me all my time to do it in three weeks.
> Alternative: but I don't think it's feasible. I know *I mean* this is the first time I've done it, and I'm not in a main line paper, but I'm sure it'll take me all my time to do it in three weeks (example taken from Fox Tree and Schrock 2002: 731).

In (9a) Original, *you know* comments on what is meant by "look right", whereas in (9a) Alternative it comments on what "after dinner" means (in other words, they differ in what they take within their *scope*; see Section 2.5 above). In (9b) Original, *I mean* comments on why the speaker says "I don't think it's feasible", without overwriting the statement, but in (9b) Alternative *I mean* comments on "I know", retrospectively treating it as a false start.

In addition, as both manual and automatic annotation will illustrate, there is no principled basis on which one could exclude from the functional class of DMs connectives such as *however*, *after all*, *consequently*, and a whole range of other items characteristic of written style, some of which (e.g. *besides*, *however*, *moreover*) are in fact included in Brown and Yule's above mentioned list of characteristics of *written* language.

3. Automatic semantic annotation: Testing its methods and precision

There are a variety of computerized semantic tagging (CST) systems, including artificial intelligence-based, knowledge-based, corpus-based, and semantic taxonomy-based systems (for an overview, cf. e.g. Prentice 2010). The present analysis draws on the results gained from the UCREL Semantic Analysis System (USAS), which has the major advantage of combining these approaches. Furthermore, USAS groups lexical items in terms of a taxonomy of semantic fields and assigns semantic categories to all words, including grammatical and other procedural (non-propositional) items, which is relevant for the present study in view of the fact that the lexical items under scrutiny are highly procedural and semantically bleached (cf. Section 2 above).

USAS system uses an automatic coding scheme of 21 semantic fields, subdivided into 232 subcategories. For reasons of brevity, only the tags that have been associated with the DM types under analysis will be discussed – the complete coding scheme can be found at http://ucrel.lancs.ac.uk/usas/. USAS uses disambiguation methods including part-of-speech tagging, general likelihood ranking, multi-word-expression extraction, domain of discourse identification, and contextual rules (for a detailed discussion, cf. Rayson et al. 2004). Previous evaluations of the accuracy of the system reported a precision value of 91% (ibid.), i.e. a 9% margin of error applying to lexical items across the board (including propositional and non-propositional items).

The research questions in the present study are as follows:

1. Are the disambiguation methods USAS uses sufficient for filtering out non-DM tokens of the most frequent DM types?

2. Does the margin of error reported to apply in general apply to the identification of DMs as well?

3. Are individual DMs identified/tagged with a similar margin of error?

4. If individual DMs are tagged with varying precisions by USAS, what formal-functional properties of the relevant DMs might explain the differences?

4. Corpus and methodology

In the course of the research, two sub-corpora of the same size (100,000 words each) were used:

– a corpus of the official transcripts of 37 confrontational type of mediatized political interviews (henceforth MPI sub-corpus) selected from BBC's *HardTalk* and *Newsnight* (available at: http://bbc.co.uk);

– a corpus of the official transcripts of 50 celebrity interviews (henceforth CI sub-corpus) downsampled from CNN's *Larry King Live* (available at: http://www.cnn.com).

The two sub-corpora have been extensively studied in previous research; thus, the results of automatic tagging have been compared to findings based on manual annotation and a combination of quantitative and qualitative methods (cf. Furkó and Abuczki 2014, 2015). Previous research was aimed at finding genre-specific patterns of DM use in the two sub-corpora, which has informed the present paper in terms of the D-values as well as the functional distribution of individual lexical items (see Section 5 below).

The research process has been as follows: in order to identify and compare the USAS tags of oral DMs in the two sub-corpora, the semantic tags assigned to frequent DMs (e.g. *I mean*, *you know*, *in other words*, *so*, *well*) were considered, and then these semantic tags were used to identify further types and tokens relevant to discourse marking. What was found was that 95.1% of the instances of DMs trawled from the two sub-corpora through this method are either tagged with Z4, described in the USAS manual as the "discourse bin" (including items such as *oh*, *I mean*, *you know*, *basically*, *obviously*, *right*, *yeah*, *yes*) or with A5.x, described as "evaluative terms depicting quality" (including DMs such as *well*, *OK*, *okay*, *good*, *right*, *alright*). The frequency of the relevant tags across the two sub-corpora was compared as well as the ratio between DM-relevant tags (i.e. Z4 and A5.x) and non-DM relevant tags (e.g. B2, I1.1, T1.3, etc.; see below for details).

In the second stage, a representative sample of 400 tokens in the MPI sub-corpus were manually annotated using a numeric code of 1 for DM and 2 for non-DM tokens with a view to comparing the results of automatic and manual tagging. When deciding if an individual token is a DM or not, the criterial features described above (see Section 2) were applied by a single expert annotator. The tokens that were selected for the sample were weighted for their frequency in the corpus, while DM and non-DM tokens were included in equal proportions. For example, the 429 tokens of *well* comprise 19.6% of all automatically tagged items, and thus 78 tokens, (39 A5.1-tagged and 39 non-A5.1 tagged by USAS) were included in the sample.

5. Findings

Table 1 below summarizes the raw frequency of the relevant lexical items' DM- and non-DM-related USAS tags. Since both sub-corpora were compiled in a way that they are of the same size of 100,000 words, these raw frequencies can be compared as if normalized.

Table 1. *Summary of DM- and non-DM-related semantic tags assigned to the most frequent DM types in the MPI and CI sub-corpora*

Lexical item	Frequency of DM-related tag in the MPI	Frequency of DM-related tag in the CI	Frequency of non-DM-related tag in the MPI	Frequency of non-DM-related tag in the CI
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

As a first step, the ratio of DM and non-DM tokens of individual items was compared with the results of previous research, in the course of which DMs in the same sub-corpora were manually annotated (cf. Furkó and Abuczki 2014). In order to gauge the categorial multifunctionality of DMs, the measure of D-function ratio or D-value (a term proposed by Stenström 1990) was used. An individual item's D-value is calculated as a quotient of the number of tokens that fulfil discourse-pragmatic functions and the total number of occurrences in a given corpus. The D-value of *oh*, for example, is 1 (100%) in the London-Lund Corpus since it is used exclusively as a DM, whereas *well* showed a D-value of 0.86 as 14% of its tokens serve non-DM (adverbial, nominal, etc.) functions (ibid.).

If we calculate the D-values of individual DMs based on the above values and compare them to the findings of previous research, we see that the results of automatic annotation and manual annotation converge to a great extent. *Mean*, for example, has a D-value of 0.808 in the MPI corpus based on automatic annotation (calculated as the number of Z4 tags divided by all tokens of *mean*, i.e. 141), while manual annotation yielded a D-value of 0.797 (cf. Furkó and Abuczki 2014: 50). Similarly, manual annotation yielded a D-value of 0.82 for *well* in the MPI corpus (Furkó and Abuczki 2014: 54), while *Table 1* yields a D-value of 0.839 for this lexical item (360 Z4 tags divided by the total number of tokens, i.e. 429).

The table also correctly predicts that most of the lexical items under scrutiny have higher D-values in the CI sub-corpus than in the MPI sub-corpus, which is explained by the fact that there is a higher degree of conversationalization in celebrity interviews, i.e. they are more similar to spontaneous, informal, face-to-face conversations (cf. Furkó 2017). For example, the D-value of *well* is 0.92, the D-value of *mean* is 0.851 in the CI sub-corpus based on automatic annotation (312 A5.1 tags divided by a total of 339 tokens, 201 Z4 tags divided by a total of 236 tokens, respectively).

In the second stage of the research, a representative sample of tokens in the MPI were manually annotated using numeric 1 for DM tokens and 2 for non-DM uses. With a view to comparing the results of automatic and manual annotation, all DM-related tags (Z4 and A5.x) yielded by USAS were re-coded as numeric 1, while non-DM tags (B2, I1.1, T1.3, etc.) were re-coded as 2. Consequently, the extracted list of the corresponding manual and automated tags was entered into a reliability calculator (Freelon's ReCal 2 for 2 coders) in order to calculate inter-annotator agreement statistics. *Table 2* below shows the result.

Table 2. *Inter-annotator agreement between automated and manual tagging of DM/non-DM tokens*

	Percent Agreement	Scott's Pi	Cohen's Kappa	N Agreements	N Disagreements	N Cases	N Decisions
Variable (DM/non-DM)	92.75	0.854519	0.854527	371	29	400	800

Although the above intercoder agreement values appear high (cf. Spooren and Degand 2010), it is important to note that there is a great degree of variation in the precision with which individual DMs are tagged by USAS. On the one hand, there are DMs, such as *I mean* and *you know,* whose DM and non-DM uses are disambiguated with surprising precision (resulting in a kappa score of <.98, i.e. close to perfect intercoder agreement between USAS and the human annotator). This is probably due to two of the disambiguation methods USAS applies: firstly, its multi-word-expression extraction algorithm and its core component of MWE lexicon (cf. Rayson et al. 2004) and, secondly, the fact that POS tagging enables the parser to differentiate between syntactically integrated tokens that are monotransitive (and are thus followed by their nominal or clausal complements) and syntactically non-integrated ones that are marked by the absence of complements. On the other hand, there are lexical items that are invariably tagged with the same (sometimes DM-relevant, other times non-DM relevant) tags regardless of their syntactic (non-)integration and functional scope. For space considerations, only two examples will be given, one for DM-relevant invariant tagging and one for non-DM-relevant invariant tagging.

An example for the former is *actually*, which might be used as a DM that has the ensuing discourse unit in its scope (10a) or as an adverbial modifier that has scope over the verb it modifies as in 10b below (all extracts are from the USAS-tagged CI corpus, emphases are mine):

> (10a) 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
> (10b) 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

An example for non-DM relevant invariant tagging is *now*, which can be used as a DM that marks topic shift (11a) or as a circumstance adverb (11b). However, USAS does not usually distinguish between DM and non-DM uses of *now*, both being labelled as T1.1.2, i.e. as "general terms relating to a present period/point in time":

> (11a) 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*** what_Z8 about_Z5 her_Z8f and_Z5 the_Z5 Kennedy_Z1mf 's_Z5 ?
> (11b) 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. Conclusions, utility and limitations of using USAS as a pre-annotation tool

In the paper, it was argued that discourse markers are notoriously difficult to identify for humans and computers alike; there are several borderline phenomena, fuzzy boundaries, and cases of ambiguity resulting from DMs' inherent, criterial features. In answer to the research questions posed in Section 3 above, it can be observed that the disambiguation methods' automatic annotation uses are efficient for filtering out non-DM tokens of the most frequent DM types: thus, automatized annotation enables the researcher to obtain an adequate global picture of the D-values of most of the lexical items that are frequently used as DM types.

We have also seen that the margin of error reported to apply in general also applies to the identification of DMs collectively, and, in the case, of multi-word units, such as *you know* and *I mean*, individually as well. However, we find a great degree of variation in the precision/margin of error with which non-multi word DMs are tagged. Such varying precisions are mostly due to DMs' criterial features of source category layering, syntactic non-integration, variable/functional scope, all of which challenge the disambiguation methods USAS applies, with special reference to part-of-speech tagging, general likelihood ranking, and multi-word-expression extraction.

While DMs will continue to puzzle humans and computers alike, we can safely say that automatized methods can take us one step closer to drawing boundaries between propositional and non-propositional, syntactically-semantically integrated and interpersonal-textual uses of lexical items, which, in addition to some issues in genre analysis discussed above, might have important implications for applied linguistic concerns as far apart as language acquisition, natural language processing, and critical discourse analysis (cf. Furkó 2017).

References

Aijmer, Karin. 2002. *English Discourse Particles: Evidence from a Corpus*. Amsterdam and Philadelphia: John Benjamins.

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

Arial, Mira. 1998. Discourse markers and form-function correlations. In Jucker, Andreas H. and Ziv, Yael. (eds), *Discourse Markers: Descriptions and Theory*. Pragmatics and Beyond Series, 57, 223–260. Amsterdam and Philadelphia: John Benjamins.

Beeching, Kate. 2016. *Pragmatic Markers in British English – Meaning in Social Interaction*. Cambridge: Cambridge University Press.

Blakemore, Diane. 1987. *Semantic Constraints on Relevance*. Oxford: Blackwell.

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

Brown, Gillian–Yule, George. 1983. *Discourse Analysis*. Cambridge: Cambridge University Press.

Carston, Roby. 1998. The semantics/pragmatics distinction: A view from relevance theory. *UCL Working Papers in Linguistics* 10: 1–30.

Crible, Ludivine. 2017. Towards an operational category of discourse markers: A definition and its model. In Chiara Fedriani and Andrea Sansó (eds), *Pragmatic Markers, Discourse Markers and Modal Particles: New Perspectives*, 99–124. John Benjamins: Amsterdam.

Dér, Csilla Ilona. 2010. On the status of discourse markers. *Acta Linguistica Hungarica* 57(1): 3–28.

Fox Tree, Jean E.–Schrock, Josef C. 2002. Basic meanings of *you know* and *I mean*. *Journal of Pragmatics* 34: 727–47.

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

Furkó, Bálint Péter. 2013. The presence and absence of pragmatic markers in naturally-occurring and scripted discourse. In Katarína Labudova and Nóra Séllei (eds), *Presences and Absences – Transdisciplinary Essays*, 23–38. Tyneside: Cambridge Scholars Publishing.

—— 2014. Cooptation over grammaticalization – The characteristics of discourse markers reconsidered. *Argumentum* 10: 289–300.

—— 2017. Manipulative uses of pragmatic markers in political discourse. *Palgrave Communications* 2017/54. *https://www.nature.com/articles/palcomms201754*. (Last accessed on: 6 September 2018).

Furkó, Bálint Péter–Ágnes Abuczki. 2014. English Discourse Markers in Mediatised Political Interviews. *Brno Studies in English* 40(1): 45–64.

—— 2015. A contrastive study of English and Hungarian discourse markers in mediatised interviews and natural conversations. *Sprachentheorie und Germanistische Linguistik* 25(2): 151–187.

Hansen, Maj-Britt Mosegaard. 1997. *Alors* and *donc* in spoken French: A reanalysis. *Journal of Pragmatics* 28: 153–187.

Heine, Bernd. 2013. On discourse markers: Grammaticalization, pragmaticalization, or something else? *Linguistics* 51(6): 1205–1247.

Infantidou-Trouki, E. 1992. Sentential adverbs and relevance. *UCL Working Papers in Linguistics* 4: 193–214.

Levinson, Stephen C. 2004. Deixis and pragmatics. In Laurence Horn, Gregory Ward (eds), *The Handbook of Pragmatics*, 97–121. Oxford: Blackwell.

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

Quirk, Randolph–Greenbaum, Sidney–Leech, Geoffrey–Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rayson, Paul–Archer, Dawn–Piao, Scott–McEnery, Tony. 2004. The UCREL Semantic Analysis System. Paper given at *Beyond Named Entity Recognition Semantic Labeling for NLP Tasks in LREC '04*, Lisbon.

Schiffrin, Deborah. 1987. *Discourse Markers*. Oxford: Blackwell.

Schourup, Lawrence. 1999. Discourse markers: Tutorial overview. *Lingua* 107: 227–265.

Sidnell, Jack. 1998. Deixis. In Jef Verschueren et al. (ed.), *Handbook of Pragmatics, 1998 Installment*. Amsterdam, Philadelphia: John Benjamins.

Spooren, William–Degand, Liesbeth. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2): 241–266.

Stenström, Anna-Brita. 1990. Lexical items peculiar to spoken discourse. In Jan Svartvik (ed.) *The London-Lund Corpus of Spoken English: Description and Research*, 137–175. Lund: Lund University Press.

Traugott, Elizabeth Gloss. 1995. The role of the development of discourse markers in a theory of grammaticalization. Paper given at the *12th International Conference on Historical Linguistics*. Manchester, 13–18 August 1995.

Waltereit, Richard. 2006. The rise of discourse markers in Italian: A specific type of language change. In Kerstin Fischer (ed.), *Approaches to Discourse Particles*, 61–76. Amsterdam: Elsevier.