Towards an operational category of discourse markers: A definition and its model

Running head: Towards an operational category of discourse markers

Abstract

The field of discourse markers (DMs) studies suffers from lack of consensus on the limits and definition of the category. There seems to be a crucial need for onomasiological studies that account for every kind of DM in crosslinguistic data. This study presents a proposal for an operational, corpus-based definition of DMs that addresses several theoretical and methodological shortcomings in the field. I claim that any categorical definition is only useful insofar as it is endorsed by an empirical model of identification and annotation. Such a model will be described and illustrated by relevant authentic examples from a pilot study on a comparable corpus of French and English interviews.

Key-words: discourse markers; linguistic categorization; corpus-based pragmatics; annotation protocol; bilingual corpus

1. Introduction

Discourse marker (DM) research today, after several decades of flourishing productivity, still faces many terminological, theoretical and methodological issues which restrain large-scale progress in the field, despite the multiplicity of theoretical frameworks and approaches taken by many valuable works (e.g. Brinton 1996; Fischer 2006; Schiffrin 1987; Waltereit and Detges 2007 to name but a few). In fact, these pragmatic elements (such as "because", "you know", "well", or "so" to give a few often cited examples) are defined by their heterogeneity and multifunctionality, which explains the proliferation of conflicting definitions. The field suffers from lack of consensus on the limits of the category, its definition and what it includes (Schourup 1999). Such differences make comparisons of results difficult, since there is usually only limited overlap between the scope of the various studies. Reasons for these discrepancies may lie in the choice of theoretical framework (e.g. coherence-based approach vs. Relevance Theory (Rouchota 1996)), restriction of items under consideration, type of data (e.g. medium, register), method and purpose of annotation, and possibly others.

In light of this rather chaotic situation, there seems to be a crucial need for functional, paradigmatic studies that include every kind of DMs, possibly in multilingual approaches for better generalization. Such endeavors would provide a solid basis for comparative or contrastive analysis between

languages and frameworks when the current state of the art is confronted to a certain particularism of approaches, responsible for a lack of communicability and exchange of research.[1] The present work puts forward a proposal for an operational definition of DMs that addresses several shortcomings in the field. The purpose of this article is thus twofold: firstly, to take a stand on methodological best practices for DM studies, and in general for corpus-based pragmatics; secondly, and more concretely, to illustrate these milestones by exposing the benefits of a corpus-based crosslinguistic definition of the category of DMs and its matching annotation scheme.

The following sections will offer a critical review of existing definitions, before presenting my proposal of an operational corpus-based definition of DMs. The remainder of the article will support the claim that any categorical definition is only useful insofar as it is endorsed by an empirical model of identification and annotation. Such a model will be briefly described and illustrated by authentic examples from a pilot study on *Backbone* (Kurt 2012), a comparable corpus of French and English interviews.

2. Definitions in contest

[1] This ambition meets the program of the COST Action IS1312 "TextLink: Structuring Discourse in Multilingual Europe", http://textlinkcost.wix.com/textlink, chair: L. Degand.

Ever since Schiffrin's (1987) seminal work focusing on discourse markers, numerous authors have offered their contribution to define this category. While we would expect converging and deeper insights into what constitutes membership in the DM category, the actual picture shows a rather confusing patchwork of interesting yet conflicting approaches to what is now feared by students (and others) as a complex object of study. The goal of this section is to dissect a selection of definitions in order to find what differentiates them, and what is their common core.

2.1 Terminology

Let us start with a terminological note. One major revealing sign of the complexity of the DM category is the proliferation of terms used to name it. Since other authors provide a list of all or most existing propositions (e.g. Aijmer & Simon-Vandenbergen 2006, Fraser 1999), I will not expand on the merits and limitations of each. I will rather focus here on the dichotomy between "discourse marker" and "pragmatic marker" (Furkó 2005) and justify my choice for the former.[2]

I would first like to claim that an "optimal" terminological choice is both possible and necessary, and that future works should strive towards adopting

[2] Other terms will be discussed in the next section where they will be reviewed for the categorical definition they refer to.

a single label to avoid further confusion on this first basic stage.[3] Such a label should be representative of all members of the category by pinpointing what they have in common, while drawing the boundaries with other categories. This very issue is at the core of my preference for "discourse marker" over "pragmatic marker" (PM).

While "PM" is used elsewhere (e.g. Aijmer & Simon-Vandenbergen 2011, Brinton 1996) to refer to a similar set of elements as the one at stake here, its main flaw is the implied inclusion of anything "pragmatic", i.e. any procedural element contributing to the interpretation of context by other means than semantic decoding. Here I follow many authors (e.g. Hansen 2006, Waltereit & Detges 2007) who assign to "PMs" a much broader definition, turning this label into an overarching category including more than DMs. As Hansen (2006: 28) explains:

> *Discourse marker* should be considered a hyponym of *pragmatic marker*, the latter being a cover term for all those non-propositional functions which linguistic items may fulfil in discourse. Alongside discourse markers, whose main purpose is the maintenance of what I have called "transactional coherence", this overarching category of functions would include various forms of interactional markers, such

[3] The "TextLink" network suggests in this matter the recourse to a more flexible term, "discourse-structuring devices", to account for the multilingual flexibility of the category.

> as markers of politeness, turn-taking etc. whose aim is the maintenance of interactional coherence; performance markers, such as hesitation marker; and possibly others.

However, unlike what Hansen seems to suggest, the actual situation is not as simple as drawing clear-cut lines between pragmatic categories, and these very frequent items of discourse are rather on a continuum of "pragmaticality", behaving like members of one functional group or another in different contexts of use. Nevertheless, this does not mean that we should stop trying to identify in an operational yet representative way the object under investigation. The term "discourse marker" will thus be used in the remainder of this article for the reasons mentioned above, hence keeping in line with the majority of works in the field, although, as we will now see, they sometimes refer to quite different sets of items.

2.2 Partial and overlapping definitions

Given the profusion of literature and the already numerous reviews of definitions, this section will not attempt to provide yet another list and does not claim exhaustivity. The aim, instead, is to offer a selection of proposals relevant for the present discussion of partial and overlapping categorizations. The selected definitions that will now be discussed and compared do not all

refer to “discourse markers” in their own terminology. However, they either claim to be references for the DM category, or are used as such in other works, when in reality they only cover a subtype of elements or overlap with another category. These definitions are shared in the field but rarely with the same understanding.

2.2.1 Discourse markers and pragmatic markers

Starting from the most general term and as mentioned above, PMs (as in Brinton 1996) refer to a broader category of procedural elements which perform very distinct functions, hardly summarized other than by being pragmatic, which is rather too inclusive and not so informative. Brinton (1996) makes an inventory of PMs in English which shows rather surprising members, such as interjections (“ah”), certain adverbs like “really”, “basically” or “almost”, response signals, etc. It includes certain conjunctions such as “and”, “because”, “if” or “so”, but not similar expressions like “although” or “while”. This inventory thus appears both incomplete and too broad to satisfy a stable categorical definition. One might even say that the very principle of an inventory defies the purpose of a functional approach, which advocates for a critical inclusion of potentially emerging forms.

Diewald (2006) provides a similar, very broad definition of the category (although referred to as “discourse particles”) which “encompasses response signals, segmentation signals, interjections, hesitation markers, etc.” (2006:

406). Any large-scale analysis, especially corpus-based studies of such inclusive categories, would result in a highly heterogeneous selection of items, answering to very different rules of use, hence not particularly informative of the individual members. It might seem practical to group elements that are complex to distinguish, but the cognitive soundness and methodological efficiency of such an approach remain to be demonstrated. In fact, to my knowledge, no corpus study has ever identified and analyzed such a large range of items in authentic data. It would seem that the merit of the PM category is therefore mainly theoretical and metalinguistic, and does not correspond to an empirically-founded category of similar expressions in language use.

2.2.2 Discourse markers and modal particles

In a later article, Diewald (2013) draws a distinction between DMs and modal particles (MPs). This other close pragmatic category is very language-specific, and it is uncertain whether it exists for French or English for instance. According to Diewald, the main difference between DMs and MPs is that the former "relate non-propositional elements which are not textually expressed", while the latter "point to propositions and speech-act alternatives which are not textually expressed but treated as 'given'" (2013: 35). However, several of these criteria are debatable (e.g. the segments connected by a DM can and often are textually expressed and propositional) and do not

allow for efficient distinctions when facing authentic data (e.g. how to operationalize the "treated as given" criterion in corpora?).

Cuenca (2013) offers a more fine-grained, prescriptive definition of these terms by adding a third category to the picture: connectives. She talks of a continuum between discourse marking and modal marking, where connectives and MPs are formally similar (syntactically integrated, occupying a prototypical position, easily identifiable grammatical classes), MPs and DMs are functionally close (having, respectively, an attitudinal and metadiscursive function, with scope over one unit), and connectives and DMs share a discourse-connecting or structural function, "since connecting at text level is what prototypical DMs do" (2013: 193). Cuenca's cline of discourse marking is appealing and takes us a step closer to the formal disambiguation of DMs against close pragmatic categories, even though their functional behavior still requires some disentangling.

2.2.3 Discourse markers and subcategories

Finally, the majority of available definitions only cover what is here considered a subcategory of DMs. Four groups (at least) can be identified in the literature, and will briefly be discussed in this section.

The first, most frequently encountered notion is that of connectives, starting with Fraser (1996) and his four categories of so-called "DMs" which connect two segments with either a topic-change, contrastive, elaborative or

inferential relation. A similar approach is taken by the Penn Discourse TreeBank (PDTB) group who extends and specifies the possible relations between two "abstract objects" signaled by connectives, and provides a detailed protocol for their annotation in written corpora (Prasad et al. 2008). By using this definition of connectives as a reference for the whole category of DMs (which is often the case, especially in studies on writing), one excludes non-relational markers such as "you know", which cannot be said to actually connect two discourse units but modify only one utterance. Still, these very frequent speech-specific expressions are often included in DM definitions and show a number of criterial features detailed in Section 3.

Another recurring distinction in the literature concerns markers which signal a relation between real-world events on the one hand, and a relation between discourse events (assumption or speech acts) on the other. The former are called "semantic" by González (2005) and "real-world relation" by Lewis (2006), while the latter are "pragmatic" and "speaker-determined" respectively. Following this view, we should exclude DMs (or occurrences of DMs) signaling a relation between two events (e.g. semantic cause) and non-relational markers altogether (e.g. "you know"): "Discourse markers in this framework, then, are discourse relational and speaker oriented" (Lewis 2006: 55). This further restriction, within what others have called connectives, turns a functional specialization into a separate category, which is unjustified given the semantico-pragmatic similarity of, for instance, a semantic concession (as

in “I’m a teacher, but I have no qualifications”) and a pragmatic concession (as in “I’m a teacher, but I’m also a parent”).

A third partial definition, especially popular in the French academia, is Vincent’s (1993) term “*ponctuants*” or punctuators, to which she primarily assigns a segmentation function, similar to that of prosodic elements: “vocally marking certain prosodic facts, which partly form the coherence of discourse by delimiting the segments” (1993: 61, my translation). This definition typically covers so-called fillers like “well” or “hum”. The issue is that it excludes DMs signaling a discourse relation, having a more prominent semantic meaning, and other metadiscursive DMs such as “sort of” which do not have a segmentation role.

Finally, Hansen’s definition of DMs points to a fourth subcategory of what she calls “transactional coherence” (2006), which roughly corresponds to discourse relations, while other elements like markers of politeness or turn-taking contribute to the “interactional coherence” of discourse. However, “transactional” and “interactional” functions are very often performed by the same DMs such as “well” or “so”, which can both be used relationally (e.g. reformulative) and non-relationally (turn-taking). Her distinction of DMs from other types of PMs, as mentioned before, is insightful, but may thus be too restrictive.

2.3 Interim discussion

There is an obvious explanation which could almost justify this multiplicity of definitions if it did not hamper scientific progress, and that is the highly multi-faceted nature of the phenomenon. Forms and functions of DMs are indeed so varied that they are suitable objects of study for frameworks as different as Relevance Theory (Sperber & Wilson 1995), Segmented Discourse Representation Theory (SDRT, Asher & Lascarides 2003), diachronic accounts in terms of grammaticalization or pragmaticalisation (e.g. Bolly & Degand 2013, Bolly 2014, Brinton 1996, Degand & Evers-Vermeul 2015, Traugott 1995) or natural language processing (e.g. Marcu 1998, Stede 2014).

Despite these discrepancies, I identified three major features that are alluded to in most definitions and that would allow for a consistent, extensive definition of DMs: syntactic integration, functional scope and multifunctionality. The preliminary claim of this article is that, by combining the values of these three parameters in a flexible way, one can obtain a more satisfying definition of the category that applies to all its members.

3. Bridging the gap: a corpus-based definition of DMs

3.1 Methodology for an extensive-intensive definition

The definitions discussed in the previous section, as well as many others in the literature, are usually of two kinds: either a theoretical, usually quite abstract account of variables that might affect the behavior of DMs, or more in-depth case studies that specify a method but only for a certain type of elements or data. The first type is rarely operationalized, while the second is hardly reproducible on a larger scale (see Bolly et al. (this volume) for an exception). To bridge the gap between these two extremes, I propose a corpus-based definition of the category of DMs which has been designed and improved through application to authentic data, and which is thought to reconcile theoretical-empirical, quantitative-qualitative and extensive-intensive purposes. As Glynn (2010: 240) observed with special reference to DMs: "The challenge is to find ways to operationalize this infamously slippery object of study, semantics".

Using an existing definition as a reference for the selection of DM tokens in authentic data soon revealed inadequate, since none of them accounted for the category's diversity in speech, nor did they provide the necessary criteria to isolate the specificities of DMs against other pragmatic categories. In corpus-based pragmatics, I argue that empirical and practical considerations must be involved early in the theoretical stages of research so that the findings are consistent with the definition, especially when the scope of the study is

inclusive and strives towards generalization. The present contrastive and paradigmatic approach thus requires a corpus-based methodology.

The procedure for the elaboration of this definition is three-fold: first, a critical review of the literature with a selection of the recurring, most relevant criteria (see previous section); then several phases of revision upon a bilingual corpus; finally an annotation experiment which identified weaker criteria and solutions to strengthen them (Crible & Zufferey 2015). Confrontation with authentic data makes it possible to better define the boundaries with other elements that one might be tempted to include in the DM category (e.g. interjections, fillers, editing expressions) and to identify, with a body of examples from the corpus, the formal and functional criteria or conditions under which some "fluctuating" elements can or cannot be categorized as DMs. In a later step, our experiment provided confirmation and strengthening of the criteria, mainly by making explicit some assumptions and biases, in order to improve the replicability of the definition.

Working with bilingual corpus data allows one to question language-specific restrictions by imposing a reliable *tertium comparationis* (Jaszczolt 2003) equally valid in all languages concerned, thus revealing the limitations of a closed-list selection. A well-documented definition can provide both the flexibility (extensiveness) and the prescriptive criteria (intensiveness) that are necessary for a synchronic paradigmatic approach to DMs.

3.2 Defining DMs with syntax and functions

The present definition is a combination of two groups of criteria, *viz.* syntactic (integration and scope) and pragmatic (multifunctionality).[4] It can be stated as follows: DMs are a grammatically heterogeneous, multifunctional type of pragmatic markers, hence constraining the inferential mechanisms of interpretation. Their specificity as part of the PM category is to *function on a metadiscursive level as procedural cues to situate the host unit in a co-built representation of on-going discourse*.[5]

Syntactically, like many PMs, they are optional (i.e. can be removed without impairing the grammatical and/or semantic structure) and relatively mobile in the utterance. However, as opposed to modal particles or interjections, they come from very diverse grammatical classes. Their integration in the syntax depends mainly on this grammatical diversity: conjunctions will mostly be well integrated with a specific position, while particles and adverbs are much freer. Another formal criterion is their fixed form, which is the result of their grammaticalization process and high frequency: this restriction prevents the selection of expressions which are either idiosyncratic or too variable in form, such as variations of general extenders ("and all that kind of jazz"). Finally

[4] Scope is both syntactic and functional, since it relies heavily on syntactic constraints but can only be identified in reference to the semantico-pragmatic interpretation of which segment the DM applies to.

[5] "Metadiscursive" is preferred over "discursive" to better capture the notion of encoding the speaker's subjectivity towards discourse, their "comments" on the message.

they have a variable scope: relational markers may take scope over two textual, simple units, or two textual more complex units like a whole information unit, or between a contextual assumption and a textual unit, thus broadening Asher's (1993) notion of "abstract objects" to "context in this wider, nonlinguistic sense" (Hansen 2006: 25); as for non-relational markers, they only apply to one unit of variable size. In any case, the host unit must be autonomous both syntactically and semantically, i.e. there must be a finite or implicit predicate, which includes subclauses but excludes a number of components such as relative clauses, infinitive phrases, and nominal phrases (except when they are acting as a-verbal predicates). This, in effect, excludes from the selection all intra-sentential conjunctions such as "cats *and* dogs" and prepositional phrases such as "because of" or "in order to".

Functionally, again like most PMs, they have a procedural meaning, i.e. they encode a constraint on "all aspects of inferential processing" (Blakemore 2002: 4). What is more specific to DMs is their multifunctionality, which can be declined in three forms: (1) the category covers items that perform many different functions; (2) a single member can perform different functions in different contexts; and (3) a single member can perform different functions simultaneously in the same context, given the great polysemy of DMs. I have structured this multifunctionality into four functional "domains" (Sweetser 1990), inspired and revised mainly from González (2005), Halliday and Hasan (1976), Redeker (1990) and Sweetser (1990):

a. *ideational*: discourse relations between real-world events (e.g. cause, contrast);
b. *rhetorical*: discourse relations between epistemic and speech-act events (e.g. conclusion), and metadiscursive functions (e.g. emphasis, approximation);
c. *sequential*: structuration of discourse segments, both for local management of small units and macro-level organization (e.g. topic-shift, turn-taking);
d. *interpersonal*: interactive management of the speaker-hearer relationship (e.g. monitoring).

Any expression that, in context, performs a function in one of these four domains and that respects the syntactic filters detailed above will be considered a DM.

3.3 Inclusions and exclusions from the category

By combining functions and syntax, the above definition draws a very diverse picture of the DM category with elements such as connectives, hedges, general extenders ("and so on"), epistemic parentheticals ("I think"), etc. The complete list of all spoken English types (1563 tokens in total) extracted from the pilot corpus is provided below:[6]

[6] See section 5 for more details on the data.

> Actually, although, and, and so on, anyway, as, as you know, because, but, equally, even though, finally, first of all, firstly, for example, hence, however, I guess, I mean, I suppose, if, if you like, if you will, in fact, in other words, indeed, kind of, nevertheless, now, oh, ok, on the other hand, or, right, say, secondly, shall we say, so, sort of, still, then, therefore, though, well, what, whereas, while, yeah, yet, you know.

What it does not include on the other hand are filled pauses like "uh", interjections, response signals, tag questions and modal particles. All these exclusions are motivated by their own set of diverse criteria, and some are only valid under specific conditions: in the list above, "oh" and "yeah" are exceptions from these restrictions because, in certain contexts, they function as DMs and not as interjections or response signals, respectively. Another, crosslinguistically motivated exclusion is that of English tag questions such as "isn't it" or "don't you" which, although they might perform interpersonal functions similar to those of DMs, do not meet the syntactic criterion of fixedness (the many possible variants depend on the syntax of the utterance). Detail of these restrictions and conditions is beyond the scope of this paper.

As a result, the flexible syntactic and functional criteria, combined with explicit restrictions, strive to isolate the specificity of DMs among the other categories of PMs with which they are often confused, while accommodating

a wide range of possible forms that DMs may take. The resulting picture of pragmatic categories can be seen in Figure 1.

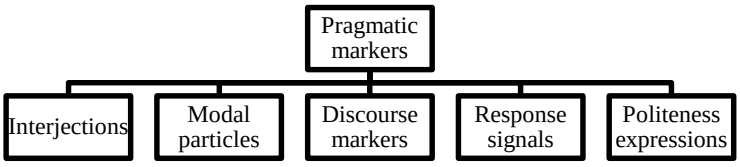

Figure 1: Members of the overarching category of pragmatic markers

The distinctions between pragmatic categories, and even between elements of a single category that may or may not be DMs depending on their context of use, do not systematically resort to the same criteria. However, all relevant criteria for these distinctions are provided by the definition. For instance, the interjection “oh”, while always syntactically optional and taking scope over an independent unit, will be discriminated on functional grounds:

(1) everybody down in the pub is saying **oh**, have you seen that new ? no, I haven’t, because I actually can’t see anymore (*Backbone* en_021 “audio-description”)

(2) all first time mums. And dads. **Oh** no no that’s not true, one of them actually has another child (*Backbone* en_014 “working mum”)

In (1), the token works at the sequential level by announcing upcoming reported speech, while in (2) it corresponds to the actual interjection, expressing the detection of an error in this case. Given the grammatical

heterogeneity of the DM category, most distinctions are functional, apart from considerations of intra- vs inter-sentential conjunctions and autonomy of the connected segment. While more subjective and less replicable than syntactic parameters (see Bolly et al. (this volume) for an assessment of similar variables in an independent annotation task), functional criteria are paramount in the process of identifying tokens of DMs in speech. This comes as no surprise since this category holds as a consistent group of elements only insofar as they share a global discourse function, and is thus a *functional category*, rather than a grammatical one.

4. From theory to practice…

To complete this rather flexible definition in a more concrete and empirical way, I suggest to combine it with a crosslinguistic annotation scheme which "translates" almost every criterion described above into its own layer of annotation. My claim is that any categorical definition is only useful insofar as it comes with a matching empirical model that helps to effectively identify and describe the members of the category.

4.1 Methodology for a crosslinguistic annotation model

Following the corpus-based methodology used for the definition, this model was elaborated from both theory and empirical testing on a French-English pilot corpus of interviews (*Backbone*, Kurt 2012). This choice of data is motivated by the intermediary level of interactivity, formality and spontaneity of spoken language in a situation of face-to-face interview, which was expected to provide examples from the whole scale of register, with both formal and more casual speech. A bilingual corpus, albeit on not-too-different languages like English and French, offers also to overcome language-specific preferences and to strive towards multilingual – if not universal – categories, values and criteria (cf. the treatment of tag questions in section 3.3).

Some variables selected in this model were borrowed and adapted from existing frameworks originally designed for either speech or writing. Additional parameters were implemented in order to complete the description of DMs' behavior and to better correspond to the definition of the category.

All variables were revised after several tests on corpus data, using the EXMARaLDA annotation software (Schmidt & Wörner 2012), as shown in Figure 2. Unclear definitions of values, implicit biases, conceptual overlapping and *ad hoc* categories were reduced to a minimum.

@@ Insert CRI1 here

Figure 1: Annotation interface of the EXMARaLDA software

For the sake of transparency and replicability, all decisions are justified and documented in a detailed coding scheme. I also recommend storing all preliminary versions of an annotation protocol, especially the modifications and their motivations. As Glynn (2010: 242) said, a coding scheme is "a crucial operationalization for quantitative semantic analysis of natural language" and therefore needs to be done carefully, and made available to the research community to avoid the multiplicity of particular, non-replicable approaches.

4.2 The model

The structure of the present model corresponds to the two groups of criteria identified in the definition, *viz.* syntax and function.

Syntactic variables include part-of-speech (POS), position and co-occurrence with another DM. The annotation of *POS* assigns a tag to the whole DM unit, and not to each component in the case of a multi-word expression. A similar approach is taken by Pitler and Nenkova (2009), who refer to this syntactic feature as "self-category": "The highest node in the tree which dominates the words in the connective but nothing else" (2009: 14). The set of POS tags used in this protocol is borrowed from the PDTB annotation guidelines in Santorini (1990) with a few adjustments.

Position of DMs is particularly challenging to annotate since they are, by definition, outside regular syntactic structures. To account for their complexity, I designed three layers of position that, once combined, provide an accurate description of their behavior in context: micro-syntax (position of the token in the smallest clause possible), macro-syntax (position of the token with respect to the root verb of the dependency structure, Lindström 2001) and turn-of-speech (position of the token within the speaker's turn, Bolly et al. (this volume)).

As for pragmatic variables, they offer three different yet related filters into the function(s) of the DM. The more specific the filter, the more informative, with more possible values and a more precise idea of what the token is performing in its context. The first filter is referred to as "*type* of DM" and addresses an issue most often left ignored in the literature: the distinction between DMs signaling a discourse relation such as cause or contrast, and DMs functioning on other semantic levels such as text-structuring, metadiscourse, interactivity. Degand and Simon-Vandenbergen (2011) were the first to propose a scale of relationality in this matter, from "non-relational" (e.g. "I think") to "strictly relational" (e.g. "because"). To operationalize this distinction as a discrete variable, the annotation scheme suggests three values: relational, non-relational and "both", which applies to tokens performing two simultaneous functions, one from each type. In this perspective, connectivity is seen as a rather inclusive notion that is "not limited to relations between

neighboring utterances" (Hansen 2006: 25) but that could include relations between assumptions or previous context.

The second filter corresponds to one of the four *domains* defined above: ideational, rhetorical, sequential and interpersonal. Type of DM and functional domain are not interdependent (i.e. a value from one layer does not systematically imply a unique value from the other layer) but are still related in the sense that, prototypically, ideational and rhetorical discourse relations (as well as some sequential functions) will be relational, while other rhetorical (metadiscursive), sequential and interpersonal functions will be non-relational.

Finally, the most informative – and most complex to code – parameter is the specific *function* of the occurrence. The present model provides a closed list of thirty functions (see Appendix 1), some of them directly borrowed from the literature (especially the PDTB and González 2005), and others which emerged from various tests on corpus data. Each function belongs to one domain, and has a prototypical type (relational or not). Based on the findings of an annotation experiment (Crible & Zufferey 2015), the list of values was operationalized with a paraphrase and prescriptive instructions on how to use the tags.[7] For instance, the potential ambiguity that arises in some under-specified contexts between relations of temporal ordering and of consequence

[7] No quantitative, statistical measure of inter-rater agreement can be provided for this study since the annotators worked with a different number of total occurrences. The results are therefore mostly qualitative, viz. revealing recurrent disagreements on weaker definitions.

(e.g. “while” and “as a consequence”, respectively) is solved and operationalized with the following instructions:

a. *Consequence*: “the connective indicates that the situation in Y is the logical effect brought about by the situation in X in the real world. Includes relations of purpose or goal. Excludes underspecified additions and temporal sequences. Paraphrased by ‘as a consequence, because of this, this happened’”;

b. *Temporal*: “the arguments are related temporally, either ordered or overlapping. Temporal bias in case of conflict with under-specified consequence. Paraphrased by ‘after/before/during this, then…’”.

All these variables and their definition are summarized in Table 1.

Syntax	POS	source grammatical class of the (head of the) DM
	Position: macro	position in the dependency structure
	Position: micro	position in the minimal clause
	Position: turn	position in the turn-of-speech
	Co-occurrence	whether the DM co-occurs with another DM
Pragmatics	Type	whether the DM is relational, non relational or both
	Domain	component of language structure affected by the marker
	Function	specific function in context

Table 1: Overview of the annotation tiers

4.3 Mapping the definition onto its annotation model

The ambition of this model is to translate the major criteria of the definition into empirically quantifiable variables. Thus, the tokens identified and annotated according to this coding scheme will in effect correspond to the description of the category provided by the definition. This mapping of definition and model concerns more specifically these four criteria: "grammatically heterogeneous", "multifunctional", "optional, relatively mobile" and "variable scope", which will be illustrated in more detail in this section.

The formal diversity of DMs is first illustrated by the POS tags, which show the full array of grammatical classes from which DMs may originate in speech. While some parsers and analysts subsume several categories under the general label "adverb" (e.g. Aijmer 1984 on "sort of", Hansen 1997 on French "donc"), in this model the original source class of the DM is preferred, to keep track of linguistic creativity and the multi-faceted form of the DM category. Hence, POS annotation is systematic ("bon" will always be coded as an adjective in French) and functionally independent, instead of mixing syntactic category and pragmatic behavior. In the following examples of "but", (3) is a typical case of introducing a contrastive relation while (4) is in final position with a more punctuating (closing) function; however they are both tagged as coordinating conjunctions:

(3) "they eventually begin to speak English **but** initially it's very hard" (*Backbone* bb_en023 "primary school")

(4) “who was a famous pirate. I don’t know much about him, **but** /” (end of turn) (*Backbone* bb_en019 “London”)

Their multifunctionality is twofold: they can perform one of 30 functions from the four domains, and they can perform two functions at once, either from the same domain or from different ones. The protocol indeed specifies the possibility to assign up to two tags simultaneously, as in (5) where “you know” both creates common ground (“monitoring” function) and introduces reported speech (“quoting” function). The number of possible values and the possibility of double tags thus reflect the multiple functions of the category (Petukhova & Bunt 2009).

(5) “they will bring to the table the actual facts, **you know**, my background is this and my husband to be’s background is that” (*Backbone* en_012 “wedding planner”)

Optionality and mobility are partially expressed through the different combinations of values from the annotation of micro- and macro-syntactic position. For instance, the conjunction “so” in medial or final (example 6) position illustrates both its optionality and mobility by taking leave of its prototypical behavior. Another example of relative mobility is the case of subordinating conjunctions such as “because” which can introduce their clause either before or after the root verb (see section 5.2).

(6) "it's a very brutal landscape a lot of it as well **so**. But my accent is very specific" (*Backbone* bb_en021 "audiodescription")

Finally, "variable scope" is covered by coding the type of DM: relational or non-relational DMs illustrate the two main kinds of possible scope, *viz.* two units (of various sizes) or one, respectively. This variation is situated at the level of the category (i.e. different members have different scopes) but also at the level of specific DMs. A number of DMs can indeed be either relational or not, depending on their function in context: "well" as turn-taking marker (non-relational) (7) or reformulation marker (relational) (8); "so" as punctuating (non-rel.) or conclusive (rel.); "but" as closing boundary (non-rel.) (cf. example 4) or contrastive marker (rel.) (cf. example 3).

(7) "are you able to say what they're involved in? /
"**well**, the ones that are here at the moment largely are…" (*Backbone* en_024 "science park")

(8) "asian speakers **well** no, asian people living in the UK want Hindi Bollywood" (*Backbone* en_021 "audiodescription")

It clearly appears that the present annotation model covers as far as possible the major aspects of the definition which it corresponds to. Such operational translation of qualitative criteria into quantifiable variables "enables verification and thus the testing of hypotheses" (Glynn 2010: 242), as we will now see.

5. … and back again: retrieving membership from annotations

One major advantage of this multi-layered model is that it keeps track of the categorical criteria of DMs in such a way that it is possible to identify different profiles of DMs. Such a flexible model asserts the unity of the DM category by applying to all its members (as presently defined), and at the same time draws a faithful portrait of their diversity, both formally and functionally. The general benefit of this approach is thus to be able to retrieve subcategories, either top-down (e.g. extracting all "relational" DMs) or bottom-up (explore corpus-driven clusters of features).

It also allows different researchers to use the same model and the same annotated dataset, by filtering out unwanted profiles that do not fit into their own framework. The following sections will illustrate the empirical and theoretical benefits of some of these filters with examples from the training corpus. This data, as mentioned before, consists of spoken interviews in French and English taken from the *Backbone* project (Kurt 2012): 27 transcripts in total, amounting to 156 minutes of recording in each language (approximately 28,000 words each), and 3,157 tokens of DMs manually identified and annotated. Scarce quantitative observations will be provided below for illustrative purposes only.

5.1 Relational and non-relational types

The most interesting example of the relational – non-relational scale is the DM "so", which illustrates the three possible values of this functional variable: relational, non-relational and both at the same time. It appears that in speech, the conjunction is very frequent (18.62% of all English DMs with 291 occurrences, the second most frequent token after "and") and can appear in contexts where the semantic meaning of consequence is weak, if present at all, and replaced by a more structuring function such as punctuating or closing. The following examples show the full range of possible scopes found in the corpus of interviews:

(9) "then you hit the tower of London which at the moment there's an outdoor skating ring. It's winter in London now **so** they pop up all over the place" (*Backbone* en_019 "London")

(10) "I deal with disputes, **so** civil disputes" (*Backbone* en_009 "lawyer")

(11) "since university, back in around two thousand one. I've been living in London, **so**, you know, eight, nine years now" (*Backbone* en_019 "London")

(12) “but there is an awful lot more work in year one. And the children and myself are both noticing that, **so** (end of turn)” (*Backbone* en_023 “primary school”)

In (9), the DM expresses its prototypical meaning of ideational consequence between two simple clauses. In our data, however, this use is only the third most frequent (40 occurrences, 13.75%) after specification as in (10) (57 occurrences, 19.59%) and conclusion, its pragmatic equivalent (118 occurrences, 40.55%). In (11), “so” expresses the rhetorical conclusion inferred from the whole previous context where the speaker says where she has been living for the past years, and simultaneously serves as punctuation to hold the floor, to stall while she is mentally calculating the number of years. In (12), on the other hand, consequence is completely absent from the meaning of “so” in this context where the conjunction is in the non-prototypical final position, thus closing the unit and turn of speech.

Researchers who want to focus on relational markers only, or more interactive functions, or on any specific function such as reformulation or consequence, can filter the annotated dataset and extract only the utterances that match their understanding of the category or their research question.

The case of “so” advocates for the inclusion of non-relational markers in the DM category, given that even originally connecting devices such as “so” can perform non-relational functions as in (12), while (11) illustrates how this

opposition is in fact a continuum of relationality, thus tending towards more flexible, inclusive boundaries of the category.

5.2 Dual position

Micro- and macro-position are especially interesting to combine in the case of subordinating conjunctions, where they give apparently contradicting information that actually accurately describe the syntagmatic mobility of these subclauses. As a reminder, the micro-level indicates the position of the DM within the minimal unit it applies to (here, the subclause), while the macro-level refers to the whole dependency structure and more precisely the root verb of the main clause. The following examples of "because" illustrate the two possible syntactic behaviors of the conjunction:

(13) "I wasn't very well, and **because** I had an emergency caesarean I physically didn't feel like myself" (*Backbone* en_014 "working mum")

(14) "in a way it's quite a good field as well **because** companies will often, when they have to make savings, intellectual property is something that they're often reluctant to cut back on" (*Backbone* bb_en016 "translator")

The governing verb in (13) "didn't feel like" is located after the subclause: the DM is thus initial both in relation to its own unit ("I had an emergency

caesarean") and to the main clause ("I physically didn't feel like myself"). However in (14), the DM depends on the previous unit "it's quite a good field" (hence end-field position) while still introducing its own (interrupted) subclause ("companies will often…").[8] A single annotation layer describing the position of DMs would fail to account for this duality of syntagmatic behavior.

In general, the macro-position reflects an opposition between integrated and non-integrated DMs, the former prototypically being relational markers and the latter mobile, free-moving elements outside the syntactic structure. As for the median position, it is usually occupied by intrusive, "disfluent" DMs, occurring where they are not expected, as in (15).

(15) "it talks about different sorts of, **well**, settings in nature really"
(*Backbone* en_023 "primary school")

Again, we see how the coding scheme provides accuracy and flexibility, in the form of complementary layers of annotation which one can either combine for a more precise description of the token's behavior, or use as filters to extract, for instance, initial or non-integrated DMs only.

5.3 Polysemous DMs

[8] In Lindström's (2001) system, syntactic position is distributed in the following slots or "fields": pre-field, initial field, middle field, end-field, post-field.

The annotation of functions in context requires deep pragmatic interpretation that needs to go beyond the basic primary meaning of a DM, however tempting it might be to always code all occurrences of an item with the same tag. “I mean” and “but” are two examples of frequent DMs that have a strong semantic core (reformulation and contrast, respectively) but are often used with a different meaning. In the corpus, they were found with four and six different functions, respectively: reformulation, specification, opening boundary and punctuation for “I mean”; contrast, concession, opposition, topic-shifting, topic-resuming and closing boundary for “but”. While all functions of “but” are relatively motivated and derived from its core adversative meaning, the uses of “I mean” are much less motivated, as is expected of this interactive, speech-specific expression. The functional spectrum of “but” is represented in Figure 3, where we see a broadening of functional scope from the core ideational meanings to more pragmatic uses in the sequential domain.

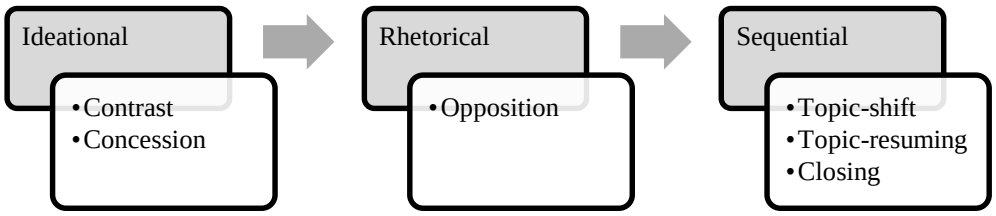

Figure 3: Functional spectrum of “but”

We observe the same scale for “I mean”, from the basic meaning of reformulation (16) to a bleached use as punctuator or turn-taking device (17).

(16) “do you find that you have quite a high turnover of businesses here, **I mean** once they've grown a little bit, do they then want to move on from the science park or do they tend to want to stay” (*Backbone* en_024 “science park”)

(17) “what are your feelings on that ?”

“**I mean**, this is very much a topic of the moment” (*Backbone* en_025 “creative writing”)

Given the great ambiguity of some of these uses and others, the annotation protocol has been improved and completed by a guide dedicated to the most frequent polysemous DMs, listing all their possible values as extracted from the pilot corpus and providing detailed criteria on how to distinguish these closely related meanings, as in Table 2 for the DM “and”.

Function	Criteria	Example
Addition	simple addition of information within the same topic	"the form of the poem reflects the substance of what you're talking about. **And** I've played around with that as well in terms of colour"
Specification	more detail, example, or particularization	"I'm interested in language itself as being multimodal **and** particularly with poetry, you've got rhythms, you've got cadence"
Consequence	logical effect brought about by the situation in S1	"that course was actually on short fiction **and** I spent quite a lot of time working on short fiction"
Temporal	chronological ordering of events	"and then I put that away in a drawer **and** I have left that since that time"

Contrast	the segments share a property which is contrasted	"you can do this in a concrete sense **and** you can do it in a slightly more implicit sense"
Topic-shift	change of topic, possible distant connection with previous topic	"I would teach my class literacy, numeracy, history, and so on. / Ok. **And** what are you doing with your class at the moment?"
Opening boundary	engage a new turn or sequence, within the same topic	"the next venture is to try to find a publisher for a small volume / **And** what's that small volume about?

Table 2: Guide to the annotation of all possible meanings of "and", with examples from *bb_en025*

5.4 Hedges: DMs or MPs?

Expressions performing hedging functions such as "sort of" or "I suppose" are often treated as a separate category (Hosman & Siltanen 2011, Liu & Fox Tree 2012). Their meaning is rooted in epistemic modality, signaling the speaker's attitude towards his/her knowledge or "the status of the proposition in terms of the speaker's commitment to it" (Palmer 1986: 54-55). Following the present inclusive approach, there seems to be no principled reason to exclude them from the category of DMs: they are optional, metadiscursive cues to interpret their context as "not reliable" or "not precise" enough.

However, epistemic modality is at the core of another pragmatic category, *viz.* modal particles. Moreover, some criteria in existing definitions would advocate for their categorization as MPs rather than DMs: they have an integrated, prototypical median position, an epistemic, metadiscursive meaning that takes scope over one unit (Cuenca 2013). To settle this

argument, we can once more rely on the parameters recommended in the present model to observe diverse authentic examples and induce what they have in common.

In the corpus, the following English types were coded as hedges: “I suppose”, “I think”, “if you will”, “sort of”, “shall we say”, “kind of”, “if you like”. We see that they are of two grammatical classes: verbal phrases or nominal constructions [NN *of*], reduced from the expression “a NN of”. Moreover, although they occur primarily in median position (94 out of 105 tokens), examples of final and initial position can also be found, as in (18) where the token applies to “having studied it…” and not what precedes.

(18) “a lot of the vocabulary that we use obviously has a Latinate base so *kind of* having studied it for three years it gave me a really good grounding” (*Backbone* en_016 “translator”)

Therefore, their relative syntactic diversity (two grammatical classes and three potential positions) would argue in favour of including them in the DM category. Modality is not the exclusive property of MPs, since DMs are also involved in attitudinal and metadiscursive functions: “both of them [DMs and MPs] may show some kind of attitude of the writer or add a nuance to what is expressed in the text” (Valdmets 2013: 127).

Arguments could be made for each categorization, and any decision would be partly arbitrary. Hedges are typical examples of problematic speech-specific elements that can or cannot be included in a category depending on how

permeable the boundaries are. For the present purposes, on the basis of their syntactic diversity and metadiscursive function, they will be considered DMs. What is more important is again to document this decision in the protocol, to motivate it with regards to the definition adopted, and to offer the possibility of filtering it out from the annotation.

6. Discussion: reliability and exhaustivity of the definition

The present proposal, while rather efficient when applied to corpus data, is not without some top-down decisions that may seem less theoretically motivated, but this cannot be avoided when dealing with such complex pragmatic categories (or any type of category). Either the definition is strict with possibly too many restrictions, or it aims for exhaustivity, with the risk of flirting with "fuzzy boundaries" (Cuenca 2013), as is the case here. Measures of agreement between different annotators thus appear necessary to assess the reliability and exhaustivity of the categorizing process. Regarding identification first, Crible & Zufferey (2015) showed that there was a substantial positive effect of training and prescriptive criteria, which led to a more operational, consensual selection of DM candidates (between 82.25% and 87.34% of relative agreement on ca. 1000-word samples). Such results support the claim of exhaustivity of the present definition, insofar as

exhaustivity is understood as inclusiveness, i.e. delimiting the boundaries between the target phenomenon and other categories. Exhaustivity as extensiveness, that is, full coverage of all potential DMs, cannot be assessed by agreement measures, if at all, given that it would require to compare the actual corpus-based selection with some pre-compiled list of reference. What can be said of this proposal is that, by applying a flexible – yet prescriptive – definition to corpus data, one selects more DM types than is usually found in other studies, all the while avoiding the inclusion of expressions from other pragmatic categories, thus furthering our knowledge of these related phenomena.

Skeptical readers may wonder whether a definition of such complex, heterogeneous phenomena is even necessary at all. I hope I have proved that this is the case, firstly for the basic reason that we, as members of a research community, need to know if we are dealing with the same category, and secondly, more importantly, given the nature of some of the members (e.g. "so") which navigate from one extreme of the category to the other. A crosslinguistic, corpus-based and paradigmatic approach such as the present one lends a certain validity to intuitive closed-lists of items, while keeping an open mind to the creativity of spontaneous speech and the emergence of new forms, on their path to be grammaticalized (e.g. Traugott 1995).

Regarding functional annotation, Bolly & Crible (2015) also concluded on the need for training even with experts in the field, although with much lower

agreement scores (K= 0.59, 60% observed agreement on 135 occurrences). If functional annotation of DMs remains a complex task (Spooren & Degand 2010), it paves the way to usage-based and contrastive research questions that need more investigation.

This analytical potential has been explored through the annotation of *DisFrEn* (Crible forthc.), a 160.000-word comparable corpus of French and English balanced across eight spoken registers (e.g. conversation, interview): 8743 DM tokens have been identified and annotated following the present annotation scheme, providing a wealth of quantitative results into the contrastive and situational distribution of DMs in various positions and functions. The major contrastive difference in this corpus lies in the high frequency of interpersonal DMs in French, a result that illustrates the crosslinguistic power of the model. French and English were also shown to differ on their combination patterns with filled and unfilled pauses, revealing contrastive usage-based prototypes (Crible et al. 2017).

Functional approaches to spoken language are only starting to emerge, given the greater challenges of this modality, and the present contribution is but a first step in this direction. Further operationalization to enhance the replicability of the functional taxonomy is particularly needed, along with intra-annotator reliability to check for consistency during the annotation process.

7. Conclusion

This article presented the design and applications of a corpus-based definition of DMs with its matching annotation model. I demonstrated how this definition, applied to bilingual corpus data, is both intensive and extensive, in the sense that it is flexible enough to cover a variety of profiles, while providing prescriptive criteria to identify tokens of DMs in natural data. I developed my claim that, in corpus pragmatics, any categorical definition is only useful insofar as it comes with an operational annotation protocol that allows a fine-grained description of the phenomena and that offers a quantifiable translation of the qualitative criteria in the definition.

This model is a first step towards a cognitive-pragmatic approach to DMs: in this growing field of linguistics, categorization models are valued for both their theoretically-sound background and their empirical validity on authentic data, with statistical and/or experimental validation. This agenda has been partly pursued in attempts to confirm the reliability of the model by applying it to the modality of gestures (Bolly et al. 2015) or sign language (Gabarró-López forthc.) and by combining it with the annotation of other surface phenomena such as disfluency markers (Crible forthc.).

References

Aijmer, Karin. 1984. “Sort of” and “kind of” in English conversation. *Studia Linguistica* 38: 118-128.

Aijmer, Karin & Simon-Vandenbergen, Anne-Marie. 2006. *Pragmatic Markers in Contrast*. Amsterdam: Elsevier.

Aijmer, Karin & Simon-Vandenbergen, Anne-Marie. 2011. Pragmatic markers. In *Discursive Pragmatics* [Handbook of Pragmatics 8], Jan Zienkowski, Jan-Ola Ostman & Jef Verschueren (eds), 223-247. Amsterdam: John Benjamins.

Asher, Nicolas. 1993. *Reference to Abstract Objects*. Kluwer: Dordrecht.

Asher, Nicolas & Lascarides, Alex. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Blakemore, Diane. 2002. *Relevance and Linguistic Meaning. The semantics and pragmatics of discourse markers*. Cambridge: Cambridge University Press.

Bolly, Catherine. 2014. Gradience and gradulaness of parentheticals. Drawing a line in the sand between phraseology and grammaticalization. *Yearbook of Phraseology* 5: 25-56.

Bolly, Catherine & Crible, Ludivine. 2015. From context to functions and back again: Disambiguating pragmatic uses of discourse markers. Paper

presented at the *International Pragmatics Association (IPrA) Conference*, July 26-31st, Antwerp, Belgium.

Bolly, Catherine & Degand, Liesbeth. 2013. Have you seen what I mean? From verbal constructions to discourse structuring markers. *Journal of Historical Pragmatics* 14(2): 210-235.

Bolly, Catherine, Gabarró-López, Silvia & Meurant, Laurence. 2015. Mapping the pragmatic functions of interactive gestures and discourse markers. Paper at the International Research Workshop CLARe, December 7-9 2015, Louvain-la-Neuve, Belgium.

Bolly, Catherine, Crible, Ludivine, Degand, Liesbeth & Uygur-Distexhe, Deniz. (this volume). Towards a Model for Discourse Marker Annotation. From potential to feature-based discourse markers.

Brinton, Laurel. 1996. *Pragmatic Markers in English. Grammaticalization and Discourse Functions*. New York: Mouton de Gruyter.

Crible, Ludivine, Degand, Liesbeth & Gilquin, Gaëtanelle. 2017. The clustering of discourse markers and filled pauses: a corpus-based French-English study of (dis)fluency. *Languages in Contrast* 17(1).

Crible, Ludivine. (forthcoming). Paradigmatic corpus annotation in *DisFrEn*: modelling discourse markers and disfluency.

Crible, Ludivine & Zufferey, Sandrine. 2015. Using a unified taxonomy to annotate discourse markers in speech and writing. In *Proceedings of the 11th*

Joint ACL-ISO Workshop on Interoperable Semantic Annotation, April 14, London, UK.

Cuenca, Maria Josep. 2013. The fuzzy boundaries between discourse marking and modal marking. In *Discourse markers and modal particles. Categorization and description* [Pragmatics and Beyond New Series 234], Liesbeth Degand, Bert Cornillie & Paola Pietrandrea (eds), 191-216. Amsterdam: John Benjamins.

Degand, Liesbeth & Evers-Vermeul, Jacqueline. 2015. Grammaticalization or pragmaticalization of discourse markers? More than a terminological issue. *Journal of Historical Pragmatics*, 16 (1): 59-85.

Degand, Liesbeth & Simon-Vandenbergen, Anne-Marie. 2011. Grammaticalization and (inter)subjectification of discourse markers. *Linguistics* 49: 287-294.

Diewald, Gabriele. 2006. Discourse particles and modal particles as grammatical elements. In *Approaches to Discourse Particles* [Studies in Pragmatics 1], Kerstin Fischer (ed.), 403-425. Amsterdam: Elsevier.

Diewald, Gabriele. 2013. Pragmaticalization (defined) as grammaticalization of discourse functions. *Linguistics* 49: 365-390.

Fischer, Kerstin. 2006. Towards an understanding of the spectrum of approaches to discourse particles: Introduction to the volume. In *Approaches to discourse particles* [Studies in Pragmatics 1], Kerstin Fischer (ed.), 1-20. Amsterdam: Elsevier.

Fraser, Bruce. 1996. Pragmatic markers. *Pragmatics* 6(2) : 167-190.

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

Furkó, Péter. 2005. The pragmatic marker – discourse marker dichotomy reconsidered. The case of "well" and "of course". PhD dissertation, University of Debrecen.

Gabarró-López, Sílvia. (forthcoming). Marqueurs du discours en langue des signes de Belgique francophone (LSFB) et langue des signes catalane (LSC) : les "balises-listes" et les "palm-ups". In *Marcadores del discurso y lingüística contrastiva en las lenguas románicas,* Oscar Loureda, Guillermo Álvarez Sellán & Martha Rudka (eds). Frankfurt : Iberoamericana/Vervuert.

Glynn, Dylan. 2010. Testing the hypothesis. Objectivity and verification in usage-based Cognitive Semantics. In *Quantitative methods in cognitive semantics: Corpus-driven approaches* [Cognitive Linguistic Research 46], Dylan Glynn & Kerstin Fischer (eds), 239-269. Berlin: De Gruyter Mouton.

González, Montserrat. 2005. Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse studies* 7(1): 53-86.

Halliday, Mark A. K. & Hasan, Ruqyiah. 1976. *Cohesion in English*. London: Longman.

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

Hansen, Maj-Britt Mosegaard. 2006. A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of French "toujours"). In *Approaches to discourse particles* [Studies in Pragmatics 1], Kerstin Fischer (ed.), 21-41. Amsterdam: Elsevier.

Hosman, Lawrence A. & Siltanen, Susan A. 2011. Hedges, tag questions, message processing, and persuasion. *Journal of Language and Social Psychology* 30(3): 341-349.

Jaszczolt, Katarzyna. 2003. On translating "what is said": Tertium comparationis in contrastive semantics and pragmatics. In *Meaning through Language Contrast*, Katarzyna Jaszczolt & Kathleen Turner (eds), 441-462. Amsterdam: John Benjamins.

Kurt, Kohn. 2012. Pedagogic corpora for content and language integrated learning. Insights from the BACKBONE Project. *The Eurocall Review* 20(2).

Lewis, Diana. 2006. Discourse markers in English: A discourse-pragmatic view. In *Approaches to Discourse Particles* [Studies in Pragmatics 1], Kerstin Fischer (ed.), 43-59. Amsterdam: Elsevier.

Lindström, Jan. 2001. Inner and outer syntax of constructions: The case of the "x och x" construction in Swedish. Paper presented at *7th International Pragmatics Association Conference*, July 9-14, Budapest, Hungary.

Liu, Kris & Fox Tree, Jean E. 2012. Hedges enhance memory but inhibit retelling. *Psychon Bull Rev* 19: 892-898. DOI 10.3758/s13423-012-0275-1.

Marcu, Daniel. 1998. A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts. In *Discourse Relations and Discourse Markers. Proceedings of the workshop, COLING-ACL '98, August 15th, Montreal, Canada*. Manfred Stede, Leo Warner & Eduard Hovy (eds), 1-7. New Brunswick, NJ: Association for Computational Linguistics.

Palmer, Franck. 1986. *Mood and Modality*. Cambridge: Cambridge University Press.

Petukhova, Vohla & Bunt, Harry. 2009. Towards a multidimensional semantics of discourse markers in spoken dialogue. *In Proceedings of the 8th International Conference on Computational Semantics*, 157–168.

Pitler, Emily & Nenkova, Ani. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP Conference Short Papers*, 13-16.

Prasad, Rashmi, Dinesh, Nikhil, Lee, Alan, Miltsakaki, Eleni, Robaldo, Livio, Joshi, Aravind & Webber, Bonnie. 2008. The Penn Discourse Treebank 2.0. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), May, Marrakech, Morocco.

Redeker, Gisela. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14(3): 367-381.

Rouchota, Villy. 1996. Discourse connectives: What do they link? *UCL Working papers in Linguistics* 8: 1-15.

Santorini, Beatrice. 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision, 2nd printing)*. Technical Report, Department of Computer and Information Science, University of Pennsylvania.

Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Schmidt, Thomas & Wörner, Kai. 2012. EXMARaLDA. In *Handbook on Corpus Phonology*, Jacques Durand, Gut Ulrike & Gjert Kristoffersen (eds), 402-419. Oxford: Oxford University Press.

Schourup, Lawrence. 1999. Discourse markers. *Lingua* 107: 227-265.

Sperber, Dan & Wilson, Deirdre. 1995. *Relevance. Communication and cognition*. Oxford: Blackwell.

Spooren, Wilbert & Degand, Liesbeth. 2010. Coding coherence relations: reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2): 241-266.

Stede, Manfred. 2014. Resolving connective ambiguity: A prerequisite for discourse parsing. In *The Pragmatics of Discourse Coherence* [Pragmatics and Beyond New Series 254], Helmut Gruber & Gisela Redeker (eds), 121-141. Amsterdam: John Benjamins.

Sweetser, Eve. 1990. *From etymology to pragmatics*. Cambridge: Cambridge University Press.

Traugott, Elizabeth Closs. 1995. The role of the development of discourse markers in a theory of grammaticalization. Paper presented at ICHL XII, Manchester, UK.

Valdmets, Anika. 2013. Modal particles, discourse markers, and adverbs with *lt*-suffix in Estonian. In *Discourse Markers and Modal Particles. Categorization and description*, Liesbeth Degand, Bert Cornillie & Paolo Pietrandrea (eds), 107-132. Amsterdam: John Benjamins.

Vincent, Diane. 1993. *Les Ponctuants de la Langue et Autres Mots du Discours*. Québec: Nuit Blanche Editeur.

Waltereit, Richard & Detges, Ulrich. 2007. Different functions, different histories. Modal particles and discourse markers from a diachronic point of view. *Catalan Journal of Linguistics* 6: 61-80.

Appendix 1: List of functions grouped by domain

Ideational	Rhetorical	Sequential	Interpersonal
cause	motivation	punctuation	monitoring
consequence	conclusion	opening boundary	face-saving
concession	opposition	closing boundary	disagreeing
contrast	specification	topic-resuming	agreeing
alternative	reformulation	topic-shifting	elliptical
condition	relevance	quoting	
temporal	emphasis	addition	
exception	comment	enumeration	
	approximation		