Ludivine Crible* and Liesbeth Degand

Reliability vs. granularity in discourse annotation: What is the trade-off?

DOI 10.1515/cllt-2016-0046

Abstract: We report on the results of an annotation experiment comparing naïve and expert coders in a sense disambiguation task consisting in the assignment of function labels to discourse markers (e.g. *well, but, I mean*) in spoken French and English using a taxonomy specifically designed for speech. Our qualitative-quantitative assessment of its reliability led us to suggest fundamental revisions of the structure of the taxonomy, striving to find a better balance between reliability and granularity. The resulting model articulates two independent levels of annotation (domains and functions) which, once combined, provide a robust tool for the analysis of discourse markers and relate them to more general functions of spoken language.

Keywords: discourse markers, annotation, taxonomy, corpus-based, inter-rater reliability

1 Introduction

Corpora of language in use have been available for a number of decades, and their potential for linguistic studies keeps on expanding to new fields of research with more and more powerful tools and methods to describe authentic data. While every step in the process of compiling and analyzing a corpus is loaded with theoretical and methodological implications, annotation, i.e. "the process of adding [...] interpretive, linguistic information to an electronic corpus" (Leech 1997: 2), remains particularly challenging and crucial to the overall validity of the study. Whether manual or (semi-)automatic and whatever the phenomenon under scrutiny, corpus annotation always involves qualitative analysis by the researcher, either in designing the coding scheme and/or applying it to the data. Annotation is particularly challenging because it faces a dilemma between two extremes of a continuum: on the one hand, annotations should be reliable, that

***Corresponding author: Ludivine Crible,** Institute for Language & Communication, Université catholique de Louvain, E-mail: ludivine.crible@uclouvain.be
Liesbeth Degand, Institute for Language & Communication, Université catholique de Louvain, E-mail: liesbeth.degand@uclouvain.be

is, optimally objective and replicable so that other researchers would be able to reproduce them; on the other, annotations should be as fine-grained as possible to allow for detailed and informative analysis. These two ideals are somewhat contradictory since subtle distinctions might not be identified equally reliably by different annotators, leading to disagreements and poor replicability. This paper tackles the how and why of finding the balance between reliability and granularity.

We specifically address the challenges of discourse annotation, which considers phenomena above the sentence level such as information structure or coherence relations. Major contributions to this domain include the annotations of coreference and argument structure in several large corpora such as the Potsdam Commentary Corpus (Stede and Neumann 2014), the Prague Dependency TreeBank (Zikánová et al. 2015) or the RST Signaling Corpus (Das et al. 2015) as well as the functional annotation of explicit and implicit connectives in the Penn Discourse TreeBank 2.0 (henceforth PDTB, Prasad et al. [2008]). According to Spooren and Degand (2010), these phenomena are particularly complex to annotate because of the intrinsic ambiguity and underdetermination of semantic interpretation in context (2010: 251, 254). The authors therefore suggest a lower threshold of acceptability for inter-annotator agreement at discourse level than for other levels of analysis such as syntax or morphology.

The present study fits in this line of research on discourse annotation by investigating the functions of so-called discourse markers (henceforth DMs, Schiffrin [1987]), which form an open-class category of pragmatic expressions generally defined by two main features: syntactic optionality (or weak clause association, Schourup [1999]) and procedural meaning (Fraser 1996). DMs can be distinguished from other discourse phenomena by their high contextual variation and multifunctionality, ranging from relational meanings (as causal or contrastive connectives) to speech-specific uses (as markers of common ground or turn-exchange management) (Fischer 2000; Maschler and Schiffrin 2015). As a result, two main challenges arise in the study of DMs: delimitation of the linguistic category of DMs as opposed to other discourse-pragmatic devices, and sense disambiguation. Although both issues have been extensively discussed in the literature (e.g. Bolly et al. 2015, in press; Cuenca 2013; Zufferey and Degand 2013), most works are focused on writing and do not take into account the specificities of speech. Even when applied to spoken data, writing-based taxonomies (e.g. the PDTB applied to spoken Italian in Tonelli et al. [2010]) overlook the additional senses that typical connectives can express in speech, such as topic-shifting *so* or turn-taking *but*. This situation can be explained by the larger availability of written data, in contrast to the fairly

recent emergence of spoken corpora which are costlier to compile and exploit. While we acknowledge the complexity and quality of discourse annotation in written corpora, we would however like to argue that spoken DMs show a wider panel of pragmatic functions and therefore require specific annotation models, rather than mere extensions of existing written language annotation models (cf. Crible and Cuenca submitted).

A number of authors have designed functional taxonomies specifically for spoken DMs (González 2005; Maschler 2009), but they often present one of two shortcomings: either poor operationality (vague definitions, few or no criteria to guide the annotator) or low granularity (coarse-grained annotation categories). In this study, we take as starting point Crible's (in press) proposal which was originally designed to address these pitfalls in spoken DM annotation. By means of an annotation experiment comparing "naïve" and "expert" coders, we offer a quantitative-qualitative assessment of Crible's proposal and suggest a revised taxonomy in light of our findings. The new scheme will in turn be tested in a second annotation experiment in order to assess its improvement on coding reliability. This study will lead us to a methodological discussion of good practices in discourse annotation with a focus on the trade-off between granularity and reliability, as well as theoretical implications of our revised approach to the functions of (spoken) language.

In this paper, we will start by defining the concepts and frameworks involved in the study (Section 2). We will then proceed to the method (Section 3) and results (Section 4) of the annotation experiment. The revised taxonomy will be presented and tested in Section 5, and critically compared to the original in Section 6, before we conclude.

2 Definition and functions of DMs in spoken language

2.1 Identifying DMs

Discourse markers form an open-class category of pragmatic expressions such as *because, well, so* or *I mean*. Despite the formal diversity and multifunctionality of the DM category and its individual members, most authors agree on a number of core defining features: syntactic optionality, procedural (non-conceptual) meaning, scope over at least one clause (or other discourse-level constituent). As for their general function, Hansen (1998) offers a cognitively-oriented definition: "markers are best seen as processing instructions intended to aid the hearer in

integrating the unit hosting the marker into a coherent mental representation of the unfolding discourse" (1998: 236). Other characteristics such as initial position, high frequency and prosodic independence, although typical of the majority of DMs, are neither necessary nor consensual (Schourup 1999). This combination of criterial and optional features helps delineate the boundaries between DMs and other pragmatic categories such as interjections or response signals, as Crible (in press) has shown.

In our view, DMs cannot be defined by syntactic or prosodic criteria alone, since the category evolves with language change and new types of DMs emerge from non-typical word classes such as verbs (e.g. *say*) or adjectives (e.g. French *bon* 'well'), in non-typical contexts (e.g. utterance-final). In addition, crosslinguistic and typological studies have shown that some (non-European) languages resort to grammatical elements such as morphemes in Turkish or verbs in Portuguese to express "discourse-marking" meanings similar to typical conjunctions or adverbials in other languages (Zeyrek et al. 2013). Therefore, we would like to argue that what is really specific to DMs is their metadiscursive function, whereby DMs can be generally described as "fulfilling structuring functions with respect to local and global content and structure of discourse" (Fischer 2000: 20). However, while most abstract definitions agree on this core structural meaning, concrete applications to authentic data reveal a lack of consensus on the range of functions that DMs are considered to fulfill, as we will now come to see.

2.2 Annotating the functions of DMs

Functional approaches to DMs (as in Pons Bordería [2006], for instance) vary greatly from one framework to another, depending on the theoretical background and type of data they are designed for. This particularism has led to a rather chaotic state-of-the-art, which can be divided in two major groups: (i) writing-based models, restricted to relational or connective functions of DMs such as cause or contrast (e.g. Rhetorical Structure Theory, henceforth RST, Mann and Thompson [1988]; the Cognitive approach to Coherence Relations, henceforth CCR, Sanders et al. [1992]; Segmented Discourse Representation Theory, henceforth SDRT, Asher and Lascarides [2003]; or the Penn Discourse TreeBank, Prasad et al. [2008]); (ii) speech-based models, including a wider range of relational and non-relational functions typical of interactive settings such as turn-taking or monitoring (e.g. González 2005; Maschler 2009).

Within these mode-specific groups, further differences are commonplace regarding the number, categorization and types of functions included in each

model. For instance, in the PDTB, sense labels are hierarchically organized from a top-level with four generic meanings ("contingency", "expansion", "temporal", "comparison") to two or three more specific levels, whereas in the CCR, the annotation does not use end-labels but rather decomposes the meaning of the DM in "cognitive primitives" which have binary values (e.g. "source of coherence" can be either "objective" or "subjective"), while RST does not identify any top-level.[1]

Endeavors to unify and standardize the functional annotation of DMs have been proposed recently by Benamara and Taboada (2015), Lapshinova-Koltunski et al. (2015), Prasad and Bunt (2015) and Sanders et al. (submitted). However, so far, these interoperable schemes either target written corpora or the relational meanings of spoken DMs, while specific (non-relational) spoken functions still lack a similar unifying approach to date. While we acknowledge that speech and writing share a common core of relational meanings which can be annotated with the same taxonomy (e.g. Crible and Zufferey 2015; Stent 2000; Tonelli et al. 2010), Crible and Cuenca (submitted) have shown that DMs in speech present a number of specific features such as under-specification or co-occurrence, which need to be taken into account during the annotation, hence calling for speech-based models. A few exhaustive proposals are available for different spoken languages (González [2005] for English and Catalan; Maschler [2009] for Hebrew), but most works only provide a categorization at a generic level without any specific corpus-based guidelines, starting from Halliday's (1970) ideational-textual-interpersonal distinction (see also Cuenca 2013; Schiffrin 1987; Sweetser 1990). Generic categorizations as well as corpus-based proposals only partially overlap and leave considerable room for improvement in the field of spoken discourse annotation.

2.3 Crible's taxonomy

Since a full review of available definitions and taxonomies is beyond the scope of this paper, we will focus on the general approach proposed in Crible (in press) where DMs are described as functioning in four "domains": ideational (relating real-world events), rhetorical (expressing the speaker's subjectivity and meta-discursive effects), sequential (structuring local and global units of discourse) and interpersonal (managing the speaker-hearer relationship). These four domains comprise thirty specific functions, as shown in Table 1.

1 In the original model, Mann and Thompson (1988) do make a general distinction between presentational vs. subject-matter relations. However, their taxonomy has been largely extended (e.g. Marcu et al. 1999) to more labels which are no longer classified according to this top-level.

Table 1: Crible's (in press) taxonomy with four domains.

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		

In this model, domains and functions are inter-dependent: a "cause" will always belong to the ideational domain, and reversely the ideational domain can only be expressed by one of the eight functions in the left column. In particular, pairs of relations that only differ in their degree of objectivity (e.g. objective cause vs. subjective cause) are grouped in different domains and given different labels (e.g. ideational "cause" vs. rhetorical "motivation", respectively) so as to be directly recognizable. This distinction is called "source of coherence" in the CCR model (Pander Maat and Sanders 2000; Sanders et al. 1992) and was repeatedly shown (e.g. Traxler et al. 1997) to be reflected in cognitive processing, which argues for its inclusion in the taxonomy as opposed to SDRT, for instance, where objective-subjective pairs are not distinguished. All functions in one domain are considered to present the same degree of granularity, i.e. the only hierarchical distinction is between domains (generic) and functions (specific). Many functions in the sequential domain and all interpersonal functions are speech-specific, and therefore only partially included (if at all) in writing-based models.

Another feature concerns the nature of the top-level categories: here, the four domains correspond to global intentions or entities, depending on whether the speaker targets content (ideational), illocutionary value (rhetorical), discourse structure (sequential) or intersubjective inferences (interpersonal). These macro-functions, which can be said to apply to other elements of language besides DMs, are related to Sweetser's (1990) "domains of use" and Redeker's (1990) components of discourse structure and can be defined as cognitive-conceptual categories. In this, Crible's (in press) taxonomy stands in sharp contrast with the PDTB taxonomy where the top-levels in the taxonomy are semantically defined, i.e. specific functions are grouped according to the similarity of their encoded meaning: the four semantic classes are "temporal",

"contingency", "comparison" and "expansion". One problem with such a semantic hierarchy, which motivated both Crible's approach in domains and Sanders et al.'s (1992) proposal of cognitive primitives in the CCR model, is that the classification of specific senses is not always coherent: for instance, in the PDTB 2.0, concessive relations are grouped in "comparison" along with contrast but are separated from causal relations (in the "contingency" class) in spite of the shared causal inference involved in both concession and cause; similarly, alternative relations are classed as "expansion" whereas their core meaning is strongly related to an operation of "comparison" of two contents. In sum, domain-labels tend towards coherent categories in line with well-established representations of discourse (cf. Chafe 1994 adapted by Martin et al. 2014; Halliday 1970) as opposed to semantic groupings with fuzzy boundaries.

2.4 Hypotheses

Crible's (in press) taxonomy aims at a fine-grained analytical grid to encompass the functions of spoken DMs. However, the reliability of this annotation model is yet to be tested, since sense disambiguation is an interpretive task which should be controlled in the perspective of rigorous corpus linguistics. In other words, we propose to situate this model on the scale between reliability and granularity by means of an annotation experiment. The first goal of this study is therefore to provide a general quantitative measure of reliability in terms of inter-rater agreement (i.e. kappa-scores), to confront the model under scrutiny to common evaluation practices. This general score will be refined according to three main lines of investigation.

The first hypothesis concerns the expertise of annotators: prior training in discourse and corpus annotation should favor agreement between trained coders, especially when dealing with abstract concepts (such as domains), as the annotation experiment by Scholman et al. (2016) has also shown: in their study, a cohort of 40 near-naïve annotators (i.e. non-trained students in Humanities) analyzed a sample of 36 coherence relations using the CCR model, which returned only 36% to 42% of correct annotations compared to the original values assigned by experts. Thus, it was demonstrated that naïve annotators are less reliable than expert coders.[2] In the present study, the comparison of expert vs. naïve annotators serves a further methodological purpose, i.e. to identify the level of granularity beyond which functional distinctions become too subtle to be reliably identified, in the

2 Naïve annotators are sometimes used to provide a large, fast and "cheap" workforce for tasks that do not require any specific training (e.g. Nowak and Rüger 2010 on image annotation).

perspective of large-scale annotation campaigns where trained experts are not always available. The "expertise" variable partially overlaps with the amount of training which the different groups followed, with the additional nuance that one of the expert annotators designed the taxonomy and thus had a longer experience with using it. We base this hypothesis on the claim that offline annotation is by nature different from online interpretation and necessarily involves some linguistic analysis, which should be more out-of-reach for naïve coders.

Secondly, we expect that annotations at function-level will be less reliable (i.e. show a lower inter-rater agreement) than at domain-level given the difference in granularity, regardless of expertise. This hypothesis is based on three features of the taxonomy: the number of possible values at each level (4 domains vs. 30 functions); the degree of precision they both involve (annotators may perceive a sequential function without being able to identify the specific function label); and the inter-dependence of the two levels, by which a function-label necessarily implies an overarching domain, so that a disagreement at domain-level will always be reflected at function-level. This hypothesis should however be qualified if we consider objective-subjective pairs (e.g. cause-motivation, cf. above) in which case annotators might agree on the underlying function but differ in its top-level classification. Satisfying reliability of annotations at a fine-grained level would support the widespread use of detailed and well-documented taxonomies by other researchers. Low reliability, on the other hand, would suggest to take more precautions when drawing empirical conclusions from the use of this taxonomy. From a more methodological point of view, low reliability can also help us identify the "grey areas" in the taxonomies, where either further operationalization is needed, or fine-grained distinctions might appear to be superfluous.

Thirdly, this study will allow us to further investigate the complexity of particular functions and particular DMs. Our hypothesis is that low reliability does not equally affect all functions in the taxonomy, and that in particular highly polysemous DMs such as *but* or *well* are more prone to disagreements because of their variation in use. By contrast, semantically stable DMs (e.g. *you know*) should be tagged consistently. Similarly, speech-specific functions such as face-saving or turn-taking can be expected to trigger more disagreements than relational functions such as cause or contrast, since the latter are more semantically encoded and usually represent the core meaning of the DMs, while the former are often pragmatic variants in specific contexts (cf. the phatic function of French *mais* 'but', Schlamberger Brezar [2012]). Problematic functions and DMs thus identified should be taken into consideration for the revision of the taxonomy, striving towards a more efficient balance between reliability and granularity.

3 Annotation experiment

3.1 Participants

Five annotators were involved in the experiment and classified in two groups based on their expertise in discourse analysis and corpus annotation. On the one hand, three nearly-naïve coders were recruited in the framework of a Graduate seminar on Discourse Analysis. They share a homogeneous profile: native speakers of French with excellent proficiency in English, linguistics majors who took the same seminar at the same time.

On the other hand, the two "expert" or nearly-expert coders are both well trained in linguistics and corpus-based discourse analysis especially focusing on DMs in very similar frameworks. However, one of the experts had designed and used the taxonomy for previous annotation projects, hence she was naturally more familiar with its definitions, coverage and distinctions. This should be kept in mind while assessing agreement within the expert group.

3.2 Materials

For this experiment, we used a transcript of conversational data in French (from the LOCAS-F corpus, Degand et al. [2014]) and English (from the ICE-GB corpus, Nelson et al. [2002]), where the first 55 tokens of DMs were annotated in each text, amounting to a sample of 110 DM tokens. DMs were pre-identified by one of the experts: identification of items was therefore not included as a step in the experiment, so as to avoid disagreement at this preliminary level. The identification of DMs was bottom-up (i.e. not based on a closed list of items) and followed a formal-functional definition including any procedural expression which is syntactically optional, formally fixed and which fulfills a metadiscursive role in context analyzable in terms of one of the four domains in the taxonomy. This definition includes expressions such as *so, because, however, well, I mean, look*, etc. and excludes other pragmatic elements such as interjections (*ah, um*), sentence adverbials (*fortunately*) or tag questions (*isn't it*) (see Crible [in press] for more details on the definition).

Annotators had access to the aligned soundtrack under the EXMARaLDA software (Schmidt and Wörner 2012), represented in Figure 1. This figure also shows that a list of all possible values is provided on the right side of the screen as a coding template, so that annotators do not have to memorize the full taxonomy and the specific tags for each value. The English and French files were annotated by all coders independently.

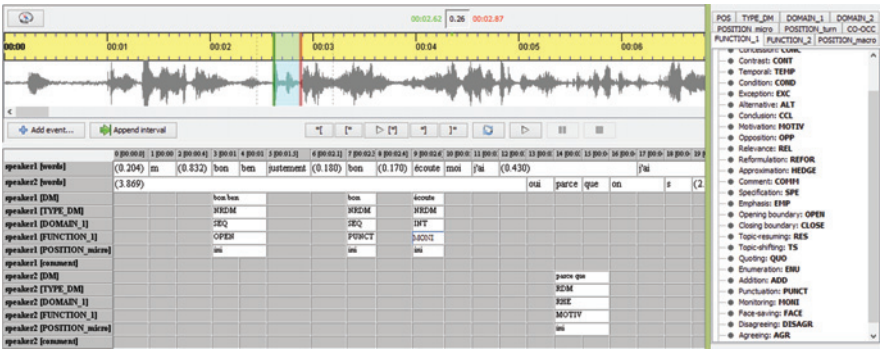

Figure 1: Annotation interface in EXMARaLDA.

Materials also included a detailed manual (Crible 2014) which provides definitions, criteria and authentic examples for all functions in the taxonomy. Based on the recommendations of a previous annotation experiment (Crible and Zufferey 2015), additional sections in the guidelines focus on highly polysemous DMs (such as *and* or *so*) and ideational-rhetorical pairs of functions (e.g. cause-motivation) which are particularly challenging to annotate (see also Scholman et al. 2016). These guidelines were browsed through collectively in the classroom in order to prevent any misunderstanding between annotators, without any further training or discussion between annotators before or during the annotation.

3.3 Instructions

All annotators were asked to perform blind annotations (i.e. without consulting other annotators at any stage of the process) on the 110 DM items pre-identified in the two samples. They had free access to the guidelines, the transcription and the audio recording without any sort of restriction, given that some DMs may require more context to be interpretable.

For each DM token, three variables were annotated on separate layers or tiers: domain, function, and position in the (sub)clause. The latter takes only four levels (initial, medial, final or independent) and is used as a control variable, given its relative simplicity compared to the functional annotations. Specific instructions for the positional variable were also available in the guidelines. Syntax is not the focus of this experiment, nor was it much developed during the seminar, but it was rather used as a formal filter that could partly explain our (dis)agreement results: non-typical positions (i.e. medial, final)

could be reserved to a limited number of possible markers and functions, hence reducing their variability and enhancing their stability across annotators.

Finally, annotators had the option to assign up to two simultaneous labels for one DM token. Double-tagging is necessary to account for the multifunctionality of DMs in some contexts, as in (1).

(1) *and there was one guy very very uhm (0.750) uh (1.670) hate to say it* ***like*** *working class*

In this example, the speaker uses a mitigating DM in order to downtone the potential depreciative connotation of the expression "working class", and is thus assigned both the "approximation" (rhetorical) and "face-saving" (interpersonal) function labels. Annotators were explicitly instructed not to use double-tagging in case of hesitations but to reserve it for true simultaneous meanings, either from the same or from two different domains.

3.4 Post-treatment

Annotations were extracted in the format of a concordancer (one DM per line) and the dataset was further transformed for the purpose of the analysis. Firstly, double tags were counted as cases of agreement when at least one of the two tags was the same. In a second step of the analysis, we simplified the dataset by reducing double-tags to single-tags since these values were too rare and skewed the statistical analyses. Additional information was semi-automatically coded in post-treatment in order to help explain the results, namely whether or not an objective-subjective pair was involved in the answers (e.g. one annotator labeled "cause" and another "motivation"), and whether or not at least one coder used a double tag.

For a more fine-grained analysis, we distinguish four levels of (dis)agreement: level 1 corresponds to full disagreement when annotators within and between the two groups (expert vs. naïve) disagree; level 2 corresponds to cases of agreement within only one of the two groups, but not between groups; level 3 includes cases of quasi-agreement between all annotators but one, regardless of their group; level 4 equals perfect agreement between all five coders.

Finally, a gold standard was decided for every item, either automatically or after discussion between the two expert annotators: for agreements of types 2 (only within experts), 3 and 4, the consensual (i.e. assigned by the majority) value was considered to be the gold standard; otherwise (types 1 and 2 within naïve coders, i. e. all the cases where the experts disagreed amongst themselves), the experts

discussed their choice until an agreement was found, which led to a considerable number of changes in the final annotations. Although the expert annotators (including the original developer of the taxonomy) were directly involved in some of these decisions, we believe that gold standards as presently computed remain useful for post-hoc analysis since they provide a reference value which will serve in the identification of problematic functions and will uncover individual differences, including within expert annotators.

4 Results

4.1 Quantitative analysis

Starting with some descriptive results, the data shows that, across the 110 annotated tokens, the five coders used the taxonomy to a relatively similar extent amounting to 28 out of the 30 possible function-labels (boiled down to 22 when looking at the final gold-standard values alone), without any significant difference or preference between annotators. However, some values were generally more frequent in the samples: in particular, the sequential domain was most frequently assigned (229/550 annotations), especially in French where sequential tags were significantly more frequent than in the English text (LL = 15.32, $p < 0.001$).[3] This result on sequential DMs is in line with Crible's (2017) large-scale annotation of a 160.000-word French-English corpus of various registers.

Turning to the quantitative evaluation of the taxonomy, inter-rater agreement analysis seems to confirm our hypothesis regarding the balance between granularity and replicability. We report in Table 2 Fleiss' (1971) kappa computed

Table 2: Inter-rater agreement scores.

Raters	Domain		Function	
	Kappa	%	Kappa	%
All	0.455	62.4	0.294	33.5
Naïve	0.475	63	0.29	33
Expert	0.563	70.9	0.406	44.5

3 Log-likelihood tests ("LL") report the significance of frequency differences in corpora of different sizes. The admitted threshold for significance at $p < 0.05$ is met when the LL value is 3.84 or above. This test was computed with the on-line calculator provided by Lancaster University, accessible at http://ucrel.lancs.ac.uk/llwizard.html

on the simplified dataset (i.e. without double tags). First, it appears clearly that annotations at a lower level of granularity (i.e. domain-level) trigger a better agreement than more specific function-labels, with a considerable decrease in kappa-scores and relative agreement for both groups of coders. The expected gap between naïve and expert annotators is also confirmed and more important at function-level, with a difference of +11 percentage points for experts. On the other hand, the relative agreement at domain-level tends to indicate that expertise is less crucial for coarse-grained annotations, with only a difference of 7% between the two groups.

Overall, these agreement scores are rather poor, below the traditional threshold of .7 recommended for discourse annotation (Spooren and Degand 2010). DM functions remain problematic to annotate even by expert coders. Compared to previous annotation campaigns, these results might not seem particularly supportive of the present functional taxonomy at first hand: in comparison with our 44.5% of agreement at function-level, Zufferey and Degand (2013) report a 66% agreement score in their experiment with expert annotators using the PDTB at a comparable level of granularity (level 2) in different languages, while Scholman et al. (2016) also present satisfying results for some of the "cognitive primitives" defined in the CCR framework (95.5% for polarity and 56.7% for source of coherence by non-expert annotators in Dutch). However, the results from the PDTB and CCR are not directly comparable to the present study for three main reasons: Crible's (in press) taxonomy (i) includes more functions (30 types against 15 in the PDTB's level 2 and binary values in CCR) thus presenting finer distinctions, (ii) covers more diverse types of functions than typical coherence relations (thus opening to non-typical markers such as particles or verb phrases) and (iii) targets spoken language which is arguably more complex or ambiguous in nature than writing (see e.g. Halliday 1987). In sum, we can draw from this quantitative analysis the (temporary) conclusion that the present model seems to favor granularity over replicability.

4.2 Qualitative analysis

To qualify these first observations, a closer investigation of the dataset reveals interesting perspectives which are relevant beyond the strict evaluation of the present taxonomy. Firstly, contrary to our expectation, good agreement is not specific to typical discourse relations but is also attested for speech-specific functions such as approximation, turn-taking or monitoring, as in (2) where all annotators agreed on the same function-label.

(2) *yeah that sounds to me like a sort of **you know** nineteenth century thing*

Secondly, again as opposed to what we expected, the objective-subjective or ideational-rhetorical distinction was not exclusively found in problematic cases (12/28 were agreements of level 3 or 4). This variable is however particularly related to the annotators' expertise, since it can explain several cases of level-2 (dis)agreement at domain-level, as in (3).

(3) <spk1> *how confusing*
<spk2> *no you don't do your hair in a barometer **so** it's not really*

In this example, the expert group agreed on a rhetorical interpretation of the DM *so* as expressing a conclusion (since the two arguments are not causally-related events but rather pertain to the epistemic domain), while the naïve annotators agreed on the ideational relation of consequence, which was probably influenced by the short length of the connected units and their closeness in the speech flow. More generally, a huge proportion of level-2 disagreements (21/23) are due to the naïve annotators at domain-level, which indicates that experts show a more consistent within-group tendency.

Another interesting result suggests that certain expressions seem to be more prone to disagreement than others. A logistic regression model testing the impact of explanatory factors (objective-subjective distinction, presence of a double tag, language, DM expression) shows that the DM itself is most predictive of annotation complexity ($C = 0.844$, $p < 0.001$). Some of the most problematic DMs (i.e. showing frequent scores of (dis)agreement levels 1 and 2) in our sample are *actually*, *but*, *well*, French *bon* ('well'), *et* ('and') or *enfin* ('I mean') as in the following example.

(4) *ils vont nous embêter toute la soirée (0.430) voire toute la nuit hein (0.180) parce que bon (0.160) **enfin** quoique ils se seront levés tôt*
'they will be annoying all evening (0.430) even all night (0.180) because well (0.170) ***enfin*** ('I mean') on the other hand they will weak up early'

Each annotator gave a different label for this example of *enfin*, which is usually a reformulative marker similar to *I mean*, here flirting with concession and mitigation. Another frequent source of disagreement is the specification label, which was almost exclusively used by one annotator. On the other hand, some functions were most frequently agreed upon and can be classified in three groups:

– functions expressed by monosemic DMs such as approximation (*sort of*), ellipsis (*and so on*, French *et tout*) or monitoring (*you know*, French *tu vois*);

- functions strongly encoded in the lexical meaning of the DM, typically coherence relations (*because, if, when*);
- very frequent functions which become familiar to the annotators, such as addition or turn-taking.

Looking at gold-standard values, we found that the rhetorical domain was frequently assigned in cases of disagreement either with the ideational or the sequential domain, which might be a sign of the complexity of this value and/or a potential bias for the rhetorical interpretation during the discussion between experts. Similarly, the most frequent gold-standard functions are often the result of the discussion after a disagreement between annotators: for instance, the eight instances where the punctuating function was assigned as gold standard all come from cases of level-1 and level-2 disagreements. Among these frequent functions, only the monitoring function shows the reversed tendency, with 6 out of 9 cases originally agreed upon.

To summarize the qualitative analysis of factors impacting the complexity of the annotation task, we ran a multiple correspondence analysis on two factors, namely syntactic position and domain (using the gold-standard annotations), to find their association with inter-rater agreement. By merging levels of (dis)-agreement in two groups (1–2 are "disagree", 3–4 are "agree"), we were able to identify two clusters of variables, as can be seen in Figure 2: disagreements

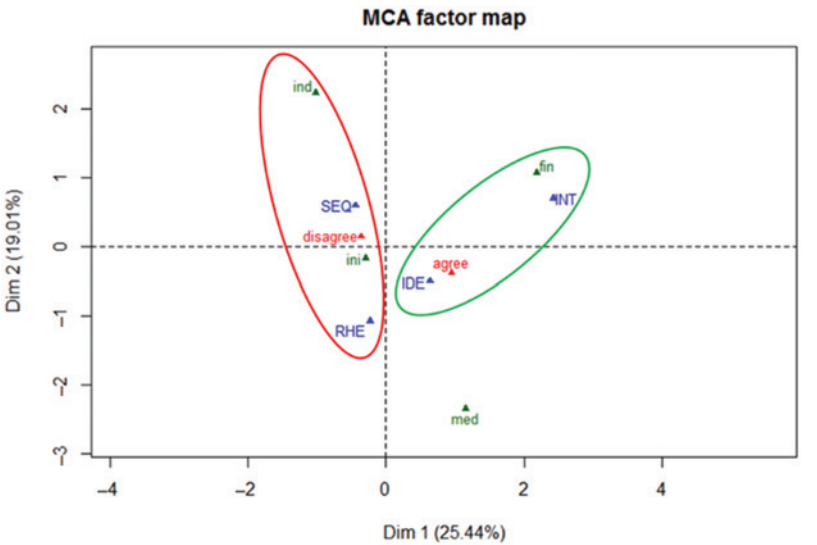

Figure 2: Multiple correspondence analysis of factors impacting annotators' agreement.

are associated with the sequential and rhetorical domains, in initial and independent positions, while agreements cluster with the ideational and interpersonal domains, in final position. This graph confirms our prior observations and tends to indicate that the most frequent values (sequential domain, initial position) are often the most complex to annotate.

Gold standard values also allow us to evaluate individual annotators by comparing their interpretations to the final decision. Tables 3 and 4 report the number of annotations which are different or similar to the gold value, by annotator, at domain- and function-level, respectively.

Table 3: Distance from gold standard per annotator at domain-level.

Annotator	Different from gold	Same as gold
exp1	22	88
exp2	22	88
naiv1	39	71
naiv2	40	70
naiv3	41	69
Total	164	386

Table 4: Distance from gold standard per annotator at function-level.

Annotator	Different from gold	Same as gold
exp1	27	83
exp2	40	70
naiv1	61	49
naiv2	70	40
naiv3	72	38
Total	270	280

We see that at domain-level, expertise comes into play by drawing a clear-cut division between expert and naïve annotators, with a sharp (1:2) difference in the ratio of "incorrect annotations" between groups, yet very similar within-group performance. The results in Table 3 would thus suggest to only involve expert annotators in (large-scale) annotation campaigns.[4] However, at function-level, Table 4 shows that the difference in performance between annotators is

4 These results on between-group performance are partially and potentially circular with the fact that some of the gold-standard values were decided by the expert annotators themselves, although this only concerns cases where they originally disagreed (see Section 3.4).

still important but more scalar and progressive: "exp1" and "exp2" show a greater difference in performance than at domain-level, and "naiv1" is somewhat intermediary between "exp2" and the other two naïve coders, who form the only coherent group in this table. Comparing individual performance to the gold-standard annotations can thus constitute a method of relative selection to identify competent annotators, possibly non-experts, for different variables at different levels of granularity. These results converge to the conclusion that discussing a gold standard is a methodological necessity, even between expert annotators, although quite costly in human and time resources.

4.3 Interim discussion

This annotation experiment has confirmed the association between high granularity and poor replicability, a conclusion which should however be nuanced by taking into consideration particular features of (spoken) DMs and their functions, as well as individual performance of annotators. In particular, we have shown that recurrent difficulties stem from the ambiguity of some frequent DMs and the lack of training of non-experts for complex notions such as the objective-subjective distinction, while other potential problems (speech-specific functions) were discarded, especially by expert annotators who show promising agreement results given the particular setting of this experiment (i.e. no joint training beforehand on a testing sample, little familiarity with the taxonomy for one of the experts and no discussion prior to or during the experiment).

Overall, functional annotation remains challenging, and this study has identified a number of problematic areas which could be the focus of extra training and more detailed annotation guidelines. Another outcome of this study is the suggestion that an intermediary level of granularity between domains and functions should enhance the replicability of the taxonomy, while maintaining a certain degree of informativeness. In this perspective, Geertzen and Bunt (2006) have proposed a weighted method to measure inter-rater agreement by taking into account the relative semantic distance between values in a taxonomy, thus being more tolerant with non-expert annotators and their "almost-correct" analysis. Their approach seems satisfying for hierarchical taxonomies that are already structured as types and subtypes of conceptually-close functions (as in the PDTB), which is not the case of Crible's (in press) model. In her broad taxonomy, the inter-dependence of domains and functions leaves little room for conceptual groupings, although a cross-domain re-organization of the values is possible and highly relevant to the search of a balance between granularity and reliability. In the next section, we put forward such a proposal.

5 Towards independent levels of annotation

5.1 Principles and structure of the revised taxonomy

Based on the quantitative and qualitative insights from the previous annotation experiment, we have revised Crible's (in press) taxonomy following a double agenda: to reduce the number of function-labels to a more manageable set and to keep the two annotation levels independent, supposing that both these principles should improve the replicability of the procedure. The main assumption of the revised taxonomy is the following: in principle, the four domains from the original taxonomy can apply to all functions; in other words, any function-label can be expressed in the ideational, rhetorical, sequential or interpersonal domain, depending on the context of use. This proposal brings functional domains to the forefront by extending their coverage, thus linking them to well-established general models of the functions of language (Halliday 1970) and planes of discourse (Schiffrin 1987; Sweetser 1990), reflecting different broad functions language in use can fulfill. In this view, linguistic categories are not binary but rather three- or four-fold, and boundaries between these categories are constantly submitted to language change, in particular through the process of inter-subjectification (Traugott 1982): in terms of the present taxonomy, DM functions tend to evolve from the ideational to the interpersonal domain.

This new perspective on the relationship between domains and functions stems from the observation that, in the original taxonomy, some functions form objective-subjective pairs, which we have identified to be problematic for non-expert annotators. Typically, this porous distinction concerns coherence relations such as cause (vs. motivation), consequence (vs. conclusion), condition (vs. relevance), contrast and concession (vs. opposition). In our view, a similar pairing can be applied to other functions from different domains, such as the alternative relation (ideational) and reformulation (rhetorical), temporal (ideational) and enumeration (sequential), or punctuation (sequential) and monitoring (interpersonal). To go even further, we put forward the hypothesis that all (or most) functions in the taxonomy, once grouped in coherent categories, can theoretically perform in each of the four domains. The resulting structure of the revised taxonomy can be found in Table 5.

Table 5: Revised taxonomy with cross-domain functions.

Ideational	Rhetorical	Sequential	Interpersonal
[addition] [alternative] [cause] [consequence] [contrast] [resuming] [temporal]	[closing] [enumeration] [topic-shift]	[concession] [opening] [specification]	[condition] [punctuation]

We can see that only fifteen function types remain from the thirty original labels after merging objective-subjective pairs (e.g. [condition] includes both condition and relevance; [cause] involves both cause and motivation), which follows the recommendation in Enschot et al. (submitted). Further reduction of the functions would be possible by merging conceptually close relations (e.g. contrast and concession), or reduce fine-grained distinctions to a more encompassing function (e.g. grouping the sequential functions [topic-shift], [opening], [closing] and [resuming] under one general [structuring] function). We believe that this new system can resolve most annotation problems at function-level by reducing hesitations and choices, all the while making relevant connections between functions that were not possible or explicit in the original model: for instance, additive DMs were originally annotated as sequential, whereas our revision implements the possibility of classifying additive relations in other domains as well. In fact, in a pilot study (Crible et al. 2016), we did find some occurrences of the same [contrast] function in different domains, as in examples (5)–(8).

(5) *nous sommes animés par le désir de participer à notre échelle au progrès de la connaissance **mais** nos liens avec l'université sont aussi fragiles*
'we are moved by the desire to participate at our own scale to the progress of knowledge ***mais*** ('but') our links with the university are fragile too'

(6) *parce que je vois encore de la poésie en cinquième ce qui peut paraître classique **mais** enfin c'est comme ça que je voulais subdiviser le le cours*
'because I do poetry again in fifth year which can seem classic ***mais*** ('but') well I wanted to divide the classlike that'

(7) <spk1> *euh j'aime les néologismes j'aime les les régionalismes mais euh je mets le point d'exlamation dessus euh pour dire euh attention*

<spk2> ***mais*** *la norme qu'est-ce qu'est-elle pour vous*

<spk1> 'uh I like neologisms I like regionalisms but uh I write an exclamation mark on them uh to say uh careful'

<spk2> ***mais*** ('but') 'the norm what is it to you'

(8) *Alors cet auditeur vigilant il va vous dire tiens euh encore Jean d'Ormesson **mais** on entend Jean d'Ormesson à chaque automne*
'well this careful listener he will tell you look uh Jean d'Ormesson again ***mais*** ('but') we hear Jean d'Ormesson every fall'

In (5), the connective "mais" expresses a contrast relation in the real world, where the fact that researchers have temporary positions in their universities contrasts with their desire to invest in teaching tasks. In (6), the contrast relation is rhetorical in that a teacher concedes that her choice to teach poetry in the fifth year might sound classic, but that this was an explicit choice. In (7), the contrast relation is situated in the sequential domain because the interviewer uses "mais" to introduce a new question, thus reorienting the topic of conversation. Finally, in (8), the contrast relation is put in the mouth of an external speaker (here the radio listener) expressing their surprise, thus expressing a relation in the interpersonal domain. So far, our data only attests the occurrence of the [punctuation] and [contrast] functions in all four domains, but further investigation of corpus data should uncover more evidence.

With this revised taxonomy, annotators can choose to start at domain-level or function-level, to annotate both levels simultaneously or independently, and could even decide to stop at one level if a particular DM token is under-specified for the other level. For instance, the difficult case of French *enfin* which we illustrated in (4) above could be assigned only the rhetorical domain in case the annotator finds it irrelevant or uncertain to be more specific. This approach is more flexible than inter-dependent or hierarchical taxonomies such as the PDTB where, in case of hesitation, the annotators can only choose to stop at a more coarse-grained level. We believe that this system with independent levels of annotation can substantially improve inter-annotator agreement provided the decision-making process is well documented and combined with adequate training and discussion (cf. Enschot et al. submitted).

5.2 Reliability of the revised taxonomy

To test whether the independence of annotation levels in the revised taxonomy improves the reliability of the method, we conducted a second annotation experiment on a sample of 83 DM tokens, previously identified in prepared monologues in French. Two linguists (only one of them is a DM expert and is familiar with the original taxonomy) independently coded the DMs at domain- and function-level, applying the list of values in Table 5 and following operational guidelines providing more ample definitions of the domains. For instance, the ideational domain is described as: *a priori* relational; it is possible to identify generally explicit S1 and S2 segments, which are adjacent by default; there is a semantic coherence relation between the segments, etc. Computing the kappa-scores returns the following agreement results: 0.78 Fleiss' kappa at domain-level and 0.66 at function-level, which is a considerable improvement of the

scores reported in Table 2 and tends to support our hypothesis of a higher reliability of this revised taxonomy with independent annotation levels.

Zooming in on the disagreements, it appears that the rhetorical domain is the most problematic one to code in the revised taxonomy, with 65 % of agreeing values against 80 % for the other three domains. This result only partially corroborates the analysis of the first annotation experiment, where both the rhetorical and the sequential domains were statistically associated with disagreements (cf. Figure 2). At function-level, the [addition] value strikes as particularly challenging to disambiguate, with only 53 % of relative agreement, most cases of disagreement concerning the [specification] and [consequence] functions. The confusion arises from the underspecification of additive DMs, typically *and* or French *et*, which are highly multifunctional and compatible with more specific pragmatic interpretations beyond mere addition, such as adding for expressing a result (9).

(9) *un projet similaire se dessine // dans le domaine de l'économie /// avec la création prochaine // d'une Louvain School of Economics /// ces premiers pas sont prometteurs /// **et** annoncent // d'autres initiatives /// tant en matière d'enseignement /// que de recherche* (LOCAS-F corpus, aca_1)
'a similar project is developed in the economics domain with the upcoming creation of a Louvain School of Economics these first steps are promising ***et*** ('and') foster other initiatives both in teaching and research'

This example illustrates a case of disagreement on *et* 'and' which was coded as ideational [addition] by annotator1 and rhetorical [consequence] by annotator2, and later resolved as ideational [consequence] after discussion: the "other initiatives" are not merely an independent fact but also the direct result of the "promising steps". We see that independent annotation levels also allow us to refine the solutions to disagreements by keeping the domain and function from two distinct annotators for a single DM token (here, the domain assigned by annotator1 and the function by annotator2).

Turning to unproblematic categories, out of the 16 occurrences of [contrast] in the sample, only two were disagreed upon by the annotators, an encouraging result which was not expected from the first annotation experiment, especially not in light of the frequent confusion between contrast and concession, as already reported in Zufferey and Degand (2013) and illustrated in (10).

(10) *et la fonction du phonostyle // c'est une fonction euh extra euh phonologique si on veut // puisque elle est destinée à montrer // ou à faire exister une communauté d'appartenance j'ai dit tout à l'heure // une profession // ou*

des choses comme ça /// euh ***mais pourtant*** */// dès qu'on a un événement de communication on a un style de parole // et c'est ce style de parole qu'on a essayé de cerner* (LOCAS-F corpus, cnf_1)
'and the function of the phonostyle it's uh an extra uh phonological function if you will because it is meant to show or to create a community I said earlier a profession or things like that uh ***mais pourtant*** ('but still') as soon as you have a communicative event you have a speech style'

In this respect, *mais* 'but' contrasts with *et* 'and' in that it is strongly associated with a contrastive meaning (100 % of occurrences in the sample), leaving only to the annotators' decision to which domain the [contrast] applies. For DMs such as *mais* with a strongly encoded function value and a variable domain (cf. examples (5)–(8) above), independent annotation levels prove once more particularly useful and more reliable.

6 Discussion

Far from solving every problem in spoken discourse annotation, the revised taxonomy still presents some benefits (as well as drawbacks) which we will now systematically compare to the original model. First, in terms of methodological soundness and overall reliability, it appears from the second annotation experiment that our revisions to Crible's (in press) taxonomy considerably restrict the risks of inter-annotator disagreements by reducing the total number of possible options (from 30 to 15 function-labels) and by keeping the two annotation levels truly independent, whereas, in the original model, a disagreement at function-level very often leads to a disagreement at domain-level.[5] This last feature also offers new possibilities for quantitative analysis of corpus data, since the researcher can now filter their annotations not only at domain-level (e. g. extracting all sequential DMs to see the distribution of functions within this domain) but also at function-level (e.g. extracting all [addition] DMs to see the distribution of domains within this function), which is especially relevant for objective-subjective pairs such as cause-motivation, or to uncover highly polysemous DMs.

5 If one aims at agreement on both domain and function simultaneously, then the actual number of possible combinations is, in principle, 15*4 (60), that is each function potentially expressed in each domain. However, agreement at each separate level (domain and function) should be improved.

Regarding granularity, the original model still remains preferable for two main reasons. First, in practical terms, inter-dependent annotations are more economical, i.e. the function-level is sufficient to deduce the corresponding domain, and labels are thus maximally informative, as opposed to the revised model where the analyst necessarily has to combine two labels in order to identify the specific value of a particular DM. Extracting annotations at a single independent level loses precision: a [contrast] can cover very different uses (see examples (5)–(8) above), and each domain can in principle include all functions in the taxonomy. Therefore, the combination of labels in the revised version is necessary during the phases of both annotation and analysis. Secondly, some relevant distinctions might be lost in the new taxonomy. We see in particular that none of the original functions from the interpersonal domain appear in the revised version, following the hypothesis that these functions are variants of labels from other domains: for instance, *monitoring* is considered to be an interpersonal use of the *punctuation* function from the sequential domain. While this position may be valid for a number of cases, it certainly loses track of the particular intersubjective meaning of expressions like *you know* as opposed to more generic "punctuating" DMs such as *well* or French *bon*. What is at stake here is what Enschot et al. (submitted) call the "external validity" of the data, i.e. "the way in which observations can be generalized to other situations outside of the specific data investigated": annotation labels should remain conceptually and cognitively accurate with respect to the data they apply to, and not be the mere results of a particular coding strategy. In our case, equating *monitoring* with [punctuation] loses some consistency with external, alternative observations of that function and the DMs which express it, but gains in return higher "internal validity", i.e. annotation reliability. Revisions like these therefore call for the necessity to combine functional annotation with some lexicographic information on the core meaning of the DM itself.

At this stage of the comparison, it would seem that the balance between granularity and reliability doesn't strongly favor one model over the other, since each version shows strengths and weaknesses at either pole of the scale. However, with the revised model, the picture is not as simple as "less informative yet more reliable", and it could be argued that the combination of functions and domains as independent layers can be even more informative than the original structure of the taxonomy. The case of [addition], for instance, reveals a more fine-grained range of uses for this relation, depending on the types of objects or events that are connected: real-world facts (ideational), arguments (rhetorical) or steps of the discourse (sequential). Integrating independent annotation levels is also highly beneficial in terms of analytical potential, since it allows to converge different types of information which gives more explanatory

power to the (statistical) modelling of variables, and may in turn validate theoretical categories by showing their associations to formal features (see e. g. Crible [2016] for an integration of DM functions and disfluency markers). Finally, as we said before, this combined approach of functions and domains may be more cognitively grounded since it echoes more general models of the functions and processes of language variation.

7 Conclusion

This study has tested the replicability of Crible's (in press) functional taxonomy for spoken DMs by means of an annotation experiment comparing expert and naïve coders. The quantitative and qualitative analysis of inter-annotator (dis)-agreements provided the grounds for a number of revisions which were developed and compared to the original model. In a second annotation experiment, we tested the replicability of the revised taxonomy and found that the independence of annotation levels and the smaller number of labels substantially improved inter-annotator agreement. While both versions show benefits and drawbacks for different aspects of the taxonomy, namely granularity and reliability, the revised model that we presented in this paper arguably outweighs the original by its flexible structure with two independent levels of annotation and analysis. We have shown how the revisions not only vouch for better replicability and agreement between annotators, but also offer a more powerful analytical tool to explore spoken discourse by connecting DMs to more general language mechanisms. As far as the trade-off between granularity and reliability is concerned, we recommend that the research question and the purposes of the linguistic analysis are not lost out of sight. Reliability and granularity are not a goal in themselves, the linguistic analysis and its explanatory power are. This being said, any linguistic analysis has to aim for (external) validity (Enschot et al. submitted). Thus, when the analysis involves subtle distinctions based on the analyst's interpretations, granularity is more important and will come at the cost of thorough discussions between analysts. If the research question involves quantitatively important data, reliability should probably receive priority, probably at the cost of granularity.

Revising annotation guidelines in order to enhance reliability is a common practice in corpus-based linguistics, and in this perspective the present paper is methodologically close to the study in Zufferey and Degand (2013): in their annotation experiments, the authors tested whether their revision of the PDTB taxonomy improved the replicability of the procedure while remaining fine-grained enough to apply to multilingual data. Their use of parallel corpora and translation

spotting techniques to identify crosslinguistic equivalents of DMs and their functions is highly interesting, although not possible for spoken data. In our approach, the suggested revisions go even further than this trade-off between granularity and reliability by proposing a new structure altogether following a strong assumption, namely that discourse domains are pervasive in all functions and correspond to general models of language use and language change. Our revision is therefore not only methodological, but also fundamentally theoretical.

All in all, comparing annotation models and frameworks is only relevant insofar as the research questions and motivations behind them are comparable. Different discourse models are designed for different research objectives. For instance, the CCR approach is theory-driven and aims at cognitive realism; the PDTB is data-driven and rather aims at practical feasibility; the SDRT is application-driven and provides a tool for full-text representation. The present approach in functional domains can be characterized as "analysis-driven" since it targets high informativeness and analytical potential in the specific field of spoken discourse analysis. Although not irreconcilable, these different agendas result in very different yet complementary models which would gain in being mapped and combined. In sum, annotation models, like corpora, are always built for a specific purpose, and there is no use in searching for one absolute, multi-purpose model that would fit any research question. That being said, we hope to have shown that our new proposal brings us closer to a satisfying balance between granularity and reliability.

A further result of importance is the observation that experts annotate systematically more reliably than naïve coders. While this might seem an obvious conclusion, we would like to discuss this result with respect to the nature of the annotation process. Namely, annotation is an off-line process that is different from the on-line interpretation process speakers and hearers engage in in daily language use. In other words, annotation is an instance of complex linguistic analysis requiring deep insight into the linguistic phenomenon under scrutiny and detailed knowledge of the coding scheme itself. Acquiring such knowledge requires intensive training of the coders, as advocated by many analysts (see e.g. Neuendorf 2002) to ensure that the annotation task is the same for all coders. Enschot et al. (submitted) point out the risk involved in intensive training, namely that annotators "code the studied phenomenon the same way because they learned to do this, which results in a high internal validity", but not necessarily in external validity. To avoid such a "learning effect", we believe further meta-analysis of the annotation process is needed. This would require that coders explicitly document the cues and procedures put to use when analyzing their data during the annotation process, as opposed to the implicitness of the usual "black-box" approach. Such a procedure is of

course time-consuming, but we believe that it will provide important information to further enrich our theoretical and empirical models of language use.

This paper paves the way for a number of research avenues, both in the area of DM research and annotation practices. First, we would like to encourage further research on the relational nature or the scope of DMs: while most writing-based definitions state that DMs (or connectives) relate two (usually adjacent) abstract objects, Crible and Cuenca (submitted) have shown that this is not always the case in speech where segments can be further apart in the discourse, interrupted or even implicit. Degand and Simon-Vandenbergen (2011) had already proposed the notion of a relational scale, from purely relational expressions like conjunctions, to non-relational DMs such as epistemic parentheticals. We believe that this notion is promising and would provide an additional independent level beyond functions and domains. Other perspectives include a large-scale evaluation of the revised taxonomy (with more annotators, more data and in more languages), as well as a proposal for a (speech-based) "common-core" taxonomy for new annotation campaigns that could apply to both spoken and written data. So far, the results of the present study can already suggest the following recommendations: allow the possibility of varying degrees of granularity; aim at informative labels with high analytical potential; integrate several annotation levels and variables to strengthen methodological and cognitive reliability. In other words, we believe that the two ends of the granularity-reliability scale can and should be combined for optimal annotation and analysis of spoken discourse.

Funding: Fédération Wallonie-Bruxelles, (Grant/Award Number: '12/17-044').

References

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Benamara, Farah & Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics (*SEM), collocated with NAACL*, Denver, USA.

Bolly, Catherine, Ludivine Crible, Liesbeth Degand & Deniz Uygur-Distexhe. 2015. *MDMA. Identification et annotation des marqueurs discursifs "potentiels" en contexte* [MDMA. Identification and annotation of "potential" discourse markers in context]. *Discours* 15.

Bolly, Catherine, Ludine Crible, Liesbeth Degand & Deniz Uygur-Distexhe. In press. Towards a model for discourse marker annotation in spoken French: From potential to feature-based discourse markers. In Chiara Fedriani & Andrea Sansó (eds.), *Discourse markers, pragmatics markers and modal particles: New perspectives*. Amsterdam: John Benjamins.

Chafe, Wallace. 1994. *Discourse, consciousness, and time*. Chicago: University of Chicago Press.
Crible, Ludivine. 2014. Identifying and describing discourse markers in spoken corpora. Annotation protocol v.8. *Technical report*, Université catholique de Louvain.
Crible, Ludivine. 2016. Discourse markers and disfluencies: Integrating functional and formal annotations. In Harry Bunt (ed.), *Proceedings of the 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-12), LREC 2016 Workshop*, 38–45.
Crible, Ludivine. 2017. *Discourse markers and (dis)fluency across registers: A contrastive usage-based study in English and French*. Louvain-la-Neuve: Université catholique de Louvain dissertation.
Crible, Ludivine. In press. Towards an operational category of discourse markers: A definition and its model. In Chiara Fedriani & Andrea Sansó (eds.), *Discourse markers, pragmatics markers and modal particles: New perspectives*. Amsterdam: John Benjamins.
Crible, Ludivine & Maria Josep Cuenca. Submitted. Discourse markers in speech: Distinctive features and corpus annotation.
Crible, Ludivine, Liesbeth Degand & Anne-Catherine Simon. 2016. Interdependence of annotation levels in a functional taxonomy for discourse markers in spoken corpora. In Liesbeth Degand, Csilla Dér, Péter Furkó & Bonnie Webber (eds.), *Proceedings of the TextLink Second Action Conference*, 36–39. Budapest: L'Harmattan Kiado.
Crible, Ludivine & Sandrine Zufferey. 2015. Using a unified taxonomy to annotate discourse markers in speech and writing. In Harry Bunt (ed.), *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11), IWCS 2015 Workshop*, 14–22.
Cuenca, Maria Josep. 2013. The fuzzy boundaries between discourse marking and modal marking. In Liesbeth Degand, Bonnie Cornillie & Paola Pietrandrea (eds.), *Discourse markers and modal particles. Categorization and description*, 191–216. Amsterdam: John Benjamins.
Das, Debopam, Maite Taboada & Paul McFetridge. 2015. *RST signalling corpus*. LDC2015T10. Distributed through the Linguistic Data Consortium.
Degand, Liesbeth, Laurence Martin & Anne-Catherine Simon. 2014. Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté [Basic discourse units and their left periphery in LOCAS-F, an annotated multigenre spoken corpus]. In *CMLF 2014 – 4ème Congrès Mondial de Linguistique Française 2014*. Berlin, Germany: EDP Sciences.
Degand, Liesbeth & Anne-Marie Simon-Vandenbergen. 2011. Grammaticalization and (inter)-subjectification of discourse markers. *Linguistics* 49. 287–294.
Enschot, Renske van, Wilbert Spooren, Antal van den Bosch, Christian Burgers, Liesbeth Degand, Jacqueline Evers-Vermeul, Florian Kunneman, Christine Liebrecht, Yvette Linders & Alfons Maes. Submitted. Taming our wild data: on intercoder reliability in discourse research.
Fischer, Kerstin. 2000. *From cognitive semantics to lexical pragmatics: The functional polysemy of discourse particles*. Berlin: Walter de Gruyter.
Fleiss, Joseph. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5). 378–382.
Fraser, Bruce. 1996. Pragmatic markers. *Pragmatics* 6(2). 167–190.
Geertzen, Jeroen & Harry Bunt. 2006. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, 126–133.
González, Montserrat. 2005. Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies* 7(1). 53–86.

Halliday, Michael. 1970. Functional diversity in language as seen from a consideration of modality and mood in English. *Foundations of Language: International Journal of Language and Philosophy* 6. 322–361.

Halliday, Michael. 1987. Spoken and written modes of meaning. In Rosalind Horowitz & S. Jay Samuels (eds.), *Comprehending oral and written language*, 55–82. New York, NY: Academic Press.

Hansen, Maj-Britt Mosegaard. 1998. *The function of discourse particles. A study with special reference to spoken standard French*. Amsterdam & Philadelphia: John Benjamins.

Lapshinova-Koltunski, Ekaterina, Anna Nedoluzhko & Kerstin Kunz. 2015. Across languages and genres: Creating a universal annotation scheme for textual relation. In *Proceedings of LAW IX at NAACL HLT 2015*, Denver, USA.

Leech, Geoffrey. 1997. Introducing corpus annotation. In Roger Garside, Geoffrey Leech & Tony McEnery (eds.), *Corpus Annotation*, 1–18. London, Longman.

Mann, William & Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3). 243–281.

Marcu, Daniel, Estibaliz Amorrortu & Magdalena Romera. 1999. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*, 48–57.

Martin, Laurence, Liesbeth Degand & Anne-Catherine. Simon 2014. Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté [Form and function of the left periphery in a multigenre annotated spoken corpus]. *Corpus* 13. 243–265.

Maschler, Yael. 2009. *Metalanguage in interaction. Hebrew discourse markers*. Amsterdam: John Benjamins.

Maschler, Yael & Deborah Schiffrin. 2015. Discourse markers: Language, meaning, and context. In Deborah Tannen, Heidi E. Hamilton & Deborah Schiffrin (eds.), *The handbook of discourse analysis*, 2nd edn., 189–221. Chichester, UK: John Wiley & Sons, Ltd.

Nelson, Gerald, Sean Wallis & Bas Aarts. 2002. *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam: John Benjamins.

Neuendorf, Kimberly. 2002. *The content analysis guidebook*. Thousand Oaks, CA: Sage Publications.

Nowak, Stefanie & Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, Philadelphia, USA.

Pander Maat, Henk & Ted Sanders. 2000. Domains of use and subjectivity. On the distribution of three Dutch causal connectives. In Bernd Kortmann & Elizabeth Couper-Kuhlen (eds.), *Cause, condition, concession and contrast: Cognitive and discourse perspectives*, 57–82. Berlin: Mouton de Gruyter.

Pons Bordería, Salvador. 2006. A functional approach to the study of discourse markers. In Kerstin Fischer (ed.), *Approaches to discourse particles*, 77–100. Amsterdam: Elsevier.

Prasad, Rashmi & Harry Bunt. 2015. Semantic relations in discourse: The current state of ISO 24617-8. In Harry Bunt (ed.), *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11), IWCS Workshop 2015*, 80–92.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi & Bonnie Webber. 2008. The penn discourse TreeBank 2.0. In *Proceedings of the 6th*

International Conference on Language Resources and Evaluation (LREC 08), Marrackech, Morroco, 2961–2968.
Redeker, Gisela. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14. 367–381.
Sanders, Ted, Vera Demberg, Jacqueline Evers-Vermeul, Jet Hoek, Merel Scholman & Sandrine Zufferey. Submitted. Unifying dimensions in discourse relations: How various annotation frameworks are related.
Sanders, Ted, Wilbert Spooren & Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15. 1–35.
Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.
Schlamberger Brezar, Mojca. 2012. Les marqueurs discursifs "mais" et "alors" en tant qu'indicateurs du degré d'oralité dans les discours officiels, les débats télévisés et les dialogues littéraires. *Linguistica* 52(1). 225–236.
Schmidt, Thomas & Kai Wörner. 2012. EXMARaLDA. In Jacques Durand, Gut Ulrike & Gjert Kristoffersen (eds.), *Handbook on corpus phonology*, 402–419. Oxford: Oxford University Press.
Scholman, Merel, Jacqueline Evers-Vermeul & Ted. Sanders 2016. A step-wise approach to discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse* 7(2). 1–28.
Schourup, Lawrence. 1999. Discourse markers. *Lingua* 107. 227–265.
Spooren, Wilbert & Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2). 241–266.
Stede, Manfred & Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 14)*, 925–929. Reykjavik, Iceland.
Stent, Amanda. 2000. Rhetorical structure in dialog. In *Proceedings of the 2nd International Natural Language Generation Conference (INLG'2000)*.
Sweetser, Eve. 1990. *From etymology to pragmatics*. Cambridge: Cambridge University Press.
Tonelli, Sara, Giuseppe Riccardi, Rashmi Prasad & Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 10)*, 2084–2090. Valletta, Malta.
Traugott, Elizabeth. 1982. From propositional to textual and expressive meanings: Some semantic-pragmatic aspects of grammaticalization. In Winfred P. Lehmann & Yakov Malkiel (eds.), *Perspectives on historical linguistics*, 245–271. Amsterdam & Philadelphia: Benjamins.
Traxler, Matthew, Michael Bybee & Martin Pickering. 1997. Influence of connectives on language comprehension: Eye-tracking evidence for incremental interpretation. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* 50A(3). 481–497.
Zeyrek, Deniz, Isin Demirşahin, Ayisigi Sevdik Çalli & Ruket Çakici. 2013. Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue & Discourse* 4(2). 174–184.
Zikánová, Šarka, Eva Hajičová, Barbora Hladká, Pavlina Jínová, Jiri Mírovský, Anna Nedoluzhko, Lucie Poláková, Katerina Rysová, Magdalena Rysová & Jan Václ. 2015. *Discourse and coherence. From the sentence structure to relations in text*. Prague: Institute of Formal and Applied Linguistics.
Zufferey, Sandrine & Liesbeth Degand. 2013. Representing the meaning of discourse connectives for multilingual purposes. *Corpus Linguistics and Linguistic Theory* 10. 1–24.