Using a unified taxonomy to annotate discourse markers in speech and writing

CRIBLE Ludivine
Catholic University of Louvain (UCL)
ludivine.crible@uclouvain.be

ZUFFEREY Sandrine
University of Fribourg
sandrine.zufferey@unifr.ch

Abstract

We report an annotation experiment aiming at assessing the use of a single functional taxonomy of sense relations for discourse markers in spoken and written data. We start by presenting an operational definition of the category of DMs and its application to identify tokens of DMs in corpora. We then present an original annotation experiment making use of a unified taxonomy to annotate written and spoken data in English and French. In this experiment, we test the reliability of the annotations made separately by two annotators and the applicability of the tag set across two languages in the spoken and written modes. Our experiment leads us to conclude that: i) spoken data is not more difficult to annotate than written data in terms of inter-annotator agreement, ii) recurrent problems are found across the two languages and modes, iii) the reliability of the annotation scheme is improved by the use of more explicit instructions and training.

1 Introduction

Discourse markers (hereafter DMs) form a functional category of lexical items including both connecting devices signaling a discourse relation (e.g. *but, or, so*) and non-relational interactive discourse markers (e.g. *you know, well*). Both types of items can be described as metadiscursive instructions given to the hearer on how to interpret an utterance (Brinton, 2008; Hansen, 2006) or in other words as items encoding procedural meaning (Blakemore, 2002; Sperber and Wilson, 1993). Existing descriptions of DMs are often designed from the perspective of either the spoken or the written mode. There is however no principled reason for this separation, as many DMs like *so* or *because* are equally used in both modes, although in some cases with partially distinct functions. In order to develop a principled comparison between the use of DMs in the spoken and the written modes, we present in this paper a first attempt to use a single taxonomy to annotate DMs across both the spoken and written modes.

While DMs have a number of syntactic and prosodic features, the annotation scheme described in this paper targets their meanings only. We report more specifically two annotation experiments that were conducted in order to evaluate the replicability of a functional tag set (Crible, 2014), originally designed for spoken French and English, to written corpora on the same languages. In the first experiment, we tested the application of definitional criteria in order to select candidate tokens of DMs in corpus data. In the second experiment, we used a functional tag set to annotate the meaning of these DMs in four corpora encompassing two languages (English and French) in the spoken and the written modes.

The paper is structured as follows. In Section 2, we introduce a functional definition of DMs and briefly discuss the taxonomy of relations used in our experiments. In Section 3, we present the data and methodology used for the selection of DMs and discuss the results from this experiment. In Section 4, we report the sense annotation experiment and compare inter-annotator agreement across the two languages and modes tested. We conclude in Section 5 and present some perspectives for future work.

2 Defining a taxonomy of discourse markers applicable to spoken and written data

Studies attempting to provide definitions for the category of DMs are numerous, but no consensus has been reached yet on the list of features characterizing this category. Definitions vary greatly depending on the framework and the type of data that is included: monolingual vs. multilingual corpora, written vs. spoken mode, genres or situations (e.g. more or less formal). This rather chaotic situation is caused by the formal heterogeneity of these pragmatic items, which can only be grouped by their overarching function, *viz.* their role as metadiscursive interpretation cues encoding the speaker's internal representation of discourse in a hearer-oriented design. Authors usually agree on including conjunctions (*but, because, although*), some adverbs (*actually, well*), particles (*oh, hum*), prepositional phrases (*in fact, in other words*) and verbal phrases (*you know, I mean*), although within syntactic and pragmatic restrictions such as "weak-clause association" (Schourup, 1999) or non-referential meaning. For the present research, we used the following definition of DMs, based on Crible (2014):

> Syntactically optional, non-truth-conditional expressions constraining the inferential mechanisms of interpretation processes. They function on a metadiscursive level as a cue to situate the host unit in a co-built representation of on-going discourse. They do so by either signaling a discourse relation between the host unit and its context, marking the structural sequencing of discourse segments, expressing the speaker's meta-comment on their phrasing, or contributing to interpersonal collaboration.

This definition is functional, inclusive and can therefore capture the various ways in which different languages encode discourse structure as well as the complexity of spontaneous oral conversations, which are a privileged source of linguistic creativity. DMs are indeed more frequent and more varied (formally and functionally) in speech, where they also co-exist with other pragmatic phenomena such as disfluencies, politeness expressions, interjections, or modal particles with which they can be particularly confusing, as noted by Cuenca (2013) who talks of "fuzzy boundaries" between modal marking and discourse marking for example.

Crible (2014) designed an annotation protocol following the above definition of the category of DMs. To structure the multifunctionality of its members, four functional "domains" (Sweetser, 1990) have been identified from a critical review of previous works (Gonzalez, 2005; Halliday and Hasan, 1976; Redeker, 1990) and their empirical soudness when applied to corpus data. These domains correspond to macro-functions of DMs and each one includes a list of possible values, twenty-nine in total:

- Ideational: relations between real-world events. Includes cause, consequence, contrast, concession, condition, alternative, temporal, exception;
- Rhetorical: relations between epistemic and speech-act events, and metadiscursive functions. Includes motivation, conclusion, opposition, relevance, reformulation, approximation, comment, specification, emphasis;
- Sequential: structuration of discourse segments. Includes: opening boundary, closing boundary, topic-resuming, topic-shifting, quoting, enumerating, punctuating, addition;
- Interpersonal: interactive management of the speaker-hearer relationship. Includes: monitoring, face-saving, agreeing, disagreeing.

This taxonomy was designed to meet the balance between extensive coverage of all possible functions of DMs as they are usually described in the literature, i.e. from coherence relations ("because") to more interactional uses ("you know"), and on the other hand, intensive, precise definition of the different categories so that they do not overlap. A similar fourfold system can be found in Haselow (2011), although without any operational criteria. These domains are defined and motivated with more detail in the annotation protocol.

The multifunctionality of DMs is also reflected in the scheme by the possibility to assign simultaneously two tags, either from the same domain or from two different ones. This accounts for the polysemy of some DMs and their ability to encode several meanings (e.g. Petukhova and Bunt, 2009), as in the following examples of (1) cause and temporal relations (both ideational) and (2) opposition (rhetorical) and topic-shift (sequential):

(1) "Rising dismay at Honohan's judgment crystallised into outright scepticism **after** an extraordinary interview with Bloomberg business news on May 28th last year." (COMTIS corpus, 210).

(2) "I think I've learnt a lot more in the intervening years and it might be nice to go back and work on those. **But** essentially since then I've been working pretty much full-time on trying to write poetry" (Backbone bb_en025 "creative writing").

As opposed to the Penn Discourse Treebank (PDTB) (Prasad et al., 2007), this model differentiates a function in one domain from its equivalent in another, for instance ideational cause and its rhetorical counterpart, motivation. Distinction of these frequent pairs at the first level of annotation allows for each tag to be autonomous and direct, while the PDTB suggests a system of levels, starting from a generic term (e.g. "contingency") and then specifying in several sublevels the particular meaning (e.g. "cause"; "reason" or "result"; "pragmatic" or "non-pragmatic"). Apart from this difference, the present model generally adopts the general approach to DMs as proposed by the theory-neutral and lexically-based framework of the PDTB[1], and more specifically its revision by Zufferey and Degand (2014).

The PDTB taxonomy was designed for written data and has scarcely been applied to spoken corpora (Demirsahin and Zeyrek, 2014; Tonelli et al., 2010). Our tag set has been adapted to speech using a corpus-based methodology: the original taxonomy was tested on spoken corpora and modified in order to better account for the specificities of this mode as they were encountered in authentic data. Therefore, the innovation of the research described here is to assess to what extent the twenty-nine functions identified by Crible (2014) are, in return, applicable to the written mode. Our goal is to reach a single multimodal annotation scheme, in order to prevent the multiplicity of frameworks and their lack of communicability[2].

3 Experiment 1: identification of candidate DMs

3.1 Data and procedure

The first experiment consisted in the identification of occurrences of discourse markers by two expert coders, with French as mother tongue and excellent proficiency in English. Although both have experience in the multilingual annotation of discourse markers, one is a specialist in written corpora while the other works with spoken corpora.

The dataset used to test the identification of DMs was comprised of four texts of ca.1000 words each, in spoken and written French and English, from the spoken corpus of face-to-face interviews *Backbone* (Kurt, 2012) and the written corpus of newspaper articles collected by the COMTIS project[3].

We proceeded in two steps: first, identification in the written texts, based on the assumption that they would be less problematic to annotate; then in the spoken texts, once potential issues had been identified. The selection on written texts was performed without prior discussion of the category, but merely using its definition from the annotation scheme as stated above. After discussion of the disagreements and identification of recurring problems, we moved on to the selection of DMs in the spoken texts.

[1] As opposed to relation-based frameworks like Rhetorical Structure Theory (e.g. Taboada, 2006) or Segmented Discourse Representation Theory (Asher and Lascarides, 2003) which analyze and annotate discourse spans or relations, rather than the discourse markers themselves, thus involving heavier theoretical background.

[2] This work is conducted as part of the ongoing COST Action Network TextLink (IS1312) "Structuring Discourse in Multilingual Europe", chair: L. Degand. `http://textlinkcost.wix.com/textlink`.

[3] `http://www.idiap.ch/project/comtis`

3.2 Results and discussion

The results from the identification experiment are reported in Table 1.

	Coder 1	Coder 2	Selected by both	Relative agreement	Missing in coder 2	Added in coder 2
EN_sp	54	70	51	82.25%	3	19
FR_sp	81	77	69	87.34%	12	8
EN_wr	20	28	19	79.16%	1	**9**
FR_wr	19	26	15	66.67%	4	**11**

Table 1: Absolute and relative agreement on the independent selection of DMs

At a general level, it is noticeable that the annotation of DMs in written texts is not easier than in spoken corpora, even though spoken data is often more diverse and complex to annotate than planned speech. DMs in writing are thus not particularly easier to identify, as demonstrated by the high number of disagreements[4] in this mode as well. However, the types of disagreement are different in the two modes. For instance, in writing, some coders include temporal connectives and prepositional phrases such as *in order to*, which can be problematic if not specifically addressed in the annotation scheme. The relevant criterion that resolved this confusion was that of semantic-syntactic independence (i.e. completion, autonomy) of the connected unit, which, in the case of *in order to*, would not be met by the following infinitive clause. On the other hand, speech-specific phenomena like turn-initial response signals (*okay, yeah*) or fillers may confuse the selection, since they are sometimes considered as DMs in the literature given their pragmatic function. Here, the annotation scheme must specify the precise conditions under which such expressions can be selected as tokens of DMs.

Finally, we also found that coders have different biases depending on their area of expertise. More specifically, coders identify more potential candidates in the modality they are more used to work with: we can observe that coder 2 (expert in writing) identified more tokens in the written texts (cf. bold-faced cells in the table). This result advocates for enhanced training and discussions even between expert coders, and a more prescriptive definition of the DM category than was originally provided by the protocol. As a result, the final version of the definition lists the following criterial features for the selection of candidate tokens:

- procedural meaning within one of the four functional domains;
- syntactic optionality: their removal does not alter the grammaticality of the utterance;
- scope over syntactically and semantically independent units: there must be a finite or implicit predicate, which excludes relative and non-finite clauses, and nominal phrases except when these are acting as a-verbal predicates;
- high degree of grammaticalization: fixed multi-word units, frequently used (not idiosyncratic) and semantically non-compositional;
- incompatibility with membership in the categories of fillers, interjections, response signals, epistemic parentheticals, general extenders, tag questions and editing terms.

Although the authors have not yet tested the extent to which this new definition improves the identification process, the boundaries between DMs and similar expressions are more directly addressed than they were before. Motivations for theses choices are detailed in the annotation protocol.

[4]Kappa scores could not be computed given the unequal number of responses between coders. The percentages represent the ratio of commonly chosen DMs on the total number of tokens selected by both coders.

4 Experiment 2: annotation of discourse functions

4.1 Data and procedure

In both languages, the corpora used in this experiment contained ca.1500 words for speech and 3100 words for writing, in different texts from the same corpora as above (Kurt, 2012 and COMTIS project). In each subcorpus (written and spoken, French and English), we annotated 100 tokens of DMs. For the spoken texts, we didn't use sound files in order to keep the annotation process as comparable as possible in both modes, even though this has been showed in previous research to increase the level of inter-annotator agreement for the identification of DMs (Zufferey and Popescu-Belis, 2004). The functional annotation was performed on DMs selected by one coder only, in order to prevent selection-related disagreements in this experiment.

As in experiment 1, we started the annotation without prior discussion of the guidelines but only used the instructions as they were provided by the annotation protocol, as any isolated researcher would do in the same situation. This was done in order to evaluate the self-sufficiency of the protocol. The instructions were presented in the form of a list of function tags (e.g. cause), the definition for each tag (e.g. "causality of two real-world events"), criteria for the use and disambiguation of tags (e.g. "applies to facts, even future or hypothetical events"), sometimes a paraphrase for specifically ambiguous functions (e.g. "this happened because") and authentic examples from the *Backbone* spoken corpus (Kurt, 2012) (e.g. "they do struggle because sometimes it's their first experience").

We performed the annotation independently in the following order: written French, spoken French, written English, spoken English. Disagreements were discussed after the annotation of each sub-corpus, thus progressively improving the scheme by making each problematic bias or boundary more explicit when possible. Cases of double tags (i.e. when two simultaneous functions were assigned to the same item) were not counted as disagreements when at least one tag was common to both coders.

4.2 Results and discussion

The results from the sense annotation experiment are reported in Table 2[5].

Corpus	Percentage of agreement
written French	44%
spoken French	52%
written English	34%
spoken English	49%

Table 2: Inter-rater agreement scores on sense annotation

These results seem to indicate that spoken data may be easier to annotate than written data, as the level of inter-annotator agreement is always higher. However, as spoken corpora were annotated after written corpora in both languages, this result might also reflect the effect of training. The latter effect was not carried over between the two languages however, as annotations in English did not lead to a better agreement than in French, even though it was performed after discussions of the two French corpora. This may be due to the fact that the annotators are not native speakers of English, which may have caused more uncertainties about the senses conveyed by DMs. Indeed, previous research has shown that learners have uncertain judgments about the correct and incorrect uses of connectives when their L1 produces negative transfer effects, even at advanced stages of language learning (Zufferey et al., to appear).

For all corpora, the sources of disagreement were located in three dimensions. The first problem was the distinction between ideational and rhetorical relations. As mentioned above, the annotation scheme

[5]Again, the incremental process of the annotation did not allow us to compute kappa scores since the successive annotation rounds were not independent of each other. However, mere percentages have been used elsewhere in similar cases, for example in the PDTB (e.g. Miltsakaki et al., 2008).

encodes this difference in the functional tags themselves, and not as a separate level as in the PDTB. Despite the benefits of a direct use of tags exposed above (see section 2), many problems originated from these disambiguations, as in examples (3) and (4):

(3) "I've begun to take my writing a little bit more seriously in the sense that I see it as part of what I do professionally as well as personally, and **so** I've started trying to develop more of a profile" (*Backbone* en_025 "creative writing")

(4) "you've got rhythms, you've got cadence, you've got rhyme schemes potentially, you've got possibilities of evoking visual scenarios, possibilities of evoking sounds and **so** it's very multimodal" (*Backbone* en_025 "creative writing")

The token "so" in (3) signals a semantic (ideational) relation between the fact of "taking one's writing more seriously" and "trying to develop a profile", while in (4), the speaker introduces more of a conclusion, an epistemic (rhetorical) consequence between a number of features of poetry and its evaluation as being multimodal. These examples illustrate how complex it is to grasp the thin line between facts of personal history (3) and personal evaluation of facts (4), as authors/speakers are somehow always involved in their discourse, although not to the same extent.

The second issue concerned the distinctions between semantically overlapping functions, such as conclusion vs. reformulation, addition vs. specification, opening boundary vs. topic-shift, which have close meanings from the same domain. Ambiguity of these functions (and of their criteria as defined in the protocol) is thus responsible for a great number of disagreements.

The third source of disagreement was our discovery of missing functions in the taxonomy, such as a tag encoding the meaning of "goal", to annotate tokens of DMs like *in order to*. This particular issue was addressed by assimilating the missing function to an existing tag ("goal" was grouped with "consequence" as was recommended by the PDTB) so that no *ad hoc* category was needed. Moreover, if certain functions simply do not exist in writing because they require a two-way interaction, some features of writing related to DMs did seem to emerge from our experiment, namely rhetorical or emphatic addition (furthermore", French "en outre", "de plus") and start of a new paragraph. The former was assimilated to the existing value "addition" with a small modification in the definition of the function, while the latter was grouped with "opening boundary" which, in speech, corresponds to a new turn of speech. Again, we chose not to create ad hoc categories but to try and integrate written specificities into the existing tag set.

We also observed that the annotation of spoken and written data involved different kinds of mode-specific problems. For instance, a recurring problem in spoken texts was the use of tags for speech-specific functions (e.g. monitoring, punctuating), given the inherent ambiguity of their "bleached" meaning and their absence in written texts. These particular functions were complex to agree upon, since their core meaning is not as explicit as a more traditional DM such as *because*, or a more monosemous expression such as *for example* which almost always expresses specification. Punctuating DMs, on the other hand, can take various forms (*well*, *I mean*, *I don't know*, *then*, etc.) and are thus less consensually identified.

Another cross-modal issue is the perceived boundary between ideational and rhetorical relations: in writing, subjectivity and interactivity are much less tangible than in speech where speakers often express their direct opinion and involve the hearer in their speech. Such medium-related tendencies led to a different bias, again reflecting each coder's expertise: coder 2 (expert in writing) would include more tokens as "pragmatic" DMs as soon as the writer's opinion is involved (as in example (5)), when coder 1 would have a more restricted understanding of "pragmatic" which is consistent with the high subjectivity of speech and requires a stronger involvement of the speaker (as in (6)), here expressed as a clear judgement or interpretation, instead of a factual event.

(5) "Les nouveaux taux devraient être supportables en Allemagne, **mais** ils vont précipiter plus avant dans le gouffre le marché immobilier et les banques" (COMTIS, 209).
*The new rates should be bearable in Germany, **but** they will plunge further into the abyss the real estate marker and the banks.*

(6) "Si tout se passe comme prévu, ce qui est d'ailleurs toujours le cas, la dette publique irlandaise atteindra les 250 milliards d'euros, **mais** ces différences sont sans importance." (COMTIS, 210).
If everything goes as planned, which is by the way always the case, the Irish national debt will reach 250 billion euros, ***but*** *these differences do not matter.*

As a result of this annotation experiment, the present annotation scheme was improved by: a greater precision in the criteria used to disambiguate similar functions (e.g. contrast vs. concession, temporal ordering vs. consequence); the systematic addition of a paraphrase for each possible value; the inclusion of specific sections in the protocol dedicated to ambiguous meanings (frequent polysemous DMs such as *and*, *but*, *so* etc. and semantic-pragmatic pairs). But this further operationalization of the taxonomy is only a qualitative, yet valuable, assessment of the methodological improvements. What inter-rater agreement analysis brings to light, and the main point of this study, is primarily the realization that many decisions that we make as annotators are implicitly biased, which leads to inevitable disagreement if not documented in the annotation scheme. Another lesson from our experiment is the necessity of training, even for expert coders, and the importance of discussing problems and decisions before launching a large-scale annotation campaign.

5 Conclusion

In this paper, our aim was to report two annotation experiments designed to assess the applicability of a functional definition of the category of DMs in order to reliably identify them in corpus data, and to assess the use of a taxonomy for DMs originally designed for speech to both spoken and written data. The results demonstrate that annotating spoken data does not lead to lower agreements compared to written data, contrary to what was expected. In addition, the differences between spoken and written data are located in the types of disagreements that they generate. However, this primarily qualitative evaluation of the taxonomy would require more data and a more systematic annotation procedure to validate these tentative results.

More generally, this pilot study makes yet another case for training and discussion while conducting annotation by several coders, and stresses the importance of a well-documented annotation scheme which provides detailed instructions and potential transfers between its tag set and other frameworks or data types, as the level of inter-annotator agreement systematically increases from the first to the second annotation performed within a language. The fact that this improvement was not carried over between the two languages reflects the fact that the marking of discourse structure is variable, even between typologically related languages (e.g. Degand, 2004; Pit, 2007; Zufferey and Cartoni, 2012), and the meanings and usage of discourse markers are therefore always at least partially language-specific. Indeed, languages vary in their encoding of discourse relations. To make a case in point, Dutch uses two specific connectives to convey ideational and rhetorical causes while English uses only one connective ("because"). French uses two specific connectives as well but one of them ("car") is also restricted to the written mode, creating register differences with Dutch.

The major outcome of this study is therefore not the quantitative reliability of the taxonomy, but rather the illustration of some methodological best practices for sense annotation in general, to raise awareness to recurring problems in discourse marker studies in particular.

Future perspectives for the annotation of DMs are the application of the coding scheme described in this paper to the modality of gestures (Bolly and Crible, 2015), the comparison of annotations performed by naive vs. expert coders (Crible and Degand, 2015), the annotation of DMs in speech with and without the help of prosody (i.e. with the sound files); and the comparison of inter-annotator agreement scores obtained by native and non-native speakers (e.g. a French coder annotating English data). Another perspective would be a comparative study between this multimodal annotation scheme and the ISO standard for discourse relations (Bunt et al., 2012) to situate the present approach within interoperable endeavours.

References

Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge : CUP.

Blakemore, D. (2002). *Relevance and linguistic meaning. The semantics and pragmatics of discourse markers*. Cambridge : CUP.

Bolly, C. and L. Crible (2015). From context to functions and back again: Disambiguating pragmatic uses of discourse markers. In *International Pragmatics Association (IPrA) Conference, July 26-31, Antwerp, Belgium*.

Brinton, L. (2008). *The comment clause in English : syntactic origins and pragmatic development*. Cambridge : CUP.

Bunt, H., R. Prasad, and A. Joshi (2012). First steps towards an iso standard for annotating discourse relations. In *Proceedings of the Joint ISA-7, SRSL-3, and I2MRT LREC 2012 Workshop on Semantic Annotation and the Integration and Interoperability of Multimodal Resources and Tools*, Istanbul, Turkey.

Crible, L. (2014). Selection and functional description of discourse markers in french and english: towards crosslinguistic and operational categories for contrastive annotation. In *International Workshop - Pragmatic Markers, Discourse Markers and Modal Particles: What do we know and where do we go from here? October 16-17, Como, Italia*.

Crible, L. and L. Degand (2015). Functions and syntax of discourse connectives across languages and genres: Towards a multilingual annotation scheme. In *International Pragmatics Association (IPrA) Conference, July 26-31, Antwerp, Belgium*.

Cuenca, M. J. (2013). The fuzzy boundaries between discourse marking and modal marking. In L. Degand, B. Cornillie, and P. Pietrandrea (Eds.), *Discourse markers and modal particles. Categorization and description*, pp. 191–216. Amsterdam : John Benjamins.

Degand, L. (2004). Contrastive analyses, translation, and speaker involvement: the case of *puisque* and *angezien*. In M. Achard and S. Kemmer (Eds.), *Language, Culture and Mind*, pp. 1–20. Stanford: CSLI Publications.

Demirsahin, I. and D. Zeyrek (2014). Annotating discourse connectives in spoken turkish. In *LAW VIII - The 8th Linguistic Annotation Workshop*, pp. 105–109.

Gonzalez, M. (2005). Pragmatic markers and discourse coherence relations in english and catalan oral narrative. *Discourse studies 77(1)*(1), 53–86.

Halliday, M. A. K. and R. Hasan (1976). *Cohesion in English*. London : Longman.

Hansen, M.-B. M. (2006). A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of french toujours). In K. Fischer (Ed.), *Approaches to discourse particles*, pp. 21–41. Amsterdam : Elsevier.

Haselow, A. (2011). Discourse marker and modal particle : the functions of utterance-final *then* in spoken english. *Journal of Pragmatics 43*(14), 3603–3623.

Kurt, K. (2012). Pedagogic corpora for content and language integrated learning. insights from the backbone project. *The Eurocall Review 20*(2), 3–22.

Miltsakaki, E., L. Lee, and A. Joshi (2008). Sense annotation in the penn discourse treebank. *Lecture Notes in Computer Science 4919*, 275–286.

Petukhova, V. and H. Bunt (2009). Towards a multidimensional semantics of discourse markers in spoken dialogue. In *Proceedings of the 8th International Conference on Computational Semantics*, pp. 157–168.

Pit, M. (2007). Cross-linguistic analyses of backward causal connectives in dutch, german and french. *Languages in Contrast 7*(1), 53–82.

Prasad, R., E. Miltsakaki, N. Dinesh, A. Lee, and A. Joshi (2007). The penn discourse treebank 2.0 annotation manual. Technical report, Institute for Research in Cognitive Science.

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

Schourup, L. (1999). Discourse markers. *Lingua* (107), 227–265.

Sperber, D. and D. Wilson (1993). Linguistic form and relevance. *Lingua 90*, 1–25.

Sweetser, E. (1990). *From etymology to pragmatics*. Cambridge : CUP.

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics 38*, 567–592.

Tonelli, S., G. Riccardi, R. Prasad, and A. Joshi (2010, may). Annotation of discourse relations for coconversation spoken dialogs. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, pp. 2084–2090. European Language Resources Association (ELRA).

Zufferey, S. and B. Cartoni (2012). English and french causal connectives in contrast. *Languages in contrast 12*(2), 232–250.

Zufferey, S. and L. Degand (2014). Representing the meaning of discourse connectives for multilingual purposes. *Corpus Linguistics and Linguistic Theory 10*.

Zufferey, S., W. Mak, L. Degand, and T. Sanders (to appear). Advanced learners' comprehension of connectives. the role of l1 transfer across online and offline tasks.

Zufferey, S. and A. Popescu-Belis (2004). Towards automatic identification of discourse markers in dialogues: the case of *like*. In *5th SIGdial Workshop on Discourse and Dialogue, Cambridge (MA)*, pp. 63–71.