Studying Discourse from Corpus and Experimental Data: Bridging the Methodological Gap

Ludivine Crible

Studying Discourse
from Corpus and Experimental Data:
Bridging the Methodological Gap

Ludivine Crible

Ghent University

Studying Discourse from Corpus and Experimental Data: Bridging the Methodological Gap

Ludivine Crible

Ghent University

· ··

This paper starts from the observation that in discourse analysis, unlike in other fields of linguistics, very few studies combine corpus-based and experimental data. It seeks to understand the reasons behind this methodological monism by taking into account the specificities of the discourse level and the benefits and drawbacks of each method. In the first part of the paper, challenges facing the combination of corpora and experimentation are developed. In the second part, a selection of mixed-method discourse-analytic studies is reviewed to illustrate the different ways in which corpora and experiments can interact, how much convergence (or lack thereof) can be found across production and comprehension paradigms, and the technical and methodological solutions that researchers have found to make the data as comparable as possible. This structured literature review concludes with the recommendation that discourse analysis can benefit from more mixed-method approaches, provided that the methods are planned with the respective affordances of each method in mind.

Keywords: discourse relations, discourse markers, corpus-based, experimental, literature review

· ··

Cet article part du constat qu'en analyse du discours, contrairement aux autres domaines de la linguistique, très peu d'études combinent des données de corpus et des données expérimentales. Il s'attache à comprendre les raisons de ce monisme méthodologique par l'examen des spécificités du niveau discursif ainsi que des avantages et inconvénients de chaque méthode. Dans la première partie de l'article, nous développons les défis auxquels la combinaison de corpus et d'expériences doit faire face. Dans la seconde, une sélection d'études multi-méthodes en analyse du discours est revue afin d'illustrer les différentes façons dont corpus et expériences peuvent interagir, la convergence (ou son absence) entre les paradigmes de production et de compréhension, et les solutions techniques et méthodologiques que les chercheurs ont trouvées pour rendre leurs données aussi comparables que possible. Cette revue de la littérature structurée se conclut par la recommandation que l'analyse du discours a beaucoup à gagner d'approches multi-méthodes, pourvu que chaque méthode soit planifiée en tenant compte des possibilités de l'autre.

Mots clés : *relations du discours, marqueurs du discours, corpus, expérimental, revue de la littérature*

1. Introduction

1 Discourse is one of the highest levels of the linguistic system above morphology, syntax or semantics, as it deals with beyond-the-clause phenomena and, within these, relations between utterances and the way they connect to build a text[1]. Without

---

the discourse level, independent sentences would only make up an incoherent juxtaposition of propositions, so that discourse is crucial to express one's ideas and intentions: as Crible and Degand (2019: online, §1) put it, "discourse is where the magic happens". Another specificity of the discourse level is that almost all elements that perform a discourse function are "born" and recruited from other, lower linguistic levels after a process of conventionalization, and are therefore on a cline from more grammatical to more discursive uses. Because of this ambivalence, discourse phenomena are often multi-faceted, multifunctional and highly sensitive to context variation, which makes them particularly challenging to study.

2   Most discourse-analytic research investigates coherence relations (causality, temporality, contrast, etc.) and their signals (connectives or discourse markers and, more recently, other segment-internal elements; Das & Taboada, 2018). The bulk of these studies explore the interplay of factors that impact the forms and functions of discourse, such as syntactic position, register, prosody, sociodemographics, and many more. Authors tend to do this by specializing in one of two methods: corpus-based analyses or experimental studies. Both methods have contributed greatly to research in discourse analysis and are highly complementary: corpus studies help describe a complex reality by giving access to large amounts of (more or less) naturally produced data, while experiments offer more control over variables and can provide insights into online processing and comprehension. Corpora and experiments each have benefits and drawbacks, yet most researchers only exploit the affordances of one approach, and mixed-method discourse studies are very rare.

3   However, this situation does not apply as much to other linguistic levels besides discourse. Gilquin and Gries (2009) pointed out that many studies combine experimental and corpus data, although psycholinguists are more familiar with this combination than corpus linguists. By sampling papers from various journals, the authors found that studies with a corpus-linguistic perspective rarely resort to experimentation and, when they do, do not exploit the full range of available methods beyond acceptability judgments. The authors further showed that corpus linguists tend to adopt an exploratory approach to the data and that most studies (60% of their sample) concern the lexical level, whereas psycholinguists mostly adopt a hypothesis-testing approach and focus more on syntax (44% of the sample that combines corpus and experimental data, against only 6% on pragmatics).

4   Such a quantitative meta-study would not be possible within the field of discourse analysis, given the rarity of mixed-method studies. While such a combination would allow researchers to overcome the limitations of each method and could cover all aspects of a phenomenon from production to online processing, only a handful of discourse studies jointly report findings from corpus-based and experimental data. There might be several explanations for this: a single team of researchers might not possess the skills and/or equipment to work with multiple techniques; access to both corpus data and participants can be costly; multi-disciplinary studies are not always welcome or valued; individuals may be interested in or prefer one method

over another. While these might be not-so-good reasons, there are other obstacles and challenges that refrain the field from embracing the full affordances of both corpora and experiments.

5   The present paper intends to identify the theoretical and methodological challenges facing mixed-method discourse analysis. It then provides a selective review of studies that have overcome these challenges, thus illustrating the multiple ways in which corpora and experiments can be combined to explore the discourse level, with a particular focus on studies investigating discourse relations and connectives. In doing so, I hope to encourage more research at the crossroads between corpus linguistics and psycholinguistics, in order to cover all aspects of discourse production, processing and comprehension in a more integrated manner.

2.   Discourse analysis and the challenges of combining corpus-based and experimental data

6   Researchers in discourse analysis face three main challenges when planning mixed-method studies. The first two are methodological in nature and relate to i) the identification and disambiguation of discourse phenomena and ii) the intricate relation between discourse relations or markers and their context. The third one is more theoretical and less specific to discourse, to a certain extent: production, perception and comprehension (offline and online) do not always match. In this section, I discuss each of these main challenges and suggest workarounds whenever possible.

2.1.   Discourse-level phenomena are costly to identify and disambiguate

7   In most cases, mixed-method studies start by reporting corpus analyses of a given linguistic element, before they confront these findings with experimental measures. The particular focus of discourse analysis, targeting phenomena beyond the clausal level, can make it more time-consuming and less straightforward for researchers to extract from corpus data the relevant structures that will be tested. This is an issue that does not affect lexical studies, where forms (lemmas) can be automatically identified by concordancers. It does concern other levels besides discourse, such as syntactic constructions (which can require a certain level of abstraction) or prosodic parameters: while regular expressions and scripts can assist these analyses, some features are complex to handle automatically and require a considerable amount of human intervention. However, whether it is lexis, syntax or prosody, most analyses focus on form, so that extraction is often the final methodological step.

8   In discourse, however, the difficulty of identification is enhanced because most studies are not only interested in forms but also in meanings and functions, which necessarily implies an extra step of semantic-pragmatic disambiguation. This is particularly true for studies investigating discourse relations and connectives, which are notoriously ambiguous and polyfunctional (e.g., Spooren & Degand, 2010). Firstly, there is no one-to-one mapping between a given connective and a given

type of relation, and most connectives can occur in multiple relations (e.g., *while* in contrastive, concessive or temporal relations; French *alors* in temporal, consequence, concessive or topic-shifting uses). Secondly, some items perform additional functions besides their discourse-relational uses, especially for discourse markers used in spoken language such as *well*, *I mean* or *you know*, which can be used for discourse functions or in their propositional sense in different contexts; this is also true for more "standard" connectives such as *so* (intensity adverb vs. causal connective) or *however* (relative pronoun vs. concessive connective).

9   Even for connectives that are monosemous, such as *by contrast* (contrast) or *because* (cause), in many cases researchers are interested in fine-grained distinctions of meaning, using notions such as Sweetser's (1990) domains of use (content, epistemic, speech act) or Sanders et al.'s (1992) cognitive primitives (objective vs. subjective, basic vs. non-basic, etc.). Most connectives can indeed express several variants of one (or more) relation, and it is this variation (and the contextual factors that explain it) that motivates most studies in discourse analysis. For instance, several studies have addressed the division of labor between the three causal connectives in French *parce que*, *car* and *puisque*, whose association to the content, epistemic and speech-act domains is far from straightforward (e.g., Lambda-l Group, 1975; Degand & Pander Maat, 2003; Simon & Degand, 2007; Zufferey, 2012). Others have applied similar classifications to large sets of expressions and entire categories, particularly in spoken discourse (Cuenca, 2013; Crible, 2018). Crible and Degand (2019) indeed report that, in French conversations, four discourse markers (namely *et* ["and"], *alors* ["so/well"], *donc* ["so"], *mais* ["but"]) can be used in all four domains included in their classification model (ideational, rhetorical, sequential and interpersonal), and 15 out of 33 marker types in their sample can function in two or more domains. Such a wide functional spectrum is typical of discourse markers, to the point that most definitions include it as a categorical criterion (e.g., Brinton, 1996; Crible, 2018).

10   This means that these and similar studies on connectives and discourse markers must not only extract but also disambiguate the item(s) under scrutiny. This crucial methodological step can be quite time-consuming and requires added expertise and workforce: i) a coding scheme must be designed, if none of the existing proposals is relevant; ii) it must be tested and operationalized on pilot data so that its application to the corpus is robust and reproducible; iii) the data can then be disambiguated; iv) at least one other analyst must disambiguate (part of) the data, in order to measure inter-annotator agreement. This last procedure tests the extent to which the two or more sets of interpretations converge, which attests to the reliability of the analysis. Spooren and Degand (2010) argue that thresholds for substantial inter-annotator agreement should be lower in discourse annotation than in other fields due to the extreme functional variation and underspecification of connectives. Thus, performing functional classification of connectives and discourse markers is not a trivial task: Crible and Degand (2017) indeed observed significant differences in agreement between experts and naïve annotators. This points to the need of adequate training in discourse analysis, training which is not necessarily available

to psycholinguists. Alternative methods of disambiguation such as crowdsourcing connective insertions may avoid the issue of training but do not solve all the problems regarding fine-grained distinctions, polyfunctionality and annotators' disagreements (Yung et al., 2019).

11   Not only is data interpretation challenging for the researchers, it is also complex to control for participants in experimental studies. One cannot assume that participants will interpret an item in the same way as an expert researcher, either because multiple interpretations are possible or because they might not concentrate on the meaning of the connective or marker and only process it in a shallow way, which is particularly problematic for the fine-grained distinctions discussed above. Therefore, when discourse meaning is a variable in the design, the experiment should be designed so that the task allows the researcher to check how the item was interpreted. There are two main options to do this. The first one is to choose a task that forces the participant to make their interpretation explicit through their answer: selecting a sentence continuation, a picture or even a gloss that represents the target meaning (e.g., Didirková et al., 2019; Morera et al., 2017). The different options presented to the participant in this case must also be pre-tested (i.e., normed by several judges) to make sure that they indeed represent what the researchers have planned.

12   The second option is to add comprehension questions after the main task (e.g., reading or judgment tasks followed by questions). Many experiments include such questions to maintain the attention level and discard inattentive participants. In the case of discourse, I would suggest that such questions are necessary to make sure that any effect that is observed is indeed related to the particular interpretation targeted by the design. Moreover, Crible and Pickering (2020) observed that subtle meaning effects can only emerge with questions that specifically aim at disambiguating the discourse relation, unlike general yes-no questions: otherwise, according to the authors, participants only process the sentences in a shallow way and do not make all possible inferences. This particularly applies to ambiguous or underspecified connectives such as *and*, which does not encode much semantic information but can still occur in relations of consequence, contrast or temporality (Luscher & Moeschler, 1990). Experiments that only rely on the stimuli to generate the intended inferences thus run the risk of overestimating the meaning effects they aim to observe, especially with passive tasks and online measures (i.e., reading times), which do not give access to interpretation unless complemented by offline measures of comprehension.

13   In sum, discourse analysis is almost always qualitative to some extent, in the sense that it involves interpreting contexts of use and classifying them along semantic-pragmatic criteria, whereas experimental studies are mostly quantitative and measure differences in numerical variables such as reaction times or ratings. Interpretation is a key component of discourse that must be operationalized and checked, whether it is performed by the corpus linguist or by the participant in an experiment. It becomes particularly challenging when meaning variation is the object of experimental manipulation, as we will now discuss.

2.2. The role of discourse context is difficult to manipulate

14   In the previous section, we saw that meaning variation is an issue that comes about both at the corpus-based and experimental stages of a mixed-method study. The second challenge facing discourse analysts is one that is more specific to experimentation and, more precisely, to the transition from the corpus to the experiment: the richness and complexity of discourse context, as observed in corpora, is often hard to fit in a controlled experimental design. One of the major differences between the two methods is how context can be accounted for and included in the study. Corpora can indeed give access to a potentially infinite number of variables, either through the metadata (time and place of writing/recording, age and gender of the speakers/writers), the extra-linguistic context (the type of text/activity, the register, the speaker roles) or features of the linguistic context such as the syntactic position of the item, its prosodic contour or co-occurrence chains. The corpus linguist often makes a selection and cannot account for all of these potential sources of influence, but she is reasonably free to cover several, whether in a qualitative way (i.e., showing examples that illustrate the role of context) or in a more quantitative way, by including these variables in multivariate statistical analyses, for instance.

15   By contrast, experimental designs typically include one or two variables of interest, rarely more. This is because psycholinguistic experiments are more "causal" in their approach, that is, they strive to explain one phenomenon by another. To do so, they need to account for the different conditions (i.e., variants of an item) and their interaction, in robust statistical models such as linear or logistic regressions, which cannot handle too many variables or variable levels. For example, if one wants to test the effect of position (initial vs. medial) on the interpretation of two connectives, then the design will have four (2×2) conditions, since both connectives will be presented in both position options. By controlling that nothing else differs in the stimuli, the experiment can be confident that any difference observed in the dependent variable (e.g., disambiguation of the connective) is indeed due to the difference between positions.

16   Variables included in experimental designs can be of different natures (linguistic or not) and are selected because of their hypothesized effect on the phenomenon. This variable selection can be based on theory or, in mixed-method studies, on corpus analyses from which significant trends emerge. In the latter case, the role of the corpus is to identify the features of context which are most likely to have a systematic impact on the phenomenon, as evidenced by clear quantitative trends in the data. Given that corpus data cannot be manipulated by the researcher, and that possibly many features not foreseen or controlled can have their own impact, it is often impossible to ascertain causality in any relation between two variables, so that the effect of variables selected from corpus data does not always translate well once included in an experimental design.

17   While independent variables in experiments are typically chosen for their hypothesized likelihood of having a significant effect, there are other considerations

that must be taken into account when designing an experiment. In particular, the ability of the variable to be manipulated systematically across items is a practical issue that bears as much weight on the final design as the more empirical matter of its significant effect. For instance, recent corpus-based studies have started to uncover the large range of lexical or syntactic elements (often called *signals*) that contribute to expressing discourse relations beyond discourse markers, yet these features of the segments can hardly be modified without strongly affecting the meaning and structure of the sentences. As Das and Taboada (2018: 767) put it:

> the manipulation of [discourse markers] (presence vs. absence) as practiced in many psycholinguistic studies (e.g., Degand & Sanders, 2002), could lead to a change in the relational meaning. The caveat for such manipulation involving other signals is probably stronger. Since other signals are typically integral part of sentences and primarily contribute to the propositional content or grammar of a sentence, removing or modifying such signals (such as lexical or syntactic) may result in significant changes in the propositional content or grammar of the sentences being compared. Thus, the experimental design would have to be more complex.

18   In other words, some features of the context cannot be manipulated without creating unwanted, confounding differences between experimental conditions that hinder their comparability and thus prevent any general conclusions from being drawn. Moreover, online measures such as reading times present the additional difficulty of requiring the critical segment to be identical across conditions, so that the manipulation must apply to other parts of the sentence before the critical segment. This is not always easy when dealing with discourse signals such as semantic relations (e.g., antonyms) or referential chains that are necessarily present in both clauses of a discourse relation (e.g., *Peter is tall but Tom is short*). If the alternative version of a condition cannot be created without completely changing the meaning of the relation or without affecting other elements in the context, then the conditions are not entirely comparable and any observed effect could not reliably be attributed to the given independent variable. This issue is particularly sensitive for discourse studies, since every linguistic level is potentially involved in discourse interpretation, so that it can never be assumed that a change at the lexical, syntactic or prosodic level will not affect participants' comprehension.

19   In sum, corpora and experiments differ wildly in their approach to context: the former can cover a large panel of contextual variables in an exploratory way, whereas the latter are much more focused on one or two features of the context for which they aim to test hypotheses. As a result, the complexity of language that is often uncovered by corpus-based discourse studies cannot always be reflected in experiments, which thus only provide a partial, narrower view of a phenomenon. The transition from corpus-based to experimental methods is thus hindered by this gap between large numbers of variables, on the one hand, and a few systematic manipulations, on the other.

2.3.   Production, perception and comprehension do not always match

20   While corpus-based studies only give us access to production data, behavioral psycholinguistic experiments can address production (through elicitation tasks), perception (judgment tasks), comprehension (interpretation questions) and online processing (reading and reaction times). There is indubitably some relation and overlap between these different processes. For instance, according to Levelt's (1989) model, speakers use their comprehension system to monitor their production. In turn, according to Pickering and Garrod (2007), we use our production habits to understand and predict language. As a result, we can reasonably expect that corpus-based findings would be informative of the factors that impact the acceptability or the comprehension of discourse relations and markers.

21   What corpora can tell us about production mainly relies on frequencies, which are then used as predictors in experimental tasks (e.g., a rare combination of variables should be less acceptable and read more slowly than a more frequent pattern). However, corpus frequencies have important limitations, especially at the lower extreme of the scale: in a study on lexical synonymy, Arppe and Järvikivi (2007) showed that relative rareness in a corpus is associated with dispreference in forced-choice tasks but not with lower acceptability ratings, as it can also be due to an effect of genre or register. So while it might be useful to complement corpus findings with acceptability results in order to investigate rare structures that cannot be found in sufficient quantities in corpora, caution must be taken in interpreting the results. The temptation to use acceptability judgments tasks (which are relatively easy to design and conduct) is thus confronted with the fact that they "reflect the general linguistic insights about what is considered possible or appropriate" (Arppe & Järvikivi, 2007: 151), which is a different process from actual language use. Such introspection might be even more complex in the case of discourse markers, which are optional and often have many alternative options (e.g., French temporal connectives *pendant que*, *tandis que*, *alors que*, *cependant*, etc.), so that judgments of acceptability might not be very stable or reliable.

22   Besides the above-mentioned shortcoming of acceptability ratings, other experimental tasks also require that the corpus-based frequency effects be large enough to be detected in lab settings, especially with online measures. Trends in production are more often relative than clear-cut, especially when it comes to discourse relations and markers: studies that investigate the distribution of functions across connectives and/or across syntactic positions (e.g., Degand, 2014) often find that most combinations of variables are possible, only to a different degree of frequency. While these relative quantitative differences are meaningful, they might not be strong enough to be reflected in significantly slower vs. faster reading times or more vs. fewer selections in forced-choice tasks.

23   It can also happen that theoretical descriptions, while valid and reliably applicable to corpus data, are not confirmed experimentally. For instance, Crible and Pickering (2020) compared reading times for additive and contrastive relations,

a pair that only differs by its polarity (positive vs. negative), following Sanders et al.'s (1992) approach. While positive and negative relations were repeatedly found to differ in their tendency to be implicit or in their age of acquisition, there was no significant reading time difference in their experiment, which led the authors to conclude that the effect of polarity does not carry over to processing cost. Similarly, Arppe and Järvikivi (2007) also pointed out that production as measured by forced-choice tasks only captures differences between two options "when the underlying individual acceptability judgments are sufficiently and significantly divergent" (Arppe & Järvikivi, 2007: 152). It is therefore crucial to select variants of a phenomenon for which the contrast is large enough to be perceived and processed differently in experiments.

24   Not only can corpus frequencies and experimental measures diverge, but within the experimental paradigm, different tasks might tap into different processes. Elicitation tasks (closest to production) will provide the most likely or most natural continuation of a given context. Acceptability or grammaticality ratings assess attitudes towards linguistic variants. It is not always the case that an option that is not spontaneously produced through elicitation would be rated as unacceptable or ungrammatical (cf. Zufferey's [2012] study on *car* vs. *parce que*, reviewed below). Similarly, offline and online measures do not always align: for instance, Crible (2021) found clear differences in coherence ratings between concessive and result relations expressed by affirmative or negative verbs, but only observed a significant increase in reading times between the two relations in the affirmative polarity, which indicates that readers can be aware of a preferred way to encode a given relation, without necessarily struggling to process the dispreferred option. By contrast, comprehension scores can be high even though reading times indicate a higher processing cost: disambiguating the referent of the pronoun *he* in *Peter yelled at John because he was tired* should lead to consensual results (*he* is Peter), although the causal segment might be read more slowly than the more likely continuation *because he broke the lamp*.

25   Overall, whether because the contrasts in production are not large enough or because different tasks give access to different processes, non-convergence between methods is a frequent possibility that should not always be considered as a sign of poor design. The selective literature review in the next section will illustrate cases of full, partial or no convergence between corpus-based and experimental discourse studies.

3.   Selective review of mixed-method discourse studies

26   Now that we have discussed some of the major challenges that mixed-method (discourse) studies might face, we can turn to exploring how previous authors have dealt with them. In doing so, I will illustrate three main ways in which corpora and experiments can be combined: i) corpora as sources for experimental stimuli, ii) corpora as sources of hypotheses to be tested experimentally and iii) corpora as

further explorations of experimental findings. I do not attempt at ranking which method is considered as primary or secondary by the authors but merely use the order of presentation of the studies in each paper as an objective sign of the relation between the datasets.

27   This literature review does not aim at quantitative exhaustivity. However, to provide some idea of the rarity of mixed-method discourse studies, I have conducted a search for publications that contain the terms "corpus and experiment discourse" and "corpus and experiment connectives" on Google Scholar. The search returned 15 different publications in total, excluding studies where "experiment" was used in a computational sense (e.g., machine-learning experiment) and studies that were not about discourse relations and/or discourse connectives. Although additional publications could probably be found in a more extensive search, this indicates that discourse studies combining corpus-based and experimental methods are indeed rare.

28   A complete review of all 15 studies is not possible within the scope of this paper and would lead to redundancy, given the similarity between some of the studies. Therefore, I focus on seven research papers which represent different mixed-method scenarios and illustrate some of the challenging issues that I addressed in the previous section. Priority was given to studies that are i) recent (the oldest in the sample was published in 2011, unlike the seminal studies by Pander Maat & Sanders [2001], Stukker et al. [2008] or Sanders [1997]), ii) focused on discourse relations (unlike Fox Tree [2015] or Döring & Repp [2020], which deal with non-relational discourse markers and modal particles) and iii) varied in the type of experimental task (all the other papers use some kind of offline acceptability task, as in Andersson [2019], Andersson & Spenader [2014], Asr & Demberg [2015] or Schumann et al. [2021]).

29   Finally, this selection does not cover mixed-method studies that are published over multiple papers. It is thus non-exhaustive and mainly aims at covering various scenarios on a fairly similar object of study, namely discourse relations and their connectives.

3.1.   Experiments using corpus-based materials

30   As discussed in the previous sections, psycholinguistic experiments often aim to test hypotheses driven from corpus-based findings and to replicate patterns observed in naturalistic production data. To do so, some studies opt for direct replication by re-using the linguistic materials extracted from the corpus as experimental stimuli. Doing so vouches for the comparability of the methods and the naturalness of the stimuli. As we will see with the first two papers, this endeavor also comes with a set of challenges.

31   In the first paper, Crible and Demberg (2020) investigate the effect of genre formality on the production and perception of the connective *and* in English. Their central hypothesis is that the underspecified uses of *and*, i.e., when it is used in relations other than its core meaning of addition, are only acceptable in informal genres, while *and* is functionally more restricted in formal texts. To test this, the

authors first conduct a corpus-based study in a spoken multi-genre corpus where they compare the functional distribution of *and* in formal (news broadcast, political speech), semi-formal (interview, lesson) and informal contexts (conversation). Some of their main observations are that i) the contrastive use of *and* is never found in the formal settings and is most frequent in informal conversations, and that ii) its use in consequence relations is equally frequent in the three genres.

32   In a second step, the authors turn to experimentation to replicate this pattern with a connective-insertion task (production) and a forced-choice task (perception). The design targets the relation between connectives (*and* vs. *but* and *so*), discourse relations (contrast vs. consequence) and text genre (formal vs. informal). While the corpus study used spoken data, the experiments focused on different degrees of formality in written texts: the Loyola CMC corpus (Goldstein-Stewart et al., 2008) was chosen as source materials because it includes, among others, chat conversations, which are both written and informal. Instances of contrast and consequence relations expressed by different connectives were extracted and pre-tested to confirm the experts' classification; this step was particularly necessary for occurrences of underspecified *and*. The original items then underwent a series of modifications: firstly, for the informal condition, some of the items were corrected for spelling and register (e.g., avoiding slang, irregular syntax). Secondly, to create the formal condition, the same items were transformed into excerpts from comments to online press articles, a genre that was assumed to be familiar to the participants (i.e., most of them would have produced an online comment before). The transformation involved using regular punctuation, removing verb contractions and changing some lexical elements into a more formal register (e.g., "locked up" vs. "imprisoned"). The authors also used two different visual displays to further represent chat messages and online comments.

33   With this set-up, they found partial convergence between the three methods: on the one hand, the low frequency of contrastive uses of *and* in the corpus was matched by a dispreference in the experimental tasks; on the other, the interaction between relations and genres observed in the corpus was not reproduced in the experiments. The authors conclude that the two genres selected for the experiments might not have been contrasted enough to trigger a different behavioral response, and that variation in the manipulations of the items can also explain the absence of a genre effect.

34   This study resonates with some of the issues discussed above and illustrates new ones, namely the importance of large contrasts between conditions in order to find significant effects in experiments, the role of context on discourse interpretation, and the complexity of stimuli creation. Crible and Demberg (2020) made several choices that targeted a compromise between experimental control and comparability with the corpus data. For instance, to mimic natural production in the experiments, they selected a formal genre that participants were likely to have produced themselves (i.e., online news comments), rather than a text type of which they would only have passive knowledge (such as news articles, which few of us produce in our lifetime).

They also restricted the excerpts to a minimal context rather than a longer passage in order to avoid the potential influence of the previous context on the interpretation, although this decision implied that the manipulations were restricted to the pair of sentences in the discourse relation and were thus not always clear or sufficient. To conclude, this first mixed-method paper includes (almost) direct re-use of corpus-based materials, complemented with manipulated versions of the same items, which resulted in only a partial convergence of the results. The central variable in the authors' design, namely text genre and its associated degree of formality, seems to be a difficult factor to manipulate in the limited set-up of experimental tasks.

35   The second paper focuses on another linguistic feature that is complex to manipulate, namely the prosody of discourse markers. Didirková et al. (2019) report on two acceptability and one prediction tasks which test the association between discourse functions and prosodic profile, focusing on the French polyfunctional discourse markers *et* ("and") and *alors* ("then/well"). The authors aim to show that, in production, there are systematic associations between the function of a marker and its prosody which are strong enough to generate preferences in perception and guided predictions: upon listening to a marker, participants should be able to disambiguate its function and anticipate the type of relation that it introduces, even if the marker is highly polyfunctional. As in Crible and Demberg (2020), the creation of the stimuli was quite complex and involved multiple steps, some of them corpus-based and others more experimental. First, the authors extracted the functional distribution of *et* and *alors* from a discourse-annotated corpus, which resulted in four functions for each marker (including two shared functions). For each original item, they constructed three new versions which each represent the other three functions of the marker and which were disambiguated by naïve annotators in order to ensure their interpretation.

36   In a second step, these items were read out loud by participants in order to extract the prosodic profile of each marker-function pair (this step is reported in Didirková et al., 2018). This procedure of elicitation is already closer to experimentation (speech production study) than to original corpus data and allowed the authors to identify three main prosodic profiles (integrated, detached and semi-detached) combining acoustic measures of marker duration, pause duration before the marker and pitch reset on the marker. The contrasts that emerged from this production step also matched semantic-pragmatic distance between the different functions of the markers (i.e., the more integrated the profile, the closer the function with respect to the core meaning). However, there was once more some variation in the extent of these contrasts across items and speakers, which made it impossible to re-use these naturalistic recordings. Therefore, in a third step, all items were recorded again by an expert speaker. During this final recording, the acoustic measures were slightly exaggerated in order to strengthen the contrasts between the three profiles. Pause duration was also automatically post-processed (i.e., made equal for all items within a profile). The items were then manually and automatically checked to ensure that they reproduced the intended prosodic profile.

37   With these materials, Didirková et al. (2019) tested whether a marker produced with its associated prosodic profile would be more acceptable and would generate more consistent predictions than when it is presented in a different prosody (the profile that is not typically produced with that marker/function). They first collected acceptability ratings and answers to a forced-choice task which presented the same excerpt in two prosodic versions (typical vs. non typical profile). Finally, the same items were used in a continuation-prediction task where participants listened only to the beginning of the excerpts up until (and including) the discourse marker: since the first part of the relation was kept constant across conditions, this was meant to test whether prosody alone can cue participants as to the type of relation that is expressed by the marker. Results showed that prosody plays a stronger role with *alors* than with *et*, for which only the integrated profile was acceptable and little consensus was found in the prediction task. Although the authors found a robust bias for the core meaning of each marker across tasks, they were able to show that polysemy and the conceptual distance between the functions of a single marker are reflected in distinct prosodic profiles, which explains the different results between *alors* (highly polysemous, prosodic contrasts) and *et* (one core meaning, one prosody). The larger convergence between tasks and methods in this study was obtained through considerable work on the stimuli, with a mix of corpus-based and experimental methods for their extraction and manipulation. The procedure used by Didirková et al. (2019) vouches for highly controlled yet naturalistic items with strong contrasts, which led to robust and important findings on the mapping between prosody and discourse.

3.2.   Experiments to confirm corpus-based findings

38   This second set of mixed-method studies includes three papers that start with corpus-based findings which are then tested experimentally. We will see that each study represents a different degree of convergence across methods. In Asr and Demberg (2020), the authors report on an impressive set of results combining corpus data, sentence continuations, coherence judgments and eye-tracking data on the contrastive vs. concessive uses of *but* and *although*. The specificity of this study is that it compares two highly similar connectives with an overlapping functional spectrum, and two discourse relations that are conceptually close (both "negative" relations). They intend to show that the meaning of a connective can be determined by studying its use in language production, and that the meaning distribution of *but* and *although* is robust and entrenched enough to create differences in acceptability and processing, despite their functional overlap.

39   Asr and Demberg (2020) first investigate how *but* and *although* are distributed in the Penn Discourse Treebank (Prasad et al., 2008), focusing on their contrastive and concessive occurrences. The corpus-based results show that *but* mostly expressed contrast while *although* is mostly found in concession, with an interaction with sentence position such that sentence-medial *although* is more balanced across contrast and concession than sentence-initial uses (mainly concessive). Following

a Bayesian probabilistic account of discourse interpretation, they then test whether the corpus-based meaning distribution of the connectives is reflected in online and offline measures of production (continuation task), comprehension (coherence judgments) and processing (eye-tracking). All the experiments converge and support the corpus data in showing a strong preference for contrastive uses of *but*, while *although* is more balanced across the two meanings with no clear preference: 50/50 continuation types in the elicited task, more varied judgment scores, no reading time difference across relations. The authors conclude that comprehenders are sensitive to distributional differences between semantically overlapping connectives through a process of entrenchment (i.e., repeated exposure to a form-function pair makes it more accessible). They stress the importance of "a carefully controlled experimental design" (Asr & Demberg, 2020: 396) to access fine-grained discourse effects.

40   This ideal situation of convergence between methods is also found to a certain extent in Zufferey's (2012) study of the French causal connectives *car*, *parce que* and *puisque* (which can all be translated by English *because*). As mentioned above, many studies have attempted to account for the fine-grained differences between the three connectives, often in terms of domains of use: very roughly, *parce que* is often seen as specialized in content relations, while *car* and *puisque* are mostly used in speech-act and epistemic relations, a division of labor which is partially supported by syntactic tests. However, corpus-based studies have observed cases where *parce que* and *car* are interchangeable, and cases where *car* and *puisque* are not, which led Zufferey to address the question once more using a combination of methods. She first compared the three connectives in speech (phone conversations) and writing (descriptive and argumentative texts). She found that, in the written corpus, *parce que* and *car* had a reverse distribution of domains, with a majority of content relations for the former and a majority of pragmatic (i.e., epistemic and speech-act) relations for the former, while *puisque* was exclusively pragmatic. In speech, however, *parce que* seems to have replaced *car* entirely and was therefore much more balanced across the three domains. Zufferey further observed that *puisque* was often accompanied by mentions of an external source or previous context, which she interprets as an "echoic" use of *puisque*.

41   The traditional distinction of domains of use (as described in the Lambda-l Group, 1975) was thus confirmed by the corpus data, although the division is much more permeable in speech. In order to find out the conditions which favor one connective over the other two, Zufferey (2012) then conducted an acceptability judgment task and an elicitation (connective insertion) task since, "by its very nature, a corpus study cannot give specific indications regarding the kinds of factors that make the use of a connective acceptable or not in a given domain" (Zufferey, 2012: 146). Items were created to represent the three domains of use, with some stimuli focusing on the *parce que* vs. *car* pair and others on the echoic nature of *puisque*. The results showed that *car* was largely preferred in the epistemic domain, that *puisque* was exclusively used in echoic contexts, whereas no consensus emerged for

parce que, which thus appears more versatile than the others. The patterns observed in the acceptability task were clearer than those of the elicitation task, which points to a partial non-convergence between the methods. In particular, attitudes towards *parce que* seem more robust than patterns of spontaneous use: "even if they hesitate about its acceptability, subjects tend to use naturally *parce que* in the epistemic domain" (Zufferey, 2012: 150). Zufferey also mentioned the difficulty of representing spoken language in writing and observed no significant differences between the written stimuli and the more casual, spoken-like contexts. Overall, the fuzziness between *car* and *parce que* in production was confirmed experimentally, and the distinct status of *puisque* was convincingly connected to the concept of echo.

42   The third and last study in this literature review where experiments strive to confirm corpus-based findings is that by Mak et al. (2013) on the impact of the semantics of the connective on referent change vs. referent continuity. The authors focus on the Russian conjunctions *i* ("and"), which favors referential continuity, and *a* ("and/but"), which favors discontinuity. They want to test whether the connectives by themselves can inform the listener as to how the discourse will unfold in terms of the next reference. To do so, they first run a crosslinguistic corpus study comparing the Russian connectives with their Dutch equivalents *en* ("and") and *maar* ("but"), for which there is no referential preference. The data show that both Russian *i* and Dutch *en* tend to occur in contexts of referential continuity, more than *a* and *maar*: against their expectations, the theoretical semantic differences between languages were not reflected in the distribution patterns.

43   Mak et al. (2013) therefore expect this similarity between languages to be reflected in processing time, as measured by an eye-tracking visual world para-digm. Participants had to look at pictures representing animals, one of which was mentioned in the first clause. The results indicate that the proportion of looks towards a different referent was higher after Russian *a* than after Dutch *maar*, and that participants looked more at the other referent after Dutch *en* than after Russian *i*. These findings are consistent with the semantic account of the connectives (referential preference) and diverge from the corpus frequencies. Contrary to Asr and Demberg's (2020) study reviewed above, corpus distribution was not a reliable predictor of processing, and the authors conclude that "frequency distributions alone do not provide a parser with sufficient cues about the probability of a reference maintenance or shift" and that "more attention should be given to factors beyond frequency of use" (Mak et al., 2013: 574).

44   The three studies reviewed in this section thus showed cases of full convergence, partial convergence and no convergence between corpus-based and experimental results, in situations where the experiment(s) were used in a second step to test hypotheses extracted from corpus data. The extent of convergence should not be taken as a scale of methodological rigor or validity, but rather illustrates the different scenarios that mixed-method discourse studies can face, and in particular point to the possible mismatch between production, attitudes and processing.

3.3.   Corpus-based studies to further explore experimental effects

45   In this final section, I will turn to two discourse studies that first report experimental findings followed by a corpus study that aims at further exploring the effects found in the first part. The order of presentation of the methods is thus reversed compared to the previous section. The first study by Evers-Vermeul and Sanders (2011) combines a production experiment on children with a corpus-based study on CHILDES (Child Language Data Exchange System), the multilingual database for child language data (MacWhinney, 2000). The study targets the age of acquisition of the different domains of use of causal connectives in Dutch. The authors first test whether four-year-olds produce epistemic relations, in an argumentative and directive task that maximizes the chance of producing this domain of use. They indeed found occurrences of epistemic uses of causal connectives, with an effect of genre (more frequent in argumentative than directive tasks). This first production experiment was complemented by a second one where they added a descriptive task in order to trigger uses in the content domain. In this data, they found epistemic relations in the speech of children as young as three years old, although in smaller amounts than the other two domains.

46   Since the experimental data do not give access to the order of acquisition between content, epistemic and speech-act relations, Evers-Vermeul and Sanders then conducted a longitudinal corpus study of children from the age of two. The authors found that some children produce content uses first, others speech-act first, so that it is impossible to conclude which of these two domains is acquired first (epistemic is never first). They conclude that content and speech-act relations have a lot in common and explain their early occurrence by the fact that they are both tied to the speaker "in the here and now" (Evers-Vermeul & Sanders, 2011: 1659). Furthermore, Evers-Vermeul and Sanders express a particular interest for methodological issues and the difference between methods and suggest that "corpus-based data show us children's earlier spontaneous use and enable us to track longitudinal developments; experiments enable us to control for context effects" (Evers-Vermeul & Sanders, 2011: 1660). Thus, although the production experiments reported in this study do not differ widely from regular corpus collection methods with children, their findings once again vouch for the triangulation of evidence to overcome the limitations of each method.

47   Moving away from production, Zufferey and Gygax (2016) present the combination of a self-paced reading task and a translation study. The two methods target the same object of study, namely perspective shift as a type of discourse discontinuity, but from two parallel vantage points: whether perspective shifts affect online processing and whether the explicitation bias is universal. To do this, the authors focus on the causal vs. confirmation uses of the French connective *en effet* ("indeed"): the former is classified as continuous in the sense that both segments of the relation share the same perspective, while the latter is discontinuous and introduces a shift of perspective (by confirming, the speaker adds credit to an external belief). Zufferey and Gygax thus

expect causal relations to be processed faster than confirmation relations, especially in implicit conditions (without *en effet*). They indeed observed a late interaction that supports this hypothesis (implicit perspective shifts are difficult to process).

48   In the second part of the paper, the authors test whether the difference between continuous and discontinuous relations is stable across languages, as it is based on universal cognitive biases. They run a study on the translations of *en effet* in English, German and Spanish, and expect to find fewer cases of implicitations (i.e., an explicit connective is removed from the translation in the target language, so the relation becomes implicit; this is a frequent phenomenon in translation) for the confirmation (discontinuous) uses. Their hypothesis was confirmed in English and German and, to a smaller extent, in Spanish, probably because of the direct translation equivalent in the former (*en efecto*) against no equivalent in English (the most frequent option is zero translation, i.e., implicit). Zufferey and Gygax conclude that "the explicit or implicit communication of discourse relations depends on cognitive principles that are coherent across languages" (Zufferey & Gygax, 2016: 548). As mentioned above, in this study, one method does not validate the other but the combination offers two relatively different elements of answer that support the same theoretical position, namely that perspective shift bears the same effects as other types of discourse continuity.

4.   Conclusions and recommendations

49   In this structured review of mixed-method discourse studies, I first identified and discussed three main challenges that the combination of corpus-based and experimental data faces, including some theoretical issues beyond the domain of discourse analysis. While methodological triangulation is quite frequent and highly beneficial for the researchers, it is not very frequent in the study of discourse relations and connectives, and I argue that this is because of the specific obstacles that emerge regarding interpretation control, the complex role of context and the fine-grained functional differences that are often observed in corpus data. In the second part, I reviewed several studies that overcame these obstacles and used corpus data either as a source of experimental stimuli, a source of hypotheses or in a more complementary manner for further explorations. While these studies show that mixed-method approaches do not always fully converge or are sometimes complex to set up, they also illustrate the scientific advances that can be reached that way, covering both production and comprehension, with a compromise between naturalness and control.

50   The discussion of challenges and the selective review together suggest a list of recommendations that I hope will help with the design of future mixed-method studies in linguistics and in discourse analysis in particular. The first one is that planning such a study requires an interdisciplinary team with different skills and resources (corpus extraction, annotation, experimental design, programming, statistical analysis): I would suggest at least one person who will act as the experimenter (and will thus be aware of the experimental conditions) and ideally two more for the

analysis (annotation) of the data, in order to avoid any bias from the conditions and to vouch for the reliability of the analysis (inter-annotator agreement). Secondly, experimental tasks should ideally involve some comprehension check in order to ensure that participants make the intended interpretation. This precaution is particularly necessary at the discourse level, since connectives and discourse markers are highly polyfunctional. This is especially true for any study involving the underspecified conjunction *and* and its equivalents in other languages (cf. Crible & Demberg, 2020; Didirková et al., 2019). In addition to finding the right task, the study must also include the right variables, that is, those that can be operationalized for both the corpus analysis and the experimental design (cf. the difficulty of manipulating text genre in Crible & Demberg, 2020). Because corpus-based studies often identify a complex interplay of variables, it might be useful to "decompose" a single corpus study into several experiments in order to cover the full range of variables and their interaction. As for the stimuli, I have emphasized the importance of using strong contrasts between conditions and the attention to controling the role of context (although the use of naturalistic examples can also be an option for some designs).

51   In sum, the perfect mixed-method study uses the right task, the right design and the right stimuli. This can only be achieved if both methods are planned upfront, while keeping each other's affordances in mind. Despite the indubitable constraints that this implies, I strongly believe that corpus and experimental data form a winning combination that must be encouraged. I strongly suspect that there are more studies than the few that I have reviewed here, and that some of them were never published (at least not in a single paper) because the results failed to converge (but see Mak et al., 2013). I hope to have illustrated that non-convergence is not necessarily a sign of poor study design and that there are good reasons for two datasets not to converge. Add to this the difficulty of finding suitable journals that are interested in mixed-method discourse studies, and only then will we grasp the full situation that perhaps sets discourse analysis apart from other fields of linguistics (cf. Gilquin & Gries, 2009) when it comes to the combination of corpus-based and experimental methods. If the papers in the present journal issue are of any indication, there is certainly a momentum for more research in this direction.

References

Andersson, M. 2019. Subjectivity of English Connectives: A Corpus and Experimental Investigation of Result Forward Causality Signals in Written Language. In Ó. Loureda, I. Recio Fernández, L. Nadal & A. Cruz (eds.), *Empirical Studies of the Construction of Discourse*. Amsterdam – Philadelphia: J. Benjamins: 299-318.

Andersson, M. & Spenader, J. 2014. Result and Purpose Relations with and without "So". *Lingua* 148: 1-27.

Arppe, A. & Järvikivi, J. 2007. Every Method Counts: Combining Corpus-Based and Experimental Evidence in the Study of Synonymy. *Corpus Linguistics and Linguistic Theory* 3 (2): 131-159.

ASR, F.T. & DEMBERG, V. 2015. Uniform Information Density at the Level of Discourse Relations: Negation Markers and Discourse Connective Omission. In M. PURVER, M. SADRZADEH & M. STONE (eds.), *Proceedings of the 11th International Conference on Computational Semantics (15-17 April 2015, London, UK)*. Stroudsburg: Association for Computational Linguistics (ACL): 118-128. Available online: https://aclanthology.org/W15-0117.pdf.

ASR, F.T. & DEMBERG, V. 2020. Interpretation of Discourse Connectives Is Probabilistic: Evidence from the Study of *But* and *Although*. *Discourse Processes* 57 (4): 376-399.

BRINTON, L. 1996. *Pragmatic Markers in English. Grammaticalization and Discourse Functions.* Berlin – New York: De Gruyter.

CRIBLE, L. 2018. *Discourse Markers and (Dis)fluency: Forms and Functions across Languages and Registers.* Amsterdam – Philadelphia: J. Benjamins.

CRIBLE, L. 2021. Negation Cancels Discourse-Level Processing Differences: Evidence from Reading Times in Concession and Result Relations. *Journal of Psycholinguistic Research* 50 (6): 1283-1308.

CRIBLE, L. & DEGAND, L. 2017. Reliability vs. Granularity in Discourse Annotation: What Is the Trade-off? *Corpus Linguistics and Linguistic Theory* 15 (1): 71-99.

CRIBLE, L. & DEGAND, L. 2019. Domains and Functions: A Two-Dimensional Account of Discourse Markers. *Discours* 24: 1-35. Available online: https://journals.openedition.org/discours/9997.

CRIBLE, L. & DEMBERG, V. 2020. When Do We Leave Discourse Relations Underspecified? The Effect of Formality and Relation Type. *Discours* 26: 1-25. Available online: https://journals.openedition.org/discours/10848.

CRIBLE, L. & PICKERING, M.J. 2020. Compensating for Processing Difficulty in Discourse: Effect of Parallelism in Contrastive Relations. *Discourse Processes* 57 (10): 862-879.

CUENCA, M.J. 2013. The Fuzzy Boundaries between Discourse Marking and Modal Marking. In L. DEGAND, B. CORNILLIE & P. PIETRANDREA (eds.), *Discourse Markers and Modal Particles. Categorization and Description.* Amsterdam – Philadelphia: J. Benjamins: 191-216.

DAS, D. & TABOADA, M. 2018. Signalling of Coherence Relations in Discourse, beyond Discourse Markers. *Discourse Processes* 55 (8): 743-770.

DEGAND, L. 2014. "So Very Fast Then" Discourse Markers at Left and Right Periphery in Spoken French. In K. BEECHING & U. DETGES (eds.), *Discourse Functions at the Left and Right Periphery: Crosslinguistic Investigations of Language Use and Language Change.* Leiden: Brill: 151-178.

DEGAND, L. & PANDER MAAT, H. 2003. A Contrastive Study of Dutch and French Causal Connectives on the Speaker Involvement Scale. In A. VERHAGEN & J. VAN DE WEIJER (eds.), *Usage-Based Approaches to Dutch. Lexicon, Grammar, Discourse.* Utrecht: LOT: 175-199. Available online: https://dspace.library.uu.nl/bitstream/handle/1874/295390/bookpart.pdf?sequence=2&isAllowed=y.

DEGAND, L. & SANDERS, T.J.M. 2002. The Impact of Relational Markers on Expository Text Comprehension in L1 and L2. *Reading and Writing* 15 (7-8): 739-757.

DIDIRKOVÁ, I., CHRISTODOULIDES, G. & SIMON, A.C. 2018. The Prosody of Discourse Markers *Alors* and *Et* in French. A Speech Production Study. In K. KLESSA, J. BACHAN, A. WAGNER, M. KARPIŃSKI & D. ŚLEDZIŃSKI (eds.), *Proceedings of the 9th International Conference on Speech Prosody 2018 (13-16 June 2018, Poznań, Poland)*. ISCA (International Speech Communication Association) Online Archive: 503-507. Available online: https://www.isca-speech.org/archive_v0/SpeechProsody_2018/pdfs/153.pdf.

DIDIRKOVÁ, I., CRIBLE, L. & SIMON, A.C. 2019. Impact of Prosody on the Perception and Interpretation of Discourse Relations: Studies on "Et" and "Alors" in Spoken French. *Discourse Processes* 56 (8): 619-642.

DÖRING, S. & REPP, S. 2020. The Modal Particles *Ja* and *Doch* and Their Interaction with Discourse Structure: Corpus and Experimental Evidence. In S. FEATHERSTON, R. HÖRNIG, S. VON WIETERSHEIM & S. WINKLER (eds.), *Experiments in Focus. Information Structure and Semantic Processing*. Berlin – New York: De Gruyter: 17-56.

EVERS-VERMEUL, J. & SANDERS, T.J.M. 2011. Discovering Domains – On the Acquisition of Causal Connectives. *Journal of Pragmatics* 43 (6): 1645-1662.

FOX TREE, J.E. 2015. Discourse Markers in Writing. *Discourse Studies* 17 (1): 64-82.

GILQUIN, G. & GRIES, S.T. 2009. Corpora and Experimental Methods: A State-of-the-Art Review. *Corpus Linguistics and Linguistic Theory* 5 (1): 1-26.

GOLDSTEIN-STEWART, J., GOODWIN, K.A., SABIN, R.E. & WINDER, R.K. 2008. Creating and Using a Correlated Corpus to Glean Communicative Commonalities. In N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS & D. TAPIAS (eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation – LREC 2008*. Paris: European Language Resources Association (ELRA): 3029-3035. Available online: http://www.lrec-conf.org/proceedings/lrec2008/pdf/771_paper.pdf.

LAMBDA-L GROUP 1975. Car, Parce que, Puisque. *Revue Romane* 10 (2): 248-280. Available online: https://tidsskrift.dk/revue_romane/article/view/29139/25890.

LEVELT, W.J.M. 1989. *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

LUSCHER, J.-M. & MOESCHLER, J. 1990. Approches dérivationnelles et procédurales des opérateurs et connecteurs temporels: les exemples de *et* et de *enfin*. *Cahiers de linguistique française* 11: 77-104. Available online: https://clf.unige.ch/index.php/download_file/view/336/153/.

MACWHINNEY, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah – London: L. Erlbaum [3rd ed.].

MAK, W.M., TRIBUSHININA, E. & ANDREIUSHINA, E. 2013. Semantics of Connectives Guides Referential Expectations in Discourse: An Eye-Tracking Study of Dutch and Russian. *Discourse Processes* 50 (8): 557-576.

MORERA, Y., LEÓN, J.A., ESCUDERO, I. & VEGA, M. DE 2017. Do Causal and Concessive Connectives Guide Emotional Expectancies in Comprehension? A Double-Task Paradigm Using Emotional Icons? *Discourse Processes* 54 (8): 583-598.

PANDER MAAT, H. & SANDERS, T.J.M. 2001. Subjectivity in Causal Connectives: An Empirical Study of Language in Use. *Cognitive Linguistics* 12 (3): 247-273.

PICKERING, M.J. & GARROD, S. 2007. Do People Use Language Production to Make Predictions during Comprehension? *Trends in Cognitive Sciences* 11 (3): 105-110.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B. 2008. The Penn Discourse TreeBank 2.0. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation – LREC 2008*. Paris: European Language Resources Association (ELRA): 2961-2968. Available online: http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

Sanders, T.J.M. 1997. Semantic and Pragmatic Sources of Coherence: On the Categorization of Coherence Relations in Context. *Discourse Processes* 24 (1): 119-147.

Sanders, T.J.M., Spooren, W.P.M. & Noordman, L.G.M. 1992. Toward a Taxonomy of Coherence Relations. *Discourse Processes* 15 (1): 1-35.

Schumann, J., Zufferey, S. & Oswald, S. 2021. The Linguistic Formulation of Fallacies Matters: The Case of Causal Connectives. *Argumentation* 35 (3): 361-388.

Simon, A.C. & Degand, L. 2007. Connecteurs de causalité, implication du locuteur et profils prosodiques: le cas de *car* et de *parce que*. *Journal of French Language Studies* 17 (3): 323-341.

Spooren, W.P.M. & Degand, L. 2010. Coding Coherence Relations: Reliability and Validity. *Corpus Linguistics and Linguistic Theory* 6 (2): 241-266.

Stukker, N., Sanders, T.J.M. & Verhagen, A. 2008. Causality in Verbs and in Discourse Connectives: Converging Evidence of Cross-Level Parallels in Dutch Linguistic Categorization. *Journal of Pragmatics* 40 (7): 1296-1322.

Sweetser, E. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge – New York – Melbourne: Cambridge University Press.

Yung, F., Scholman, M.C.J. & Demberg, V. 2019. Crowdsourcing Discourse Relation Annotations by a Two-Step Connective Insertion Task. In A. Friedrich, D. Zeyrek & J. Hoek (eds.), *Proceedings of the 13th Linguistic Annotation Workshop (Florence, Italy, 1 August 2019)*. Stroudsburg: Association for Computational Linguistics (ACL): 16-25. Available online: https://aclanthology.org/W19-4003.pdf.

Zufferey, S. 2012. "*Car, Parce que, Puisque*" Revisited: Three Empirical Studies on French Causal Connectives. *Journal of Pragmatics* 44 (2): 138-153.

Zufferey, S. & Gygax, P.M. 2016. The Role of Perspective Shifts for Processing and Translating Discourse Relations. *Discourse Processes* 53 (7): 532-555.