HAL open science

LEXCONN: a French Lexicon of Discourse Connectives

Charlotte Roze, Danlos Laurence, Philippe Muller

To cite this version:

Charlotte Roze, Danlos Laurence, Philippe Muller. LEXCONN: a French Lexicon of Discourse Connectives. MAD 2010 - 8th Workshop Multidisciplinary Approaches to Discourse, Mar 2010, Moissac, France. pp.114-125. inria-00511615

HAL Id: inria-00511615

https://inria.hal.science/inria-00511615

Submitted on 18 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LEXCONN: a French Lexicon of Discourse Connectives

Charlotte Roze[1] Laurence Danlos[1] Philippe Muller[2]
(1) Université Paris 7, Alpage
(2) Université Toulouse, IRIT & INRIA, Alpage
charlotte.roze@linguist.jussieu.fr, laurence.danlos@linguist.jussieu.fr,
muller@irit.fr

Abstract. With respect to discourse organisation, the most basic way of signalling the speaker's or writer's intentions is to use explicit lexical markers: so-called discourse markers or discourse connectives. While a lexicon of discourse connectives associated with the relations they express can be very useful for researchers, especially in Natural Language Processing, few projects aim at collecting them exhaustively, and only in a small number of languages.

We present LEXCONN, a French lexicon of 328 discourse connectives, collected with their syntactic categories and the discourse relations they convey, and the methodology followed to build this resource. The lexicon has been constructed manually, applying systematic connective and relation identification criteria, using the Frantext corpus as empirical support. Each connective has been associated to a relation within the framework of Segmented Discourse Representation Theory. We make a case for a few refinements in the theory, based on cases where no existing relation seemed to match a connective's usage.

Keywords. discourse connectives, discourse relations, lexicon, ambiguity

1 Introduction

With respect to discourse organisation, the most basic way of signalling the speaker's or writer's intentions is to use explicit lexical markers: so-called discourse markers or discourse connectives. Used to express functional relations between parts of discourse, these items can be used at the sentential level or at the level of larger textual units.

We will focus here on the basic inter-sentential level: what is expressed as a whole by two sentences in a coherent discourse. This can be recursively extended to cover an entire discourse when the same relations are applied to sets of sentences. Discourse connectives explicitly signal the presence of a discourse relation between two discourse units and more generally, they contribute to discourse coherence and mark discourse structure, at least the basic organisation mentioned in Spooren and Sanders (2008): causality, sequence, grouping, contrast.

From the reader's point of view they help to disambiguate discourses whose interpretations would be vaguer without them. For example, in (1a), two interpretations are possible:[1] either Peter can find his own way home because he is not stupid (relation *Result*), or the fact that Peter can find his own way home proves he is not stupid (relation *Evidence*). We can see in (1b) and (1c) that the connectives (which are italicized) forces one of the two interpretations.

[1]This example comes from (Wilson and Sperber, 1993).

(1) Peter is not stupid.
  a. He can find his own way home.
  b. *So* he can find is own way home.
  c. *After all*, he can find is own way home.

A lexicon of discourse connectives associated with the relations they express can be very useful for researchers in Natural Langage Processing, who aim at produce automatic discourse analysis for French. Connectives can help to select the right relation between two discourse units, as they do for speakers. Very few studies or projects aim at collecting them exhaustively, and only in a small number of languages. We will detail the gathering of such a resource for French, LEXCONN,[2] and the methodology followed. The set of functional and rhetorical relations targeted by this study is taken *a priori* from Segmented Discourse Representation Theory (Asher and Lascarides, 2003), and we will evaluate how good a fit the theory is with respect to the set of connectives under investigation.

In LEXCONN we list 328 discourse connectives, collected with their syntactic categories and the discourse relations they express. Such a resource already exists for English (Knott, 1996), Spanish (Alonso et al., 2002) and German (Stede and Umbach, 1998), but LEXCONN is the first one for French. The lexicon aims at being exhaustive. It has been constructed manually, applying systematic connective identification criteria, associating a SDRT relation, and the type (coordinating or subordinating) of this relation with each connective. We used the FRANTEXT[3] corpus as a source of examples.

The rest of the paper is organised as follows. In Section 2, we present the theoretical background of this work (SDRT) and introduce the terminology we adopt about discourse connectives. In Section 3, we detail the methodology for building the lexicon and present syntactic, semantic and discursive criteria for identification of connectives. In Section 4, we describe the second stage of our work: associating discourse relations with discourse connectives. In Section 5, we present some problematic cases for SDRT when trying to associate relations with connectives.

2 Preliminaries

Our work is in line with SDRT (Asher and Lascarides, 2003), who inherits from the Discourse Representation Theory or DRT (Kamp, 1981) and discourse analysis (Grosz and Sidner, 1986; Mann and Thompson, 1988). SDRT aims at representing discourse coherence and discourse structure. The construction of SDRS (Segmented Discourse Structures) mainly rests on the distinction between coordinating relations (like *Narration* and *Result*) and subordinating relations (like *Elaboration* and *Explanation*). This distinction allows for the definition of some important principles of the theory, such as the *Right Frontier Constraint* (RFC). According to this constraint, in the course of building an SDRS, the only available sites for attachement of new information are the last segment of the discourse context and the segments which structurally dominate it.

Following Danlos (2009), we use the following terminology. The clause where a connective appears is called its "host clause". A discourse connective/relation has two arguments which

[2]The data base is available at www.linguist.univ-paris-diderot.fr/~croze/.
[3]FRANTEXT is a textual base of French litterature. It is available at www.frantext.fr.

are the semantic representations of two discourse segments called "host segment" and "mate segment". The host segment of a connective is identical to or starts at its host clause. The mate segment is governed by constraints described in Section 3.1.

3 Building a Lexicon of Connectives

The first step of our methodology was to gather a corpus of discourse connectives candidates (about 600). To do that, we used various corpora of conjunctions of subordination and prepositions given by Eric Laporte and Benoît Sagot, the list of French discourse markers of the ANNODIS project[4] and the translated corpus of English discourse connectives built by Knott (1996).

In the database, we associate a syntactic category with each connective, which can differ a little from traditional ones: conjunction of coordination (`cco`) for connectives like *et* (*and*), *ou* (*or*) and *mais* (*but*), which are always at the beginning of their host clause, and whose mate segment is always on the left; conjunction of subordination (`csu`) for connectives like *parce que* (*because*), *même si* (*even though*) and *tandis que* (*whereas*), which are always at the beginning of their host clause, and whose mate segment can be anteposed, postposed, or internal;[5] preposition (`prep`) for the reduced forms of conjunctions of subordination when the host clause is an infinitive VP, like *afin de* (*in order to*), *pour* (*for*) and *avant de* (*before*);[6] adverb (`adv`) for connectives like *donc* (*so*), *néanmoins* (*nevertheless*) and *en tout cas* (*in any case*), which can appear in various positions in their host clause, and whose mate segment is always on the left.[7]

After gathering a corpus of candidate connectives, we have applied various criteria for the identification of connectives. In Section 3.1, we present some syntactic and semantic criteria we used for identification of connectives, and in Section 3.2, some discursive ones.

3.1 Syntactic and Semantic Criteria

The criteria we present in this Section concern three properties of discourse connectives: they are not integrated to propositional content (cleft criterion), they cannot be referential expressions (substitutability criterion), and their meaning is not compositional (compositionality criterion).

Cleft Criterion Discourse connectives cannot be focused in cleft constructions.

According to Riegel et al. (2004), the items which can be focused in cleft constructions have one of the following functions: subject, object, or adverbial. These items are inside the predicative structure. Jayez and Rossari (1996) distinguish the connectives which are integrated to the predicative structure (and which can be focused in cleft constructions) from the other ones. For example, they claim that *à ce moment-là* in (2a) is a temporal connective which can be focused

[4] ANNODIS is a project of French discourse annotation (Péry-Woodley et al., 2009).

[5] However, for some conjunctions of subordination like *comme*, the mate segment is always anteposed. For others, the mate segment can be anteposed or internal. These informations are marked in LEXCONN.

[6] There exists a few `prep` which are not linked with `csu`, e.g. *quitte à*, *quant à*.

[7] We consider as adverbs some NPs which are not introduced by a preposition, like *la preuve*, *résultat*.

in a cleft construction, see (2b). On the other hand, Bras (2008) claims that *à ce moment-là* in (2a) is not a connective, but a temporal cue: it only temporally locates events, and doesn't play any role at the discourse level. We agree with Bras contra Jayez and Rossari: *à ce moment-là* has a non-discourse usage in (2a), where it refers to the temporal location of an eventuality, while it has a discourse usage in (3a) where it cannot be clefted, see (3b). Moreover, it is referential in (3a) but not so in (3b), which goes along with the next criterion.

(2) *Il a commencé à pleuvoir.* 'It started raining.'
  a. A ce moment-là*, Marie est arrivée.* '*At that moment*, Mary arrived.'
  b. C'est à ce moment-là que *Marie est arrivée.*

(3) *Tu as l'air de penser qu'elle n'est pas honnête.* 'You seem to think she is not honest.'
  a. A ce moment-là*, ne lui raconte rien.* '*So* don't tell her anything.'
  b. # C'est à ce moment-là que *ne lui raconte rien.*

Substitutability Criterion Discourse connectives cannot be substituted by an entity (person, event, discourse unit) of the context.

Knott (1996) considers as discourse connectives some phrases like *because of this*. He keeps phrases which contain propositional anaphora in his corpus, which can be substituted by entities of the discourse context. On the contrary, we don't retain this type of phrases in LEXCONN.

To illustrate the Substitutability Criterion, consider *après ça* in (4b) and *à part ça* in (6b). On the one hand, in (4b), *ça* refers to the segment in (4a), as shown by the acceptability of (5). On the other hand, *ça* in (6b) does not refer to the segment in (6a), as shown by the inacceptability of (7). The Substitutability Criterion tells us that *après ça* is not a connective, while *à part ça* remains in the corpus of candidate connectives.

(4) a. *Bruno est allé en Argentine.* 'Bruno went to Argentina.'
  b. Après ça*, il est parti au Pérou.* '*After that*, he moved to Peru.'

(5) Après *[qu'il est allé en Argentine], Bruno est parti au Pérou.*

(6) a. *Hier soir j'ai croisé Pierre dans une boîte de nuit.* 'Last night I saw Peter in a nightclub.'
  b. A part ça *il nous dit tout le temps qu'il est fatigué.* '*Though* he always says he is tired.'

(7) # A part *[qu'hier soir je l'ai croisé dans une boîte de nuit], Pierre nous dit tout le temps qu'il est fatigué.*

Compositionality Criterion Discourse connectives are invariable.[8]

Various studies (Molinier, 2003; Cojocariu and Rossari, 2008; Nakamura, 2009) aim at showing the connecting role played by adverbials like *à ce propos* and *la preuve*, which contain (predicative) nouns. It seems that the emergence of a discursive role for these adverbials is correlated with a process of fixation. For example, the determiners and the numbers of *la preuve* and *à*

[8]Connectives cannot undergo internal modification, but some of them can be externally modified by adverbials, such as *probablement* or *certainement* for *parce que*.

ce propos (in their discourse usages) have become invariable (# *les preuves*, # *à ces propos*). These studies inspired our Compositionality Criterion: nouns contained in connectives cannot be modified by an adjective, their numbers and their determiners are invariable. This criterion allows us to retain some candidates like *en tout cas* and *résultat*: *en tout cas* in (8a) cannot be modified by an adjective in (8b), and *résultat* in (9a) is invariable, see (9b).

(8) *Je ne sais plus s'il y avait vraiment de la neige, ce Noël-là.* 'I don't know if there really was snow, that Christmas.'

a. En tout cas*, dans mon souvenir, je la vois tomber...*[9] '*In any case*, I remember seeing it falling...'

b. # En tout cas envisagé / possible*, dans mon souvenir, je la vois tomber...*

(9) *Pierre n'a pas réussi à dormir cette nuit.* 'Peter couldn't spleep last night.'

a. Résultat*, il était en retard aujourd'hui.* '*Thus*, he was late today.'

b. # Le résultat / Les résultats*, il était en retard ce matin.*

3.2 Discursive Criteria

The criteria we present in this Section only make use of discourse notions. They were applied after syntactic and semantic criteria, and helped identifying discourse relations conveyed by connectives.

Contextual Criterion If the discourse $D = c\ clause$ is coherent without other discourse context, then c is not a discourse connective.

The Contextual Criterion is the only test Knott (1996) used to build a list of English connectives. This test is insufficient to discard adverbials like *le lendemain* or *un peu plus loin*, which express temporal or spatial information. However we used Knott's test to discard some candidates.

Forced Relation Criterion Let D_a and D_b be coherent discourses with $D_a = seg_1\ seg_2$ and $D_b = seg_1\ c\ seg_2$, R_a the discourse relation which holds between seg_1 and seg_2 in D_a, and R_b the relation which holds in D_b. If $R_a \neq R_b$ then c is a discourse connective.

Consider (10a) and (10b) which differ by the presence of *malheureusement* in (10b). The segment in (10a) is an *Explanation* of the first segment (Mark will camp this summer), whereas the segment in (10b) is in a *Contrast* relation with the first segment (maybe Mark will not camp this summer). This is evidence that *malheureusement* is a connective. On the other hand, consider (10c) and (10d) which differ by the presence of *évidemment* in (10d). The presence of this adverbial doesn't change the discourse relation, which is *Result* in both cases. More generally, we found no example where the presence of this adverb changes the relations involved. This is evidence that *évidemment* is not a connective.

(10) *Marc veut faire du camping cet été.* 'Mark wants to camp this summer.'

a. *Il n'a pas beaucoup d'argent.* 'He does not have much money.'

[9] Patrick Modiano, *Un pedigree*, 2005, p. 94.

b. Malheureusement *il n'a pas beaucoup d'argent.* '*Unfortunately* he does not have much money.'
c. *Il faut qu'il économise de l'argent.* 'He must save up money.'
d. Evidemment*, il faut qu'il économise de l'argent.* '*Of course*, he must save up money.'

Coherence Criterion If seg_1 seg_2 is incoherent and seg_1 c seg_2 is coherent, then c is a discourse connective.

Beaulieu-Masson (2002) gives a study of connectives like *à propos*, *à ce propos* and *au fait*, which force discourse coherence. For example, in (11), the presence of *à propos* helps linking the segment in (11b) to the segment in (11a). Without it, the discourse would be incoherent. The Coherence Criterion is inspired from this study. It can be used for various connectives. For example, *ceci dit* in (12a) is a discourse connective (which mark the relation *Opposition*), because if it is deleted, the discourse becomes incoherent, see (12b).

(11) a. *Boris, Je prends des gouttes pour stimuler mon appétit, mais les résultats sont lents, très lents.* 'Boris, I take drops to stimulate my appetite, but the results are slow, very slow.
b. A propos*, vers quel moment crois-tu que tu pourras venir ?*[10] '*By the way*, when can you come ?'

(12) *Ce serait vraiment utile pour nous d'aller à cette réunion*. 'It would be really useful for us to go to this meeting.'
a. Ceci dit*, on peut s'en passer.* '*But* we can do without it.'
b. # *On peut s'en passer.* 'We can do without it.'

After we applied these criteria, 328 candidates were kept as connectives.[11]

4 Associating Relations with Connectives

After building the list of French discourse connectives, we tried for each connective to determine which discourse relation(s) it expresses, observing the contexts where it appears in discourses from the FRANTEXT corpus. To do this, we used a set of 15 discourse relations defined in SDRT, which are of various kinds: temporal (*Narration*, *Background* (*backward* or *forward*), *Flashback*), causal (*Result*, *Explanation*, *Goal*), structural (*Parallel*, *Contrast*, *Elaboration*, *Continuation*), logical (*Alternation*, *Consequence*), metatalk (*Result**, *Explanation**). Each relation is typed (coordinating or subordinating), and has semantic effects.

4.1 Tests for Relations Identification

In order to identify the discourse relation conveyed by a connective, we tried to use the following clues.

[10] Lydia Flem, *Lettres d'amour en héritage*, 2006, p. 127.
[11] The list of discourse markers from ANNODIS project contains about 60 connectives.

Attachment Test This test helps to determine the type of the relation (Asher and Vieu, 2005). As we said in Section 2, in SDRT, relations are either coordinating or subordinating. This distinction is essential for the RFC: if the relation between two discourse segments (π_1) and (π_2) is subordinating, a third segment (π_3) can be attached to (π_1), whereas if it is coordinating, (π_3) cannot be attached to (π_1), because (π_1) is no longer available for attachment. We used this test to identify the type of relation expressed by connectives.

Substitution Test If two connectives are substituable for each other in most of the discourse contexts they appear in, e.g. the discourse interpretation is unchanged, they probably express the same discourse relation. This test is inspired from Knott (1996). However, given that our goal is not to build a taxonomy of connectives/discourse relations we did not use more subtle relationships than contingent substituability (such as synonymy, hyponymy or hyperonymy).

For example, the Substitution Test tells us that *dès lors que*, *puisque* and *étant donné que* have one discourse usage in common: in (13), they are substituable for each other without changing the discourse interpretation (they express $Explanation*$).

(13) *Brillant résultat de quinze ans de diplomatie gaulliste, mais résultat inévitable,* dès lors que / puisque / étant donné que *nous avons toujours placé (...) les apparences au-dessus des réalités* ...[12] 'This is the brilliant outcome of fifteen years of Gaullist diplomacy, but this is inevitable, *given that* we always preferred appearances to reality.'

Semantics Effects In SDRT, discourse relations have semantic effects. Some relations (such as $Background$, $Explanation$ and $Flashback$) set temporal constraints on the eventualities they link. For example, $Flashback(\alpha, \beta)$ implies a temporal precedence between e_α and e_β.[13] Relations such as $Result$ and $Explanation$ can also establish causal relationships between eventualities. For instance, $Result(\alpha, \beta)$ implies a causal link between e_α and e_β.

4.2 Ambiguity

The database contains 328 connectives, and 428 usages of connectives: connectives are ambiguous. We describe here two types of ambiguity.

Some connectives can establish more than one discourse relation. For instance, *si* has a conditional usage (see (14)), in which its mate segment can be anteposed, postposed or internal. It also has a concessive usage (see (15)), in which its mate segment can only be anteposed. In the same way, the adverb *aussi* expresses $Result$ when it is in initial position of its host clause and $Parallel$ when it is not in initial position.

(14) Si *je ne reçois pas très vite de l'aide, nous courons au désastre.*[14] 'If nobody comes to my help very soon, we're doomed.'

(15) *Quand j'étais un jeune garçon, j'ai manié indéfiniment les vieux fascicules de cette revue.* Si *j'étais trop jeune pour les bien comprendre, j'en recevais toutes sortes de rêves...*[15] 'When I was a boy, I handled old issues of this magazine endlessly. If I was too young to understand them, I drew all kinds of dreams from them...'

[12] Pierre Mendès-France, *Oeuvres complètes. 6. Une vision du monde.*, 1974-1982, 1990, p. 133

[13] e_α and e_β are the eventualities described in the segments α and β.

[14] Patrick Rambaud, *La Bataille*, 1997, p. 228, CHAPITRE V, Seconde journée.

[15] Inspired from: Maurice Barrès, *Mes Cahiers - Tome 14 : 1922-1923*, 1923, p. 163, 46ème cahier.

In LEXCONN, such informations about the position of the mate segment of subordinating conjunctions and the position of adverbs in their host clause are encoded by specific attributes/features (`position-sub` and `position-adv`). However, for many ambiguous connectives, the usage cannot be selected by surface clues like the connective's position or the mate segment's position and depends more on discourse content.

Some other connectives such as *et* (*and*) present a second type of ambiguity : they have discourse and non-discourse usages. These non-discourse usages are frequent for adverbials and are not represented in LEXCONN. However we kept in the lexicon non-discourse usages for connectives like *à ce moment-là* ($Result*$) and *en même temps* ($Opposition$), which often express strictly temporal relations (e.g. temporal simultaneity).

We now give quantitative data about ambiguous connectives:[16] 73 connectives (23,7%) have more than one discourse usage and 14 connectives (4,2%) have discourse and temporal usages. Concerning ambiguity between discourse usages, two cases must be distinguished: the case where a connective establish discourse relations of the same type (coordinating or subordinating) and the case where a connective establish relations with different types. The first case seems less problematic than the second in an NLP perspective, because it doesn't implies structural ambiguity. Only 6,2% from the total number of connectives are in the second case.

4.3 Relations Frequency

We cannot yet know the frequency of each discourse connective in terms of occurrences in a real corpus (this work has to be done using LEXCONN and the ANNODIS corpus), but we now can give the frequency of each discourse relation in terms of number of connectives. These frequencies are given in Table 1.[17] Some of the relations are defined in SDRT and listed above, but some of them are not and are detailed in Section 5.

About 28% connectives are "contrastive" ones, e.g. they express either *Contrast* (formal contrast) or *Opposition* (violation of expectation) or *Concession* (these relations are grouped together in ANNODIS corpus). What we can say is that there are many ways of expressing contrastive relations, maybe because they are difficult to express without a discourse connective. On the contrary, *Elaboration* has a low frequency in terms of connectives.

5 Problematic Cases for SDRT

This stage led to the following result about discourse relations: some discourse connectives appear in contexts where no relation defined in SDRT can hold. In other words, although this work is in line with SDRT, the set of discourse relations defined in the theory is insufficient for describing the contributions to discourse interpretation of all French discourse connectives. Two cases must be distinguished. First, the case where we can introduce relations that are not defined in SDRT. These relations are generally defined in Rhetorical Structure Theory (Mann and Thompson, 1988). Second, the case for which it seems impossible to associate any relation to a discourse connective.

[16] We do not consider connectives marked as "unknown" in the counts.

[17] Notice that we distinguish several usages for some connectives.

Relation	Number	Percentage
Opposition	41	9,5
Result	35	8,1
Concession	32	7,4
Continuation	32	7,4
Explanation	28	6,5
Goal	25	5,8
Condition	25	5,8
Explanation∗	24	5,6
Narration	23	5,4
Unknown	21	4,9
Contrast	17	4,0
$Background_b$	15	3,5
$Temporal_{location}$	14	3,3

Relation	Number	Percentage
Parallel	13	3,0
Elaboration	11	2,6
Result∗	11	2,6
Summary	11	2,6
Flashback	10	2,4
Detachment	9	2,1
Alternation	9	2,1
Consequence	7	1,6
$Background_f$	7	1,6
Evidence	7	1,6
Rephrasing	6	1,4
Digression	6	1,4
Total	428	100%

Table 1: Relations frequencies: number and percentage of connectives.

5.1 Introducing New Relations in SDRT

We introduced six relations in LEXCONN which are not defined in SDRT. These relations are: *Concession* (même si, bien que), *Opposition* (cependant, malgré tout), *Summary* (en gros, globalement), *Detachment* (quoi qu'il en soit, de toute manière), *Digression* (à propos, au fait), and *Rephrasing* (enfin, tout au moins). These relations were introduced because no relation defined in SDRT can represent the contributions to discourse interpretation of some connectives, which can be grouped together with respect to the contexts where they appear.

For example, connectives like *bien que* or *même si* are considered in ANNODIS as markers of the coordinating relation *Contrast*. However, they express a subordinating relation, as shown in (16): the segments (π_1) and (π_3) are linked by the relation *Result*, therefore the relation between (π_1) and (π_2) is necessarly subordinating. The discourse structure associated with (16) is shown in Figure 1.

(16) a. *Pierre m'a aidé à repeindre la chambre* 'Peter helped me repaint the bedroom' (π_1)

b. bien qu'*il ait beaucoup de boulot en ce moment*. '*even though* he has a lot of work at this time.' (π_2)

c. Du coup*, c'est déjà terminé !* '*Thus* it is already over.' (π_3)

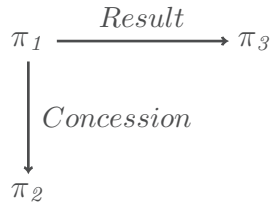

Figure 1: Graph Representation for (16)

In addition, these connectives link segments which don't have necessarly similar semantic structures, while *Contrast* must link segments with some structural similarities. In conclusion, *bien*

que and *même si* cannot express the coordinating relation *Contrast*: they express the subordinating relation *Concession* (which is defined in RST) that we introduced in LEXCONN.

5.2 Unknown Relations

For 21 connectives (about 6%), the associated discourse relation in LEXCONN is `unknown`. Among these connectives, there are adverbs (*en fait*, *au moins*), conjunctions of subordination (*avant même que*, *à mesure que*), and prepositions (*quant à*, *quitte à*). Each connective associated with `unknown` verifies the criteria we presented in Section 3, but any possible relation is insufficient for describing the semantics of the connective.

For example, *à mesure que*, whose meaning is non-compositional, as shown by the inacceptability of (17b), and which doesn't contain a referential expression, as shown in (17c), is a connective. However, whatever relation we try to associate with it (*Simultaneity*, *Explanation*, or even *Parallel*), some semantic information is lost, i.e. the fact that there is a simultaneous temporal progression between the two events involved. As a consequence, *à mesure que* is associated with `unknown`.

(17) *Tes digressions s'allongeaient* 'Your digressions got longer and longer'

a. à mesure que *tu finissais les alcools de ta mère.*[18] '*as and when* you finished your mother's alcohols.'

b. # à la mesure que *tu finissais les alcools de ta mère*.

c. # à cette mesure-là.

6 Conclusion

Building a French lexicon of discourse connectives brought several results. It involved a systematic methodology to identify discourse connectives and associate discourse relations to them, resting on various studies about connectives and corpus-collected examples. In addition, it shows which connectives remain to be studied in detail (especially connectives whose function is "unknown" so far). A statistical analysis of the resulting lexicon allowed us to quantify several things, such as the importance of the various discourse relations in terms of the number of connectives associated with them, and a count of ambiguous connectives.

Despite these results, LEXCONN has to be improved: some information has to be added. For example, some information about ambiguity between discourse and non-discourse usage has to be introduced. This improvement will be possible with other linguistic analysis, but also with automatic analysis on ANNODIS corpus: we could examine the link between position in the host clause and discursive/non-discursive role for adverbials.

However, LEXCONN already constitute a precious resource for NLP. It might help for discourse markers annotation in ANNODIS, in which connectives are not yet marked. A statistical analysis of the connectives on corpus can also be useful, for example concerning connective's frequency. Such analysis could help answering the following question: are ambiguous connectives the most frequent ones?

[18] Edouard Levé, *Suicide*, 2008, p. 29.

References

Laura Alonso, Irene Castellón, and Lluís Padró. Lexicón computacional de marcadores del discurso. *SEPLN, XVIII Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 2002.

Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

Nicholas Asher and Laure Vieu. Subordinating and coordinating discourse relations. *Lingua, Elsevier*, 115(4):591–610, 2005.

Anne Beaulieu-Masson. Quels marqueurs pour parasiter le discours ? *Cahiers de Linguistique Française*, 24:45–71, 2002.

Myriam Bras. *Entre relations temporelles et relations de discours*. Dossier d'HDR, Université de Toulouse le Mirail, 2008.

Corina Cojocariu and Corinne Rossari. Constructions of the type *la cause/la raison/la preuve* + utterance: grammaticalization, pragmaticalization, or something else? *Journal of pragmatics*, 40:1435–1454, 2008.

Laurence Danlos. D-STAG: a formalism for discourse analysis based on SDRT and using synchronous TAG. In *Proceedings of the 14th Conference on Formal Grammar (FG'09)*, pages 1–20, 2009.

Barbara Grosz and Candace Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

Jacques Jayez and Corinne Rossari. *Donc* et les consécutifs, des systèmes de contraintes différentiels. *Linguisticæ Investigationes*, XX:117–143, 1996.

Hans Kamp. Evénements, représentations discursives et référence temporelle. *Langages*, 64: 34–64, 1981.

Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh, 1996.

William Mann and Sandra Thompson. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8:243–281, 1988.

Christian Molinier. Connecteurs et marqueurs énonciatifs : Les compléments figés formés à partir du nom *propos*. In *Actes du Colloque Grammaires et Lexiques Comparés*, volume 26, pages 15–31. Conenna, Mirella and Laporte, Éric, 2003.

Takuya Nakamura. Observations sur la prédication : prédicat verbal, prédicat nominal avec verbe support et prédicat nominal sans verbe support. In *Actes du Colloque International Supports et prédicats non verbaux dans les langues du monde*, Paris, France, 2009.

Marie-Paule Péry-Woodley, Nicholas Asher, P. Enjalbert, Farah Benamara, Myriam Bras, Cécile Fabre, Stéphane Ferrari, Lydia-Mai Ho-Dac, Anne Le Draoulec, Yann Mathet, Philippe Muller, Laurent Prévot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez Couret, Laure

Vieu, and Antoine Widlöcher. ANNODIS : une approche outillée de l'annotation de structures discursives (poster). In *Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlis, France, 2009.

Martin Riegel, René Rioul, and Jean-Christophe Pellat. *Grammaire méthodique du français*. Presses universitaires de France, Paris, France, 2004.

Wilbert Spooren and Ted Sanders. The acquisition order of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics*, 40:2003–2026, 2008.

Manfred Stede and Carla Umbach. Dimlex: A lexicon of discourse markers for text generation and understanding. In *In Proceedings of the Joint 36th Meeting of the ACL and the 17th Meeting of COLING*, pages 1238–1242, 1998.

Deirdre Wilson and Dan Sperber. Linguistic form and relevance. *Lingua*, 90:1–25, 1993.