Discours
Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics

10 | 2012
Multidisciplinary Perspectives on Signalling Text Organisation

LEXCONN: A French Lexicon of Discourse Connectives

Charlotte Roze, Laurence Danlos and Philippe Muller

Electronic version
URL: http://journals.openedition.org/discours/8645
DOI: 10.4000/discours.8645
ISSN: 1963-1723

Publisher:
Laboratoire LATTICE, Presses universitaires de Caen

Electronic reference
Charlotte Roze, Laurence Danlos and Philippe Muller, « LEXCONN: A French Lexicon of Discourse Connectives », *Discours* [Online], 10 | 2012, Online since 16 July 2012, connection on 19 April 2019.
URL : http://journals.openedition.org/discours/8645 ; DOI : 10.4000/discours.8645

discours
Revue de linguistique, psycholinguistique et informatique
http://discours.revues.org/

LEXCONN: A French Lexicon of Discourse Connectives

Charlotte Roze
ALPAGE, UMR-I 001 INRIA
Université Paris Diderot, Sorbonne Paris Cité

Laurence Danlos
ALPAGE, UMR-I 001 INRIA
Université Paris Diderot, Sorbonne Paris Cité

Philippe Muller
IRIT
Université de Toulouse

Charlotte Roze, Laurence Danlos et Philippe Muller, « LEXCONN: A French Lexicon of Discourse Connectives », *Discours* [En ligne], 10 | 2012, mis en ligne le 16 juillet 2012.

URL : http://discours.revues.org/8645

Titre du numéro : *Multidisciplinary Perspectives on Signalling Text Organisation*
Coordination : Shirley Carter-Thomas & Frédéric Landragin

revues.org
CENTRE POUR L'ÉDITION ÉLECTRONIQUE OUVERTE
CENTRE FOR OPEN ELECTRONIC PUBLISHING

discours

Presses universitaires de Caen

LEXCONN: A French Lexicon of Discourse Connectives

Charlotte Roze
ALPAGE, UMR-I 001 INRIA
Université Paris Diderot, Sorbonne Paris Cité

Laurence Danlos
ALPAGE, UMR-I 001 INRIA
Université Paris Diderot, Sorbonne Paris Cité

Philippe Muller
IRIT
Université de Toulouse

With respect to discourse organization, the most basic way of signaling 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, identification of connectives

1. Introduction

1 With respect to discourse organization, the most basic way of signaling 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.

2 We will focus here on the basic inter-sentential and intra-sentential levels: what is expressed as a whole by one or 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. They contribute to discourse coherence and mark discourse structure, at least the basic organization mentioned in Spooren and Sanders' (2008) study: causality, sequence, grouping, and contrast.

3 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) select one of the two interpretations.

[1a] Peter is not stupid. He can find his own way home.

[1b] Peter is not stupid. *So* he can find his own way home.

[1c] Peter is not stupid. *After all*, he can find his own way home.

4 A lexicon of discourse connectives associated with the relations they express can be very useful for researchers in Natural Language Processing (NLP), who aim at producing 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 construction 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 [SDRT] (Asher & Lascarides, 2003), and we will evaluate how good a fit the theory is with respect to the set of connectives under investigation.

5 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, Castellon & Padro, 2002) and German (Stede & 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 with each connective an SDRT relation, and the type (coordinating or subordinating) of this relation. We used the FRANTEXT [3] corpus as a source of examples.

6 The rest of the paper is organized 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 discourse criteria for the identification of connectives. In section 4, we describe the second stage of our work:

1. This example comes from Wilson and Sperber (1993).
2. The database is available at http://www.linguist.univ-paris-diderot.fr/-croze/D/Lexconn.xml.
3. FRANTEXT is a textual base of French literature. It is available at http://www.frantext.fr.

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

7 Our work is in line with SDRT (Asher & Lascarides, 2003), which inherits from Discourse Representation Theory or DRT (Kamp, 1981) and discourse analysis (Grosz & Sidner, 1986; Mann & 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), first mentioned by Polanyi (1985). According to this constraint, in the course of building an SDRS, the only available sites for attachment of new information are the last segment of the discourse context and the segments that structurally dominate it. In SDRT, discourse relations are established by linguistic knowledge and world knowledge. For example, in the discourse [Max fell. John pushed him.], the relation *Explanation* is inferred by the Push Causal Law: the pushing can cause the falling. Discourse relations have semantic effects: for example, the relation *Explanation* has temporal and causal effects on the eventualities described in its arguments.

8 In accordance with 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 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 (in [1b], the host segment of *so* is the second sentence). The mate segment is governed by constraints described in section 3.1 (in [1b], the mate segment of *so* is the first sentence).

3. Building a lexicon of connectives

9 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 subordinating conjunctions and prepositions given by Éric Laporte (Université Paris-Est, LIGM, CNRS) and Benoît Sagot (INRIA Paris-Rocquencourt, ALPAGE, Université Paris VII), the list of French discourse markers of the ANNODIS project[4] and the corpus of English discourse connectives built by Knott (1996), that we translated manually.

10 In the database, we associate a syntactic category with each connective, which can differ a little from traditional ones: coordinating conjunction (cco) for connectives

4. ANNODIS is a project of French discourse annotation (Péry-Woodley et al., 2009). In the annotation manual, a list of possible markers was given for each discourse relation.

such as *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; subordinating conjunction (csu) for connectives such as *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 subordinating conjunctions when the host clause is an infinitive VP, such as *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].

11 After gathering a corpus of candidate connectives, we applied various criteria for their identification. In section 3.1, we present some syntactic and semantic criteria, and in section 3.2, some discourse ones.

3.1. Syntactic and semantic criteria

12 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). These criteria enabled certain connectives to be discarded from the list of candidates.

3.1.1. Cleft Criterion

13 Discourse connectives cannot be focused in cleft constructions.

14 According to Riegel, Rioul and Pellat (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 [2b] is a temporal connective which can be focused in a cleft construction, see [2c]. On the other hand, Bras (2008) claims that *à ce moment-là* in [2b] is not a connective, but a temporal cue: it only temporally locates events, and does not 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 [2b], where it refers to the temporal location of an eventuality, while it has a discourse usage in [3b] where it cannot be clefted, see [3c]. Moreover, it is referential in [3b] but not so in [3c], which goes along with the next criterion.

---

5. For some subordinating conjunctions, the mate segment is always anteposed (*comme*). For others, the mate segment can be anteposed or internal. This information is marked in LEXCONN.

6. There exist 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*.

[2a] Il a commencé à pleuvoir.
'It started raining.'

[2b] *À ce moment-là*, Paul est arrivé.
'At that moment, Paul arrived.'

[2c] C'est *à ce moment-là* que Paul est arrivé.

[3a] Tu as l'air de penser qu'elle n'est pas honnête.
'You seem to think she's not honest.'

[3b] *À ce moment-là*, ne lui raconte rien.
'So don't tell her anything.'

[3c] # C'est *à ce moment-là* que ne lui raconte rien.

3.1.2. Substitutability Criterion

15 Discourse connectives cannot be substituted (partly or entirely) by an entity (person, event, discourse unit) of the context.

16 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 have not included this type of phrases in LEXCONN.

17 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 unacceptability of [7]. The criterion tells us that *après ça* is not a connective, while *à part ça* remains in the corpus of candidate connectives.

[4a] Bruno est allé en Argentine.
'Bruno went to Argentina.'

[4b] Après ça, il est allé au Pérou.
'After that, he moved to Peru.'

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

[6a] Hier soir j'ai croisé Pierre dans un bar.
'Last night I saw Peter in a bar.'

[6b] *À part ça*, il nous dit tout le temps qu'il est fatigué.
'Though he always says he's tired.'

[7] # À part [*qu'hier soir je l'ai croisé dans un bar*], Pierre nous dit tout le temps qu'il est fatigué.

3.1.3. Compositionality Criterion

18 Discourse connectives are invariable [8].

19 Various studies (Molinier, 2003; Cojocariu & 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 discourse role for these adverbials is correlated with a process of grammaticalization. For example, the determiner and the number of *la preuve* and *à 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 number and determiner are invariable. This criterion allows us to keep some candidates like *en tout cas* and *résultat*: *en tout cas* in [8b] cannot be modified by an adjective in [8c], and *résultat* in [9b] is invariable, see [9c].

[8a] 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.'

[8b] En tout cas, dans mon souvenir, je la vois tomber...
'At any rate, I remember seeing it falling...'

[8c] *# En tout cas envisagé/possible*, dans mon souvenir, je la vois tomber...

[9a] Pierre n'a pas réussi à dormir cette nuit.
'Peter couldn't sleep last night.'

[9b] Résultat, il était en retard aujourd'hui.
'As a result, he was late today.'

[9c] *# Le résultat/Les résultats*, il était en retard ce matin.

3.2. Discourse criteria

20 The criteria we present in this section make use of the notions of discourse coherence and incoherence. They were applied after syntactic and semantic criteria, in order to verify that each connective in the lexicon plays a role in discourse interpretation. They also helped in identifying the relations conveyed by the connectives.

3.2.1. Contextual Criterion

21 If the discourse *D=c seg* (where *c* is the candidate connective and *seg* its potential host segment) is coherent without any other discourse context, then *c* is not a discourse connective.

---

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

22 The Contextual Criterion is the only test Knott used to build a list of English connectives. This test is insufficient to discard adverbials such as *le lendemain* and *un peu plus loin* which express temporal or spatial information.

3.2.2. *Forced Relation Criterion*

23 Let D_a and D_b be coherent discourses with $D_a = seg_1\ seg_2$ and $D_b = seg_1\ c\ seg_2$ (where seg_2 is the potential host segment of c), 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 differs from R_b then c plays a role in discourse interpretation.

24 Consider [10b] and [10c] which differ by the presence of *malheureusement* in [10c]. The segment in [10b] is an *Explanation* of the first segment (Mark will go camping this summer), whereas the segment in [10c] is in a *Contrast* relation with the first segment (maybe Mark will not go camping this summer). This is evidence that *malheureusement* is a connective. On the other hand, consider [10d] and [10e] which differ by the presence of *évidemment* in [10e]. The presence of this adverbial does not 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.

[10a] Marc veut faire du camping cet été.
'Mark wants to go camping this summer.'

[10b] Il n'a pas beaucoup d'argent.
'He has not got a lot of money.'

[10c] Malheureusement il n'a pas beaucoup d'argent.
'Unfortunately, he has not got a lot of money.'

[10d] Il faut qu'il économise de l'argent.
'He must save up money.'

[10e] *Évidemment*, il faut qu'il économise de l'argent.
'Obviously, he must save up money.'

3.2.3. *Coherence Criterion*

25 If $seg_1\ seg_2$ is incoherent and $seg_1\ c\ seg_2$ is coherent, then c is a discourse connective.

26 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 to link the segment in [11b] to the segment in [11a]. Without it, the discourse would be incoherent. The Coherence Criterion is based on this study. It can be used for various connectives. For example, *ceci dit* in [12b] is a discourse connective (marking the relation *Opposition*), because if it is deleted, the discourse becomes incoherent, see [12c].

[11a] Boris, je prends des gouttes pour stimuler mon appétit, mais les résultats sont lents, très lents.
'Boris, I'm taking drops to stimulate my appetite, but the results are slow, very slow.'

[11b] *À propos*, vers quel moment crois-tu que tu pourras venir ?
'By the way, when can you come?'

[12a] Ce serait vraiment utile pour nous d'aller à cette réunion.
'It would be really useful for us to go to this meeting.'

[12b] *Ceci dit*, on peut s'en passer.
'But we can do without it.'

[12c] # On peut s'en passer.

27 After applying these criteria, 328 candidates were kept as connectives [9].

4. Associating relations with connectives

28 After building the list of French discourse connectives, we tried to determine for each connective what 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*), and metatalk (*Result**, *Explanation**). Each relation is typed (coordinating or subordinating), and has semantic effects.

4.1. Tests for identifying relations

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

4.1.1. Attachment Test

30 This test helps to determine the type of the relation (Asher & Vieu, 2005). As mentioned 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 seg_1 and seg_2 is subordinating, a third segment seg_3 can be attached to seg_1, whereas if it is coordinating, seg_3 cannot be attached to seg_1, because seg_1 is no longer available for attachment. We used this test to identify the type of relation expressed by connectives.

4.1.2. Substitution Test

31 If two connectives are substitutable for each other in most of the discourse contexts they appear in, e.g. the discourse interpretation is unchanged, they probably express

9. The list of discourse markers from the ANNODIS project contains about 60 connectives.

the same discourse relation. This test is based on Knott. However, given that our goal is not to build a taxonomy of connectives/discourse relations we did not use more subtle relationships than contingent substitutability (such as synonymy, hyponymy or hyperonymy).

32 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 substitutable 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...
'This is the brilliant outcome of fifteen years of Gaullist diplomacy, but this is inevitable, given that we always preferred appearances to reality...'

4.1.3. Semantic effects

33 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* implies a temporal precedence between the eventualities it describes. Relations such as *Result* and *Explanation* can also establish causal relationships between eventualities.

4.2. Ambiguity

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

35 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 in 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.
'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...
'When I was a boy, I used to leaf again and again through back issues of this magazine. Although I was too young to understand them, they filled me with all kinds of dreams...'

36 In LEXCONN, such information about the position of the mate segment of subordinating conjunctions and the position of adverbs in their host clause is encoded by specific attributes/features (position-sub and position-adv). However, for many ambiguous connectives, the usage cannot be selected by surface clues

such as the connective's position or the mate segment's position and depends more on discourse content.

37 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 can express a temporal location.

38 We now give quantitative data about ambiguous connectives [10]: 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 establishes discourse relations of the same type (coordinating or subordinating) and the case where a connective establishes relations of different types. The first case seems less problematic than the second in an NLP perspective, because it does not imply any structural ambiguity. Only 6.2% of the total number of connectives are in the second case.

4.3. Frequency of relations

39 Table 1 gives the frequency of each discourse relation in terms of number of connectives [11]. Some of the relations listed are defined in SDRT, but some of them are not and are detailed in section 5.

Relation	Number	%	Relation	Number	%
Opposition	41	9.5	*Parallel*	13	3.0
Result	35	8.1	*Elaboration*	11	2.6
Concession	32	7.4	*Result**	11	2.6
Continuation	32	7.4	*Summary*	11	2.6
Explanation	28	6.5	*Flashback*	10	2.4
Goal	25	5.8	*Detachment*	9	2.1
Condition	25	5.8	*Alternation*	9	2.1
*Explanation**	24	5.6	*Consequence*	7	1.6
Narration	23	5.4	*Background(f)*	7	1.6
Unknown	21	4.9	*Evidence*	7	1.6
Contrast	17	4.0	*Rephrasing*	6	1.4
Background(b)	15	3.5	*Digression*	6	1.4
Temporal	14	3.3	**Total**	**428**	**100%**

Table 1. Frequencies of relations: number and percentage of connectives

10. We do not consider connectives marked as "unknown" in the counts.

11. Note that we distinguish several usages for some connectives.

5. Problematic cases for SDRT

40 Associating discourse relations with connectives leads to the following conclusion: 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 of all French discourse connectives to discourse interpretation. 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 [RST] (Mann & Thompson, 1988). Second, the case where it seems impossible to associate any relation to a discourse connective.

5.1. Introducing new relations in SDRT

41 In LEXCONN, there are six relations which are not defined in SDRT: *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). They 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 in which they appear.

42 For example, *bien que* or *même si* are considered in ANNODIS as possible markers of the coordinating relation *Contrast*. However, they express a subordinating relation, as shown in [16]: the segments [16a] and [16c] are linked by the relation *Result*, therefore, according to the RFC, the relation between [16a] and [16b] is subordinating. In addition, these connectives link segments which do not necessarily present 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*: in LEXCONN, they are associated with the subordinating relation *Concession* (which is defined in RST).

[16a] Pierre m'a aidé à repeindre la chambre...
'Peter helped me repaint the bedroom...'

[16b] ...bien qu'il ait beaucoup de boulot en ce moment.
'...even though he has a lot of work at the moment.'

[16c] Du coup, c'est déjà terminé!
'Thus it is already over!'

5.2. Unknown relations

43 For 21 connectives (about 6%), the associated discourse relation in LEXCONN is unknown. Among these connectives, there are adverbs (*en fait*, *au moins*), subordinating conjunctions (*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.

44 For example, *à mesure que*, whose meaning is non-compositional, as shown by the unacceptability of [17c], and which does not contain a referential expression, as shown in [17d], 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.

[17a] Tes digressions s'allongeaient...
'Your digressions got longer and longer...'

[17b] ...à mesure que tu finissais les alcools de ta mère.
'...as and when you finished your mother's alcohols.'

[17c] # *à la mesure que* tu finissais les alcools de ta mère.

[17d] # *à cette mesure-là*.

6. Conclusion

45 Building a French lexicon of discourse connectives has produced several results. It required a systematic methodology to identify discourse connectives and associate discourse relations to them, resting on various studies concerning 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 points, 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.

46 Despite these results, some information must be added in LEXCONN, in particular about ambiguity between discourse and non-discourse usage. It will be possible with further linguistic analysis, but also with automatic analysis on the ANNODIS corpus: the link between position in the host clause and discourse/non-discourse role for adverbials must be studied.

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

References

Alonso, L., Castellon, I. & Padro, L. 2002. Lexicón computacional de marcadores del discurso. *SEPLN, XVIII Congreso anual de la Sociedad Española para el Procesamiento del Lenguaje Natural.*

Asher, N. & Lascarides, A. 2003. *Logics of Conversation*. Cambridge – New York: Cambridge University Press.

Asher, N. & Vieu, L. 2005. Subordinating and Coordinating Discourse Relations. *Lingua* 115 (4): 591-610.

Beaulieu-Masson, A. 2002. Quels marqueurs pour parasiter le discours ? *Cahiers de linguistique française* 24: 45-71.

Bras, M. 2008. *Entre relations temporelles et relations de discours*. "Habilitation à diriger les recherches" report. Université de Toulouse II-Le Mirail.

Cojocariu, C. & Rossari, C. 2008. Constructions of the Type *la cause/la raison/la preuve* + Utterance: Grammaticalization, Pragmaticalization, or Something Else? *Journal of Pragmatics* 40 (8): 1435-1454.

Danlos, L. 2009. D-STAG: A Formalism for Discourse Analysis Based on SDRT and Using Synchronous TAG. In P. de Groote (ed.), *Proceedings of the 14th Conference on Formal Grammar (FG'09)*. Rocquencourt: INRIA: 1-20. Available online: http://hal.inria.fr/docs/00/51/17/28/PDF/D-STAG-FG_09.pdf.

Grosz, B.J. & Sidner, C.L. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12 (3): 175-204.

Jayez, J. & Rossari, C. 1996. *Donc* et les consécutifs. Des systèmes de contraintes différentiels. *Linguisticæ Investigationes* 20 (1): 117-143.

Kamp, H. 1981. Événements, représentations discursives et référence temporelle. *Langages* 15 (64): 39-64.

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

Mann, W. & Thompson, S. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text* 8 (3): 243-281.

Molinier, C. 2003. Connecteurs et marqueurs énonciatifs : les compléments figés formés à partir du nom *propos*. *Linguisticæ Investigationes* 26 (1): 15-31.

Nakamura, T. 2009. 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)*.

Péry-Woodley, M.-P. et al. 2009. ANNODIS : une approche outillée de l'annotation de structures discursives. In *Actes de la 16e conférence sur le Traitement automatique des langues naturelles (TALN 2009 – Senlis, 24-26 juin 2009)*. Available online: http://www-lipn.univ-paris13.fr/taln09/paper/paper_TALN_52.html.

Polanyi, L. 1985. A Theory of Discourse Structure and Discourse Coherence. In *Proceedings of the 21st Meeting of the Chicago Linguistics Society*.

Riegel, M., Pellat, J.-C. et Rioul, R. 2004. *Grammaire méthodique du français*. Paris: PUF.

Spooren, W. & Sanders, T. 2008. The Acquisition Order of Coherence Relations: On Cognitive Complexity in Discourse. *Journal of Pragmatics* 40 (12): 2003-2026.

Stede, M. & Umbach, C. 1998. DIMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. In *Proceedings of the Joint 36th Meeting of the ACL and the 17th Meeting of COLING*. 1238-1242. Available online: http://acl.ldc.upenn.edu/P/P98/P98-2202.pdf.

Wilson, D. & Sperber, D. 1993. Linguistic Form and Relevance. *Lingua* 90 (1-2): 1-25.