Exploring a corpus annotated in causal discourse relations for the study of causal lexical clues

Caroline Atallah, Myriam Bras, Laure Vieu

▶ **To cite this version:**

Caroline Atallah, Myriam Bras, Laure Vieu. Exploring a corpus annotated in causal discourse relations for the study of causal lexical clues. Final Action Conference TextLink 2018 : Cross-Linguistic Discourse Annotation: applications and perspectives, Lydia-Mai Ho-Dac (University of Toulouse, CLLE-ERSS); Philippe Muller (University of Toulouse, IRIT), Mar 2018, Toulouse, France. hal-02982984

HAL Id: hal-02982984
https://hal.science/hal-02982984

Submitted on 3 Nov 2020

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.

Exploring a corpus annotated in causal discourse relations for the study of causal lexical clues

Caroline Atallah[1], Myriam Bras[2], Laure Vieu[3]

[1] LIDILE, Université de Rennes, France
[2] CLLE, Université de Toulouse, CNRS, UT2J, France
[3] IRIT, CNRS, Université de Toulouse, France

1 Introduction

Usually, the study of Discourse Relations (DRs) is based on Lexical Clues (LCs) commonly associated with these DRs, like connectives. For example, a corpus study of causal DRs can be done from the analysis of some connectives commonly associated with causality, like *because*. Such a semasiological approach, that proceeds from a given LC towards DRs, has a significant advantage: it is much easier to locate LCs than DRs in a corpus.

The approach presented here complementarily exploits two types of analysis. We first adopt an onomasiological approach, that proceeds from a given DR towards LCs. In other words, we analyze all the occurrences of this DR in a corpus in order to identify all the LCs that contribute to the DR interpretation. Then, the results of these first analyses are completed by a semasiological analysis: each LC that has been identified is projected on the corpus in order to determine whether it specifically marks the given DR or not.

The onomasiological approach requires working on data that have previously been annotated with DRs. Before the ANNODIS corpus was built (*ANNotation DIScursive de corpus*; Péry-Woodley et al., 2009, 2011; Afantenos et al., 2012), such data did not exist for French and an onomasiological approach, as presented above, was simply impossible for this language. This rather new methodology has already been applied to a few DRs on the ANNODIS corpus (see Vergez-Couret, 2010, for an application to *Elaboration* DR). We propose to focus here on a specific family of DRs: causal DRs, and to base our study on a corpus specifically annotated with causal DRs: the EXPLICADIS corpus (*EXPLication et Argumentation en DIScours*; Atallah, 2014; Atallah, 2015).

2 The EXPLICADIS corpus

In the ANNODIS project, 86 texts were segmented into Elementary Discourse Units (EDUs) and then annotated with a tagset of DRs inspired by SDRT relations (*Segmented Discourse Representation Theory*, Asher and Lascarides, 2003). The EXPLICADIS corpus has been built in the continuity of ANNODIS: the 86 texts were reused and re-annotated with a more complete and accurate new set of causal DRs.

Then 31 more texts were added, segmented and annotated in order to provide a better representation of different text genres: narrative, expositive and argumentative. The whole EXPLICADIS corpus includes 117 texts, 4,580 EDUs and 39,103 tokens.

This new set of causal DRs was adopted in order to remedy the difficulties experienced by ANNODIS annotators with the first set of DRs and to adequately account for the data in a semantically clear set of relations (Atallah, 2014; Atallah et al., 2016). It includes, like the previous one, two types of relations: *Explanation* relations (noted further Rh_Exp) and *Result* relations (noted further Rh_Res)[1]. The new set is original because it distinguishes within both rhetorical types four subtypes of DRs:

- content-level DRs that involve a causal link between the eventualities that are described in the propositional content: *Explanation (α,β)* (1) and *Result (α,β)* (2);
- epistemic DRs that involve a causal link between knowledge items and beliefs: $Explanation_{ep}$ *(α,β)* (3) and $Result_{ep}$ *(α,β)* (4) ;
- inferential DRs that involve a causal link between knowledge items: $Explanation_{inf}$ *(α,β)* (5) and $Result_{inf}$ *(α,β)* (6);
- speech-act (or pragmatic) DRs that involve a causal link between an eventuality that is described in the propositional content and a speech act: $Explanation_{prag}$ *(α,β)* (7) and $Result_{prag}$ *(α,β)* (8).

A total of 319 causal DRs were annotated using this tagset, including 186 Rh_Exp relations and 133 Rh_Res relations. Examples of each type of these DRs are presented below:

(1) [L'armée est déçue,]$_{\alpha}$ [il n'y a aucun viol, aucun pillage, aucun meurtre.]$_{\beta}$
([The army is disappointed,]$_{\alpha}$ [there is no rape, no looting, no murder.]$_{\beta}$)

(2) [le côté gauche de la voiture a mordu l'accotement.]$_{\alpha}$ [L'automobile a perdu sa roue gauche.]$_{\beta}$
([the left side of the car hit the roadside.]$_{\alpha}$ [The car lost its left wheel.]$_{\beta}$)

(3) [Ce phénomène semble se confirmer à Mariana,]$_{\alpha}$ [où on peut observer deux voies parallèles à la sortie sud de la ville.]$_{\beta}$
([This phenomenon seems to be confirmed in Mariana,]$_{\alpha}$ [where two parallel roads can be observed at the south exit of the city.]$_{\beta}$)

(4) [Or la psychomécanique répond à ces deux types d'exigences.]$_{\alpha}$ [Il serait donc intéressant de regarder si les outils théoriques qu'elle a développés permettent de rendre compte de certaines observations faites par la neuropsychologie.]$_{\beta}$
([Yet psychomechanics meets these two types of requirements.]$_{\alpha}$ [It would therefore be interesting to examine whether the theoretical tools it has developed are able to account for certain observations made by neuropsychology.]$_{\beta}$)

[1] « Rh_ » is put for « Rhetorical ». We consider that *Explanation* relations and *Result* relations do not simply differ in the order of presentation, but rather in the rhetorical choice of presentation.

(5) [BITNET était différent d'Internet]$_{\alpha}$ [parce que c'était un réseau point-à-point de type « stocké puis transmis ».]$_{\beta}$
([BITNET was different from Internet]$_{\alpha}$ [because it was a point-to-point network of "stored and transmitted" type.]$_{\beta}$)

(6) [La première exposition avicole de Belfort date de 1922.]$_{\alpha}$ [Cela fait donc plus de trois-quarts de siècle que la digne société du même nom encourage, dans la région, les éleveurs amateurs.]$_{\beta}$
([The first avicultural Belfort exhibition dates back to 1922.]$_{\alpha}$ [Therefore, the honorable society of that name has been supporting farmers for more than seven decades.]$_{\beta}$)

(7) [Mais que ces derniers se rassurent,]$_{\alpha}$ [il y aura encore deux autres tours pour se rattraper.]$_{\beta}$
([These can rest assured[2] that]$_{\alpha}$ [there will be two more rounds to catch up.]$_{\beta}$)

(8) [Suzanne Sequin n'est plus.]$_{\alpha}$ [...] [Nos condoléances.]$_{\beta}$
([Suzanne Sequin is gone.]$_{\alpha}$ [...] [Our condolences.]$_{\beta}$)

After annotation, each of these DRs has been analyzed in order to identify lexical clues (LCs). Within LCs, we draw a line between *clues* and *markers*. We consider that a *clue* is a linguistic unit that plays a potential role in the DR interpretation; while a *marker* has an established function in discourse interpretation, it plays a primordial role in the inference of a DR (Vergez-Couret, 2010; Péry-Woodley, 2000). Thus, for us, a clue is just a potential marker.

To identify causal LCs, we tried to spot every LCs that could have helped to guide our interpretation to a causal DR during the annotation process. It is important to note that those LCs are not necessarily responsible (on their own) for the inference of the causal DR. We consider that actually, in most cases, it is a whole bundle of clues that contributes to the inference of a DR. Thus, by *LCs* we do not mean *discourse markers*, but a simple clue that accompany the DR. To determine the discursive function(s) associated with a LC requires a more in-depth study than the one presented here, a semasiological study of bigger data.

The onomasiological approach we first adopted has its own advantages. For example, it allowed us to study causal DRs associated to LCs but also DRs being annotated without the help of any LC. Those represent 38.87% of the annotated causal DRs. We noticed that the presence of LCs was related to the rhetorical choice, the type of causal DR, but also the text genre. The methodology adopted also helped listing causal LCs, and thereby noticing that LCs associated with Rh_Exp DRs were more diversified (31 LC types for 186 DR occurrences) than LCs associated with Rh_Res (21 LC types for 133 DR occurrences). This observation must however be considered carefully, given the small size of the corpus.

[2] This translation does not keep the imperative form of the verb, impossible in English with a third person. The French construction is similar to a English "But rest assured," in which the imperative is directed to the addressee instead.

3 The LEX-PLICADIS database

We compared our LC list with another existing inventory: LexConn (Roze, 2009; Roze et al., 2012). This resource lists French connectives and associates each of them to one or more DRs[3]. The causal DRs used in LexConn is the classical SDRT set, only including two types of causal DRs: content-level and speech-act relations. Thus, to compare EXPLICADIS LCs with LexConn LCs, we consider that LexConn speech-act causal DRs correspond to one of the three following types of DRs: epistemic, inferential or speech-act DRs.

Among the 52 different LCs we identified, 23 LCs were not recorded at all in LexConn and 4 were listed but not associated with causality. We therefore decided to complete LexConn with EXPLICADIS data in order to create a new database: LEX-PLICADIS.

To fill it, we completed the onomasiological analysis with a semasiological one. We first projected each LC identified on the whole EXPLICADIS corpus, in order to verify whether it was specialized in the expression of causality or not. Results were then compared with LexConn. We also analyzed the 70 LCs that were associated with causality in LexConn but not in EXPLICADIS. Naturally, the absence of an association between a LC and a DR in EXPLICADIS does not question the information listed in LexConn. Such a study should be continued on a larger annotated corpus. We therefore decided to be as exhaustive as possible and to record all the causal LCs identified in EXPLICADIS and/or in LexConn, specifying if it was associated in each resource to:

- a content-level DR;
- an epistemic causal DR;
- an inferential causal DR;
- a speech-act (or pragmatic) causal DR;
- a non-causal DR.

The complete database includes 120 causal LCs, among which 67 LCs associated with Rh_Exp DRs and 53 with Rh_Res DRs. We provide in table 1 an excerpt that concerns the 52 LCs we identified in EXPLICADIS and associated with the expression of causality.

Table 1. Excerpt of the LEX-PLICADIS database

LC	Rh_Exp DRs				other DRs
	content-level	epistemic	inferential	speech-act	
à cause de	L- E+	L- E-	L- E-	L- E-	L- E-
à la suite de	L- E+	L- E-	L- E-	L- E-	L- E+
avec	L- E+	L- E-	L- E-	L- E-	L- E+

[3] It is interesting to note that LexConn had been partly built on the basis of the LCs listed in the ANNODIS annotation guide.

car	L- E+	L* E+	L* E+	L* E-	L- E-
comme	L+ E+	L* E-	L* E-	L* E-	L+ E+
conséquence de	L- E+	L- E-	L- E-	L- E-	L- E-
d'autant plus que	L+ E+	L- E-	L- E-	L- E-	L- E-
d'autant que	L+ E-	L- E+	L- E-	L- E-	L- E-
dans la mesure où	L- E-	L* E+	L* E-	L* E-	L+ E-
de	L- E+	L- E-	L- E-	L- E-	L- E+
dès que	L+ E+	L- E-	L- E-	L- E-	L+ E+
des suites de	L- E+	L- E-	L- E-	L- E-	L- E-
devant	L- E+	L- E-	L- E-	L- E-	L- E+
du fait de	L- E+	L- E-	L- E-	L- E-	L- E-
en+Verb-*ANT [gerund]*	L+ E+	L- E-	L- E-	L- E-	L+ E+
en effet	L- E+	L* E+	L* E+	L* E-	L- E+
en raison de	L- E+	L- E-	L- E-	L- E-	L- E-
en témoignage de	L- E+	L- E-	L- E-	L- E-	L- E-
étant donné	L- E-	L- E+	L- E+	L- E-	L- E-
étant donné que	L+ E-	L- E+	L- E-	L- E-	L- E-
faute de	L+ E+	L- E-	L- E-	L- E-	L- E-
grâce à	L- E+	L- E-	L- E-	L- E-	L- E-
le temps de	L- E+	L- E-	L- E-	L- E-	L- E+
par	L- E+	L- E-	L- E-	L- E-	L- E+
parce que	L+ E+	L* E+	L* E+	L* E-	L- E-
pour	L- E+	L- E-	L- E-	L- E-	L+ E+
pour des raisons (de)	L- E+	L- E+	L- E-	L- E-	L- E-
puisque	L+ E+	L* E+	L* E-	L* E-	L- E-
si... c'est que	L- E+	L- E-	L- E-	L- E-	L- E-
suite à	L- E+	L- E-	L- E-	L- E-	L- E+
vu	L- E-	L- E+	L- E-	L- E-	L- E-

LC	Rh_Res DRs				other DRs
	content-level	epistemic	inferential	speech-act	
à ce rythme	L- E-	L- E-	L- E+	L- E-	L- E-
ainsi	L+ E+	L- E+	L- E+	L- E-	L- E+
alors	L+ E+	L* E-	L* E-	L* E-	L+ E+
au point que	L+ E+	L- E-	L- E-	L- E-	L- E-
au prix de	L- E-	L- E+	L- E-	L- E-	L- E-
aussi [initial position]	L+ E-	L- E+	L- E-	L- E-	L- E-
avec pour conséquence	L- E+	L- E-	L- E-	L- E-	L- E-
c'est pourquoi	L+ E+	L- E+	L- E-	L- E-	L- E-
conduisant à	L- E+	L- E-	L- E-	L- E-	L- E-
de sorte que	L+ E+	L- E-	L- E+	L- E-	L- E-
dès lors	L- E+	L* E-	L* E-	L* E-	L- E-
donc	L+ E+	L* E+	L* E+	L* E-	L- E+
d'où	L+ E-	L- E-	L- E+	L- E-	L- E+

et	L- E+	L- E-	L- E-	L- E-	L+ E+
jusqu'à ce que	L+ E+	L- E-	L- E-	L- E-	L+ E+
pour	L- E+	L- E-	L- E-	L- E-	L+ E+
preuve que	L- E-	L* E+	L* E-	L* E-	L- E-
résultat(s)	L+ E+	L- E-	L- E-	L- E-	L- E-
si bien que	L+ E+	L- E-	L- E-	L- E-	L- E-
tant que	L- E+	L- E-	L- E-	L- E-	L+ E-
tel(les)... que	L- E+	L- E-	L- E-	L- E-	L- E-

"L-": LC absent in LexConn "E-": LC absent in EXPLICADIS
"L+": LC present in LexConn "E+": LC present in EXPLICADIS
"L*": LC associated in LexConn with a speech-act causal DR

The study of the repartition of each LC allowed us to test some hypotheses formulated in the literature. For example, we found, for Rh_Exp DRs, that the values originally associated with *parce que* and *car* (*because*) (Groupe λ-l, 1975; Degand and Fagard, 2008) still persisted: *car* is more subjective than *parce que* (Simon and Degand, 2007), ie it is more often associated with epistemic DRs than content-level DRs. And we found, for Rh_Res DRs, that *donc* (*therefore*) was specialized in inferential DRs. *Donc* forces some sort of inferential reading: in a content-level DR, the effect described is presented as an inevitable event, and in an epistemic DR, the conclusion is presented as an obvious and indisputable fact (Hybertie, 1996).

4 Conclusion

To build the new resource LEX-PLICADIS, onomasiological and semasiological approaches were used complementarily. Thanks to the onomasiological analysis, which consists in a sort of exhaustive exploration of the corpus, we got results that could not have been obtained otherwise, such as DR occurring without LC. It also enabled us to add to LexConn many associations between LCs and DRs that had not been envisaged. It was important to complete and test the LexConn proposals for the causality domain. The same work should be done with other domains in a method akin to the ASFALDA French FrameNet project's one (Djemaa et al., 2016).

However, as an onomasiological approach requires a corpus annotated with DRs and as such a corpus requires a long and hard work, it implies to work with small quantity of data and to accept that the corpus, because of its size, presents limitations. Therefore, the onomasiological study must be considered and adopted as a first exploratory and non-exhaustive phase of the analysis, which can be then completed by a semasiological study on a bigger corpus.

References

1. Afantenos, S. D., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, L.-M., Le Draoulec, A., Muller, P., Péry-Woodley, M.-P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M. et Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organization : the ANNODIS corpus. *In Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, pages 2727–2734, Istanbul, Turkey.
2. Asher, N., and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.
3. Atallah, C. (2014). *Analyse de relations de discours causales en corpus : étude empirique et caractérisation théorique*. Thèse de Doctorat, Université de Toulouse.
4. Atallah, C. (2015). La ressource EXPLICADIS, un corpus annoté spécifiquement pour l'étude des relations de discours causales. In *Actes de TALN 2015* (pp. 551–557). Caen.
5. Atallah, C., Vieu, L., Bras, M. (2016). Formal characterization of a new set of causal discourse relations. *NISM 2016 (New Ideas in Semantics and Modeling)*, Paris, 7-8 septembre 2016.
6. Degand, L. et Fagard, B. (2008). (Inter)subjectification des connecteurs : le cas de *car* et *parce que*. *Revista de Estudos Linguìsticos da Universidade do Porto*, 3(1):119–136.
7. Djemaa, M., Candito, M., Muller, P. and Vieu, L. (2016). Corpus annotation within the French FrameNet: a domain-by-domain methodology. *In* Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Goggi, and Grobelnik, editors, *Language Resources and Evaluation Conference (LREC)*, pages 3794–3801, Portoroz, Slovenia, 23-28 May 2016. *European Language Resources Association (ELRA)*.
8. Groupe λ-l (1975). *Car*, *parce que*, *puisque*. *Revue Romane*, 10(2):258–280.
9. Hybertie, C. (1996). *La conséquence en français*. L'essentiel français. Ophrys, Paris/Gap.
10. Péry-Woodley, M.-P. (2000). *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*. Mémoire d'HDR, Université Toulouse II Le Mirail, Toulouse.
11. Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M. and Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *TAL*, 52(3):71–101.
12. Péry-Woodley, M.-P., Asher, N., Enjalbert, P., Benamara, F., Bras, M., Fabre, C., Ferrari, S., Ho-Dac, L.-M., Le Draoulec, A., Mathet, Y., Muller, P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M., Vieu, L. and Widlöcher, A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. *In Actes de TALN 2009*, Senlis, France.
13. Roze, C. (2009). *Base lexicale des connecteurs discursifs du français*. Mémoire de Master 2, Université Paris Diderot, Paris.
14. Roze, C., Danlos, L. and Muller, P. (2012). LEXCONN: A French Lexicon of Discourse Connectives. *Discours*, (10).
15. Simon, A. C. et 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(03):323–341.
16. Vergez-Couret, M. (2010). *Etude en corpus des réalisations linguistiques de la relation d'Elaboration*. Thèse de Doctorat, Université Toulouse II Le Mirail.