

Vers le FDTB : French Discourse Tree Bank

Laurence Danlos Diégo Antolinos-Basso Chloé Braud Charlotte Roze
ALPAGE, Université Paris Diderot, 175 rue du Chevaleret, 75013 Paris
prenom.nom@linguist.jussieu.fr

RÉSUMÉ

Nous présentons les premiers pas vers la création d'un corpus annoté en discours pour le français : le French Discourse TreeBank enrichissant le FTB. La méthodologie adoptée s'inspire du Penn Discourse TreeBank (PDTB) mais elle s'en distingue sur au moins deux points à caractère théorique. D'abord, notre objectif est de fournir une couverture totale d'un texte du corpus, tandis que le PDTB ne fournit qu'une couverture partielle, qui ne peut donc pas être qualifiée d'analyse discursive comme celle faite en RST ou SDRT, deux théories majeures sur le discours. Ensuite, nous avons été amenés à définir une nouvelle hiérarchie des relations de discours qui s'inspire de RST, de SDRT et du PDTB.

ABSTRACT

Towards the FDTB : French Discourse Tree Bank

We present the first steps towards creating an annotated corpus for discourse in French : the French Discourse Treebank enriching the FTB. Our methodology is based on the Penn Discourse Treebank (PDTB), but it differs in at least two points of a theoretical nature. First, our goal is to provide full coverage of a text, while the PDTB provides only partial coverage, which can not be described as discourse analysis such as the one made in RST or SDRT, two major theories on discourse. Second, we were led to define a new hierarchy of discourse relations which is based on RST, SDRT and PDTB.

MOTS-CLÉS : Discours, corpus annoté manuellement, analyse discursive, PDTB, RST, SDRT.

KEYWORDS: Discourse, manually annotated corpus, discourse analysis, PDTB, RST, SDRT.

1 Introduction

Dans l'idée de disposer de corpus annotés pour le français, nous avons l'objectif de développer le FDTB (French Discourse Tree Bank), un corpus annoté pour l'analyse discursive. Le FDTB s'inspire du PDTB (Penn Discourse Tree Bank, (PDTB Group, 2008)) qui ajoute une couche d'annotation discursive (manuelle) sur le PTB-v2 (Penn Tree Bank, (Marcus *et al.*, 1999)), corpus anglais tiré du *Wall Street Journal* annoté manuellement pour la morpho-syntaxe. De même, le FDTB ajoute une couche d'annotation discursive sur le FTB (French Tree Bank, (Abeillé *et al.*, 2003)), corpus français tiré du journal *Le Monde* annoté manuellement pour la morpho-syntaxe¹.

Le PDTB est présenté à la Section 2. Si le FDTB s'en inspire, il s'en départit néanmoins sur

1. Le corpus Annodis (Péry Woodley *et al.*, 2009) est aussi un corpus français annoté manuellement pour l'analyse discursive. Le FDTB s'en distingue principalement sur deux points : (i) le fait de disposer d'une analyse syntaxique manuelle du corpus et (ii) de s'appuyer fortement sur les connecteurs de discours (explicites et implicites).

certains choix méthodologiques (Section 3), entre autres, sur le fait que nous voulons une “couverture totale” des articles journalistiques du corpus et non une ‘couverture partielle’ comme obtenue dans le PDTB. Enfin, nous concluons sur quelques remarques concernant les premières annotations effectuées. Soulignons que celles-ci ne sont que préliminaires, la réalisation d’un corpus comme le FDTB étant un travail de longue haleine qui n’a commencé que depuis six mois (pour l’instant au sein d’une seule équipe).

2 Présentation du PDTB

Le PDTB s’appuie sur trois principes de base qui font consensus dans la communauté travaillant sur le Discours :

- (i) Un connecteur de discours est un prédicat sémantique à deux arguments dénotant des “objet abstraits” (notés AO dans (Asher, 1993) qui a introduit cette notion). Pour une annotation discursive, il faut donc identifier dans le texte d’entrée les connecteurs de discours - ceux-ci appartiennent à une liste fermée d’éléments regroupant principalement conjonctions de coordination et de subordination et certains adverbiaux - , vérifier qu’ils sont bien employés comme connecteurs de discours², et enfin délimiter les emplacements de texte correspondant aux arguments de ces connecteurs.
- (ii) Un connecteur de discours lexicalise une “relation de discours” (“relation rhétorique”) qui appartient à une liste fermée d’éléments organisés dans une hiérarchie arborescente, où les feuilles sont les relations de discours, les nœuds intermédiaires des classes de relations de plus en plus générales en montant vers la racine. Il faut donc pour chaque connecteur de discours indiquer quelle(s) relation(s) de discours il lexicalise en précisant une feuille de la hiérarchie en cas de certitude ou en remontant dans la hiérarchie en cas de doute.
- (iii) Les relations de discours ne sont pas forcément lexicalisées par un connecteur de discours (explicite) : elles sont alors inférables³. On parle alors de “connecteur implicite” (ou de “connecteur vide \emptyset ”). Pour une annotation discursive, il faut donc identifier les positions où on doit insérer un connecteur implicite, et appliquer aux connecteurs implicites les traitements décrits en (i) et (ii) pour les connecteurs explicites⁴.

En plus des annotations venant des trois principes de base décrits ci-dessus, sont annotées dans le PDTB pour chaque connecteur de discours (explicite ou implicite) la source de la relation de discours en jeu et la source de chacun de ses arguments. Ainsi, pour l’exemple (3a) de (Prasad *et al.*, 2006) où le connecteur *while* est souligné, ses arguments Arg1 et Arg2 repérés par les segments de texte respectivement en italiques et en gras, le “segment attributif” *purchasing agent said*, placé dans une boîte, se contente d’indiquer que la source de Arg2 n’est pas le locuteur

2. Un même mot ou groupe de mots peut avoir un emploi comme connecteur de discours et un emploi non discursif, comme illustré pour *à ce moment-là* en (1) : en (1a), il s’agit d’un emploi discursif mais pas en (1b), (Roze, 2009).

- (1) a. Tu as l’air de penser qu’elle n’est pas honnête. *A ce moment-là*, ne lui raconte rien.
- b. Il a commencé à pleuvoir. Marie est arrivée *à ce moment-là*.

3. Ainsi en (2a), le locuteur demande à son interlocuteur d’inférer que le contenu propositionnel de la seconde phrase est la cause du contenu propositionnel de la première phrase sans que la relation de discours *Explication* liant ces deux AO soient explicitement indiqués par un connecteur, comme elle l’est en (2b).

- (2) a. Fred n’était pas en forme aujourd’hui. Il a mal dormi la nuit dernière.
- b. Fred n’était pas en forme aujourd’hui parce qu’il a mal dormi la nuit dernière.

4. En plus des connecteurs explicites ou implicites, le PDTB fait aussi appel aux notions *AltLex*, *EntRel*, et *NoRel*, mais nous n’avons pas la place de les présenter dans cet article.

mais les purchasing agents. Les annotations basiques du PDTB sont donc complétées par d'autres informations, représentées ici sous forme de tableau donné en (3b). Ce tableau indique la valeur du trait [Source] pour la relation REL (marquée par *while* et identifiée comme étant *Contraste*), pour Arg1 et pour Arg2. La valeur "Wr" est utilisée pour l'auteur ("writer") du texte, "Inh" indique que la valeur de [Source] est héritée de celle de REL, "Ot" ("other") est utilisée pour un (ou des) individu(s) autre(s) que l'auteur (il s'agit des *purchasing agents* pour Arg2). Le tableau comporte d'autres informations relatives à la factivité et la polarité, les traits [Type], [Polarity] et [Determinacy], mais nous ne les détaillerons pas ici ⁵.

(3)a. *Factory orders and construction outlays were largely flat in December while*

purchasing agents said **manufacturing shrank further in October.**

	REL	Arg1	Arg2
b. [Source]	Wr	Inh	Ot
[Type]	Comm	Null	Comm
[Polarity]	Null	Null	Null
[Determinacy]	Null	Null	Null

Quelques données quantitatives concernant le PDTB issues de (Prasad *et al.*, 2008) : 18459 connecteurs explicites — appartenant à une liste fermée de 100 éléments — et 16224 connecteurs implicites ont été annotés. Les relations de discours sont au nombre de 30 réparties dans quatre classes majeures (*TEMPORAL*, *CONTINGENCY*, *COMPARISON* et *EXPANSION*). Ce travail d'annotation fut mené sur plusieurs années par des chercheurs senior et des doctorants ⁶. L'accord inter-annotateur est : pour les relations de discours de 77 % sur les feuilles de la hiérarchie et de 90% sur les quatre classes majeures ; pour les empan des arguments de 90.2 % pour les connecteurs explicites et de 85.1 % pour les implicites. Un outil d'annotation ANNOTATOR a été développé ainsi qu'une interface de visualisation des annotations, qui propose entre autres de voir pour chaque argument de connecteur l'analyse syntaxique proposée dans le PTB.

L'exploitation des annotations du PDTB a donné lieu à multe voix de recherche (en plus des nombreuses statistiques données en annexes du manuel d'annotation (PDTB Group, 2008)), citons le parsing de discours (Lin *et al.*, 2011), la classification des connecteurs (Pitler *et al.*, 2010), et d'autres récapitulées dans (Webber *et al.*, 2011).

Des projets analogues pour d'autres langues (dont le chinois, le turc et l'hindi) sont en cours de développement. Et nous nous attaquons maintenant au français, en effectuant quelques choix différents de ceux du PDTB, non pas tant à cause de différences entre les langues concernées — l'anglais et le français sont relativement très proches — mais pour des raisons méthodologiques ou théoriques ⁷.

5. Comme expliqué dans (Danlos, 2011), les informations de factivité utilisées dans FactBank (Sauri et Pustejovsky, 2009) sont plus élaborées que celles du PDTB et nous projetons donc d'utiliser les informations de factivité à la FactBank plutôt que celles du PDTB.

6. D'après A. Joshi (pc) l'idée de départ était de confier ce travail à des étudiants undergraduate, comme ce fut le cas pour l'annotation morpho-syntaxique du PTB. Elle fut abandonnée car l'analyse discursive est trop difficile : elle repose en effet beaucoup sur de l'interprétation qui peut déraiper dans une subjectivité totale et improductive si elle n'est pas guidée par des principes coercitifs reposant sur des connaissances solides en morpho-syntaxe, sémantique et pragmatique.

7. Cette position ne va pas sans inconvénients, par exemple le fait de ne pas pouvoir comparer directement le PDTB au PDTB qui fait référence. Néanmoins, ces inconvénients sont contre-balançés par le fait qu'il semble nécessaire de faire avancer les connaissances sur l'analyse discursive.

3 Différences entre le FDTB et le PDTB

Nous ne sommes pas d'accord avec tous les choix méthodologiques — et les principes théoriques sous-jacents — du PDTB, et nous nous orientons donc vers d'autres choix décrits ci-dessous.

3.1 Couverture partielle versus totale

Notre objectif dans le FDTB est d'avoir une “couverture totale” du texte annoté (i.e. un article *du Monde*). La couverture totale s'oppose à la “couverture partielle” réalisée dans le PDTB. L'équipe du PDTB s'est fixée comme objectif d'annoter les arguments de certains connecteurs (plus précisément, les connecteurs explicites appartenant à une liste d'une centaine d'éléments, plus certains connecteurs implicites mais pas tous (PDTB Group, 2008)) : leur objectif n'est pas d'obtenir une analyse discursive complète du texte comme celle obtenue en RST (Mann et Thompson, 1987; Taboada et Mann, 2006) ou SDRT (Asher et Lascarides, 2003) où un graphe discursif connexe couvre **tous** les segments du texte (au même titre qu'une analyse syntaxique d'une phrase couvre **tous** les mots de la phrase). Cette position du PDTB est clairement revendiquée (Webber *et al.*, 2011), mais ce n'est pas la nôtre car seule une analyse complète d'un texte permet de rendre compte de sa cohérence et d'en extraire des informations ou de le résumer adéquatement, par exemple.

Une couverture totale demande de se départir du PDTB sur les points suivants :

- Annoter tous les connecteurs explicites et non pas simplement ceux appartenant à une liste de 100 éléments définis comme les plus fréquents⁸. Le lexique LEXCONN répertorie une liste aussi exhaustive que possible des connecteurs du français : il comporte plus de 300 éléments (Roze, 2009; Roze *et al.*, 2010). Le FDTB va annoter tous les connecteurs de LEXCONN — après désambiguation entre emplois discursifs versus non discursifs.
- Annoter plus de connecteurs implicites que dans le PDTB. Le PDTB a fortement restreint les positions où il était licite de poser que deux AO sont reliés par un connecteur vide (PDTB Group, 2008). Par exemple, les connecteurs vides des discours en (4) ne sont pas considérés alors qu'il est clair qu'une relation de discours existe entre les deux AO à gauche et à droite du connecteur \emptyset . Les positions où un connecteur vide doit être inséré demandent une étude linguistique qui débouche sur une insertion automatique du connecteur \emptyset demandant le moins possible de révision manuelle (Antolinos-Basso, 2012). Signalons que dans le PDTB, il est demandé aux annotateurs de préciser le sens d'un connecteur implicite en insérant un (ou deux) connecteur(s) explicite(s) ainsi qu'en indiquant la ou les relation(s) de discours exprimée(s). Il semble inutile d'insérer des connecteurs explicites à la place des connecteurs vides : on peut simplement demander aux annotateurs de préciser la ou les relation(s) de discours exprimée(s) (après avoir éventuellement testé si la présence de tel ou tel connecteur explicite n'altère pas la sémantique du texte).

- (4) a. Fred a tiré sur Marie, \emptyset la tuant.
b. Fred a tué Marie, \emptyset en lui tirant dessus.
c. Fred a fait la vaisselle, \emptyset passé l'aspirateur, et lavé les carreaux.

8. Nous ne disposons pas pour le français de la liste des 100 connecteurs les plus fréquents. Certes, on peut effectuer un simple comptage des occurrences des (suites de) mots correspondant aux connecteurs, mais ce comptage ne prendrait pas en compte qu'une même suite de mots peut correspondre à un emploi discursif ou non, voir à ce moment là dans (1). Pour déterminer les 100 connecteurs les plus fréquents du français, il faut disposer d'annotations comme celle du FDTB.

- Supprimer les notions de Sup1 et Sup2. Citons (Joshi *et al.*, 2006) qui argument pour un “principe de minimalité” : “Only as many clauses and/or sentences should be included as are minimally required for interpreting the relation. Any other span of text that is perceived to be relevant (but not necessary) should be annotated as supplementary information : Sup1 for material supplementary to Arg1, Sup2 for material supplementary to Arg2.” Ce principe de minimalité ne peut donner lieu qu’à des interprétations plutôt subjectives, dont la subjectivité est cependant limitée par des consignes relevant de la syntaxe, par exemple : “An argument includes any non-clausal adjuncts, prepositions, connectives, or complementizers introducing or modifying the clause” (Joshi *et al.*, 2006)⁹. La question suivante se pose : quel est le rôle de Sup1 ou Sup2 par exemple dans l’exploitation du PDTB pour les techniques d’apprentissage supervisé reposant sur les données annotées ? Nous n’avons trouvé aucune information sur la question dans les nombreux articles exploitant les résultats du PDTB (Section 2) et nous subodorons que les segments annotés Sup1 et Sup2 sont tout bonnement et simplement ignorés. De plus, Sup1 et Sup2 ne sont pas pris en compte dans l’accord inter-annotateurs sur les empanns des arguments des connecteurs (Prasad *et al.*, 2008). Il semble donc que les notions de Sup1 et Sup2 n’ont été utilisées que pour le confort des annotateurs. C’est une raison valable dans l’objectif d’une couverture partielle que s’est fixé le PDTB, mais qui ne trouve guère de justification dans l’objectif d’une couverture totale où les segments de texte qui auraient été identifiés comme Sup1 ou Sup2 doivent être intégrés dans l’analyse discursive totale du texte.

3.2 Hiérarchie des relations des connecteurs

La hiérarchie des relations de connecteurs établie dans (PDTB Group, 2008, page 27) est discutable, ne serait-ce que parce qu’elle ne semble pas avoir été établie sur la base des relations de discours (et de leur organisation en classes) définie en RST (Mann et Thompson, 1987; Taboada et Mann, 2006) ou SDRT (Asher et Lascarides, 2003), les deux grandes théories sur le discours. Or ces théories présentent des classifications intéressantes sur des dimensions orthogonales :

- RST distingue les relations de discours ‘informationnelle versus présentationnelle’ en s’appuyant sur le fait que certaines relations n’établissent qu’une relation sémantique (informationnelle) entre les contenus propositionnels des ses arguments tandis que d’autres (présentationnelles) demandent de faire appel à des actes de parole. Ainsi la relation de discours lexicalisée par *donc* est informationnelle en (5a), tandis qu’elle est présentationnelle en (5b) puisqu’elle sous-entend l’acte de parole *J’en déduis (donc qu’il a plu)*.
 - (5)a. Il a plu. Donc, les routes sont mouillées.
 - b. Les routes sont mouillées. Donc, il a plu.
- SDRT distingue les relations véridicales des relations non véridicales. La notion de véridicalité y est définie par la formule $R(\alpha, \beta) \Rightarrow K_\alpha \wedge K_\beta$, qui se glose ainsi : si $R(\alpha, \beta)$ est vraie alors les contenus propositionnels des arguments, notés K_α et K_β , sont vrais. Comme montré dans (Danlos et Rambow, 2011), cette définition ne tient pas dès qu’on quitte les assertions de l’auteur, voir (6) avec $\text{NARRATION}(\alpha, \beta)$ où NARRATION , qualifiée de véridicale en SDRT, n’implique pas que K_β soit vrai.

9. Nous sommes actuellement en train de rédiger des consignes analogues pour le français, avec à nouveau des différences par rapport au PDTB. Ainsi, contrairement au PDTB, nous excluons tout connecteur des empanns pour les arguments Arg1 et Arg2 d’un connecteur, ce qui va de pair avec l’objectif d’obtenir une couverture totale.

(6) [Fred ira à Dax pour Noël.]_α Jane pense qu'*ensuite* [il ira à Pau.]_β

Il n'empêche que la définition de véridicalité posée en SDRT tient quand on considère que α et β sont des assertions de l'auteur, et qu'il semble pertinent de distinguer les relations véridicales (e.g. *Narration*) des autres (e.g. *Condition* illustrée dans [Marie viendra à ma fête]_α si [Fred vient]_β.)

Nous avons donc défini une nouvelle hiérarchie des relations de discours qui s'appuie sur les classifications orthogonales des relations de discours définies en RST et SDRT et sur celle du PDTB. Cette hiérarchie, donnée en appendice et décrite dans (Danlos et Roze, 2011), distingue cinq classes de relations véridicales (*CAUSALE*, *TEMPORELLE*, *COMPARAISON*, *EXPANSION* et *ADDITIVE*) et une classe de relations non-véridicales. Dans chaque classe, des sous-classes distinguent éventuellement les relations informationnelles/sémantiques (e.g. *RÉSULTAT*) des relations présentationnelles/pragmatiques (distinguées par le signe * e.g. *RÉSULTAT**). Cette hiérarchie contient 31 feuilles auxquelles il faut ajouter la relation UNKNOWN qui est utile pour certains connecteurs dont le sens est très spécifique (e.g. *au fur et à mesure que* qui s'emploie dans un énoncé comme *Son débit de parole augmentait au fur et à mesure qu'il vidait la bouteille*) ou lorsqu'un annotateur n'est satisfait par aucune relation de la hiérarchie¹⁰.

4 Premières expériences et perspectives futures

Pour nous familiariser avec l'outil d'annotation — ANNOTATOR adapté pour les besoins du FDTB —, identifier les problèmes récurrents d'annotation en vue de rédiger un manuel d'annotation complet et précis (voir note 6), nous avons mené quelques expériences sur certains connecteurs explicites. Ces premières expériences ont montré une grande divergence dans ce qui est annoté Sup1 ou Sup2, ce qui nous conforte dans l'idée d'abandonner tout bonnement et simplement ces notions. Elles ont aussi montré qu'il n'était pas adéquat de travailler connecteur par connecteur. Dans la suite du développement du FDTB, nous demanderons aux annotateurs de travailler article par article, en annotant les connecteurs explicites et implicites avec l'objectif d'une analyse discursive totale.

Nous invitons toute personne intéressée par les connecteurs de discours et/ou l'analyse discursive à nous contacter afin de participer à cette entreprise de longue haleine aux résultats attendus.

Remerciements

Nous remercions M. Djemaa, C. Dordonne, A. Hu, M. Majdoub, L. Patris, C. Ribeyre, F. Sari, et Y. Sun, étudiants de Master2 dans le cursus de Linguistique Informatique de l'Université Paris Diderot, qui ont servi de cobayes (en binôme) pour une annotation expérimentale des connecteurs *afin de*, *lorsque*, *puisque* et *en revanche*. Le FTB comporte aux alentours de 80 occurrences de chacun de ces connecteurs.

10. Un annotateur peut choisir deux éléments de la hiérarchie (mais pas plus que deux, est-ce une limite?) pour un même connecteur (explicite ou implicite). Par exemple, pour *alors que* en (7) il peut choisir *OVERLAP* et *CONTRASTE* sans avoir à trancher arbitrairement entre les deux.

(7) En ce moment, Fred est complètement déprimé, *alors que* Marie est en pleine forme.

Références

- ABEILLÉ, A., CLÉMENT, L. et TOUSSENEL, F. (2003). Building a treebank for french. In ABEILLÉ, A., éditeur : *Treebanks*. Kluwer Academic Publishers, Dordrecht.
- ANTOLINOS-BASSO, D. (2012). Les connecteurs implicites dans le FDTB. Mémoire de Master, Université Paris Diderot.
- ASHER, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.
- ASHER, N. et LASCARIDES, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge.
- DANLOS, L. (2011). Analyse discursive et informations de factivité. In *Actes de TALN 2011*, Montpellier, France.
- DANLOS, L. et RAMBOW, O. (2011). Veridicality of discourse relations and factuality information. In *Proceedings of the fourth workshop on Constraints in Discourse (CID 2011)*, Agay, France.
- DANLOS, L. et ROZE, C. (2011). Hiérarchie des relations de discours dans le FDTB. Rapport technique, ALPAGE, Université Paris Diderot.
- JOSHI, A., PRASAD, R. et WEBBER, B. (2006). Discourse annotation : Discourse connectives and discourse relations. In *Tutorial at the Association for Computational Linguistics*, Sydney, Australia.
- LIN, Z., NG, H. T. et KAN, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT)*, Portland, OR.
- MANN, W. et THOMPSON, S. (1987). Rhetorical structure theory. In KEMPEN, G., éditeur : *Natural Language Generation*, pages 85–95. Martinus Nijhoff Publisher, Dordrecht.
- MARCUS, M., SANTORINI, B., MARCINKIEWICZ, M. A. et TAYLOR, A. (1999). Building a treebank for french. In *Treebank-3*. Linguistic Data Consortium, Philadelphie.
- PDTB GROUP (2008). The Penn Discourse Treebank 2.0 annotation manual. Rapport technique, Institute for Research in Cognitive Science, University of Philadelphia.
- Péry WOODLEY, M.-P., ASHER, N., ENJALBERT, P., BENAMARA, F., BRAS, M., FABRE, C., FERRARI, S., HO DAC, L.-M., Le DRAULEC, A., MATHET, Y., MULLER, P., PRÉVOT, L., REBEYROLLE, J., TANGUY, L., Vergez COURET, M., VIEU, L. et WIDLÖCHER, A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *Proceedings of TALN 2009*, pages 190–196, Senlis, France.
- PITLER, E., LOUIS, A. et NENKOVA, A. (2010). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, Suntec, Singapore.
- PRASAD, R., DINESH, N., LEE, A., JOSHI, A. et WEBBER, B. (2006). Attribution and its annotation in the Penn Discourse Treebank. *Revue TAL*, 47(2).
- PRASAD, R., DINESH, N., LEE, A., MILTSAKAKI, E., ROBALDO, L., JOSHI, A. et WEBBER, B. (2008). The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- ROZE, C. (2009). LEXCONN : Base lexicale des connecteurs discursifs du français. Mémoire de Master, Université Paris Diderot.
- ROZE, C., DANLOS, L. et MULLER, P. (2010). LEXCONN : a French lexicon of discourse connectives. In *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010)*, Moissac, France.
- SAURÍ, R. et PUSTEJOVSKY, J. (2009). FactBank : A corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.
- TABOADA, M. et MANN, W. (2006). Rhetorical Structure Theory : Looking back and moving ahead. *Discourse Studies*, 8(3):423–459.
- WEBBER, B., EGG, M. et KORDONI, V. (2011). Discourse structure and language technology. *Natural Language Engineering*, 1(1):1–54.

Hiérarchie des relations de discours

Classe CAUSALE

- Type sémantique
 - RÉSULTAT
 - EXPLICATION
 - MOYEN
 - BUT
- Type pragmatique
 - RÉSULTAT*
 - EXPLICATION*
 - BUT*
- Type sémantico-pragmatique
 - EVIDENCE

Classe TEMPORELLE

- Type sémantique
 - OVERLAP
 - PRÉCÉDENCE
 - SUCCESSION

Classe COMPARAISON

- Polarités opposées
 - CONTRASTE
 - CONCESSION
 - CONCESSION-AVANT
 - CONCESSION-ARRIÈRE
- Polarités égales
 - PARALLÈLE
- Type pragmatique
 - CONTRASTE*

Classe EXPANSION

- Type Elaboration
 - ELAB-ENTITÉ
 - ELAB-ÉVÈNEMENT
 - ELAB-INSTANCE
 - EXCEPTION
- RÉSUMÉ
- EVALUATION

Classe ADDITIVE

- Type Liste
 - ENUMERATION
 - CONTINUATION
- DISGRESSION

Classe NON VERIDICALE

- Alternatives
 - ALTERNATIVE
 - ALTERNATIVE*
- CORRECTION
- Sémantiques Conditionnelles
 - UNLESS
 - OTHERWISE
 - CONDITION
 - CONSÉQUENCE-HYPOTHÉTIQUE
- Pragmatiques Conditionnelles
 - CONDITION*