FDTB1 : Annotation des connecteurs de discours dans un corpus français

L. Danlos, M. Colinet et J. Steinlin (ALPAGE)

Abstract

This paper presents the annotation of discourse connectives in the corpus FTB (French Treebank) already annotated for morpho- syntax. This is the first step in the full discursive annotation of this corpus. The method consists in projecting on the corpus the items that are listed in LexConn, a lexicon of French connectives, and then filtering the occurrences of these elements that do not have a discursive use. More than 10K connectives have been annotated.

Keywords : discourse connectives, discourse annotation, grammar and discourse

Résumé

Cet article présente l'annotation des connecteurs de discours dans le corpus FTB (French Treebank) déjà annoté pour la morpho-syntaxe. C'est la première étape dans l'annotation discursive complète de ce corpus. Il s'agit de projeter sur le corpus les éléments répertoriés dans LexConn, lexique des connecteurs du français, et de filtrer les occurrences de ces éléments qui n'ont pas un emploi discursif mais par exemple un emploi d'adverbe de manière ou de préposition introduisant un complément sous-catégorisé. Plus de 10 000 connecteurs ont été annotés.

Mots-clés : connecteurs de discours, annotation discursive de corpus, grammaire et discours

1 Introduction

Le projet FDTB (French Discourse Treebank) s'inscrit dans la lignée du projet PDTB, Penn Discourse Treebank (Prasad *et al.* (2008)) qui a consisté à ajouter une couche d'annotation discursive sur le PTB (Penn Treebank), corpus composé d'articles du *Wall Street Journal*, déjà annoté en morpho-syntaxe. De même, le projet FDTB consiste à ajouter une couche d'annotation discursive sur le FTB, French Treebank (Abeillé *et al.* (2003)), corpus composé d'articles du journal *Le Monde* annoté en morpho-syntaxe.

L'annotation complète du PDTB ou FDTB consiste à repérer les connecteurs ("explicites" et "implicites" [1]), et à annoter leurs sens et leurs arguments. Des expériences préliminaires d'annotation du FDTB (Danlos *et al.* (2012)) ont montré qu'il était difficile d'effectuer toutes ces opérations en une seule passe car

1. Un connecteur implicite n'est pas réalisé : c'est le connecteur vide qu'on observe entre deux phrases simplement juxtaposées dans une parataxe. A l'inverse, un connecteur explicite est un item lexical non vide.

de nombreux items lexicaux sont ambigus entre un emploi comme connecteur de discours et un emploi non discursif ; par exemple, la conjonction de coordination *et* est connecteur en (1) et non-connecteur en (2) :

(1) Paul a fini d'écrire son article et il est parti en vacances.

(2) Paul et Marie sont très bons amis.

Des adverbes comme *ainsi* ou *alors* présentent une ambiguïté similaire comme décrit respectivement dans (Molinier (2013); Karssenberg & Lahousse (2014)) et (Bras (2008); Degand & Fagard (2011)). De ce fait, il est apparu qu'il valait mieux effectuer l'annotation du FDTB en commençant par une première étape, appelée FDTB1, qui consiste uniquement à repérer tous les connecteurs de discours du corpus. C'est cette étape que nous présentons ici. Signalons que l'annotation du PDTB n'est pas passée par cette première étape : seuls les 100 connecteurs anglais considérés comme les plus fréquents ont été annotés. Nous ignorons comment la fréquence des connecteurs anglais a été déterminée vu l'ambiguïté dont nous venons de parler et nous ne pouvons déterminer la fréquence des connecteurs français qu'après une étude comme celle menée ici : l'annotation décrite ici permettra de déterminer les 100 connecteurs français les plus fréquents, au moins dans un corpus journalistique comme *Le Monde*, voir Table 7 à la Section 7 [2].

Les connecteurs de discours sont des items lexicaux qui permettent d'exprimer explicitement les relations discursives entre deux segments du discours, "élémentaires" ou "complexes" [3]. Le travail effectué dans le FDTB1 s'appuie sur LexConn (Roze *et al.* (2012)), un lexique qui recense de la manière la plus exhaustive possible les connecteurs du français. Cette première version de LexConn compte 325 items qui sont listés avec leur catégorie syntaxique et la ou les relations de discours qu'ils peuvent exprimer, ce qui débouche sur 430 sens de connecteurs. Les catégories syntaxiques sont au nombre de quatre : conjonctions de coordination, conjonctions de subordination, prépositions (introduisant un VP à l'infinitif ou au participe présent) et adverbiaux (catégorie qui regroupe principalement des adverbes simples et des syntagmes prépositionnels).

Les principes qui ont guidé l'élaboration de LexConn sont résumés à la Section 3. Un de ces principes est qu'un segment de discours élémentaire doit comporter un syntagme verbal VP (à temps fini ou non). Ce principe a éliminé de LexConn des prépositions comme *à cause de* ou *en raison de* qui ne peuvent introduire que des syntagmes nominaux (SN). Ce principe a aussi été appliqué dans le FDTB1 : les éléments de LexConn qui n'ont pas portée sur un VP dans le corpus ont été éliminés automatiquement et n'ont pas été projetés dans le FDTB1 ; à titre d'illustration, pour la préposition *pour*, seules ses occurrences introduisant un VP à l'infinitif ont été projetées dans la FDTB1,

2. Nous avons aussi annoté, en suivant les principes du FDTB1, une sous-partie du corpus Sequoia (Candito & Seddah (2012)) qui a été annoté pour la morpho-syntaxe en suivant les mêmes principes que le FTB mais sur des textes d'un genre différent, à savoir des textes de Wikipedia.fr et du journal *L'Est Républicain*. Malheureusement, nous n'avons pu annoter qu'une sous-partie de Sequoia car le reste du corpus est un ensemble de phrases isolées qu'on ne peut pas annoter discursivement. La partie annotée de Sequoia étant 20 fois plus petite que le FDTB1, nous ne sommes pas en mesure de donner des fréquences significatives pour les connecteurs de discours sur des corpus de genres différents.

3. Un segment de discours est complexe s'il couvre plusieurs segments élémentaires contigus reliés eux-mêmes par des relations discursives.

en excluant les *pour SN*. Un tel filtrage automatique bénéficie de l'annotation morpho-syntaxique du FTB et s'effectue avec l'outil Tregex (Levy & Andrew (2006)).

Après avoir projeté automatiquement dans le FTB les éléments de LexConn respectant le principe ci-dessus, le travail du FDTB1 se ramène à une tâche de désambiguïsation : les occurrences des éléments de LexConn projetées dans le corpus sont-elles employées comme connecteurs ou non ? Les tâches de désambiguïsation ont été les suivantes :

- désambiguïsation morpho-syntaxique (Section 4), par exemple pour les homonymes comme *bref* qui peut être un adjectif ou un adverbe connecteur,
- désambiguïsation entre grammaire et discours (Section 5) pour les adverbiaux qui peuvent avoir un emploi comme connecteur et un emploi d'ajout ou de complément à l'intérieur de leur phrase hôte,
- désambiguïsation entre grammaire et discours (Section 6) pour les prépositions (et quelques conjonctions de subordination) qui peuvent avoir un emploi comme connecteur et un emploi de complément sous-catégorisé par un élément (verbal, nominal, adjectival ou adverbial) de leur phrase hôte.

La première tâche de désambiguïsation peut être faite automatiquement grâce à l'annotation morpho-sytaxique du corpus, les deux autres sont délicates car elles touchent à des questions difficiles concernant la notion de connecteur de discours.

Mais avant d'entrer dans le cœur de la présentation du FDTB1, il est utile de présenter la suite de ce projet, c'est-à-dire l'annotation complète du FDTB. En effet, il faut garder à l'esprit que, même si le FDTB1 débouche sur un corpus annoté qui est déjà une ressource pouvant être utile à différents chercheurs, ce n'est qu'une première étape qui a été faite avec un objectif plus large pouvant avoir guidé certains choix.

2 Annotation complète du FDTB

Pour arriver à une annotation discursive complète à partir du FDTB1, trois tâches seront à effectuer :

1) annotation du sens et des arguments des connecteurs "explicites" repérés dans le FDTB1,
2) identification des "AltLex" et des connecteurs "implicites",
3) annotation du sens et des arguments des éléments identifiés à l'étape 2.

La première tâche sera effectuée dans l'esprit du PDTB, avec quelques modifications mineures (Danlos (2012)) concernant la hiérarchie des sens de connecteurs et l'annotation de leurs arguments, notés Arg1 et Arg2 où Arg2 est le segment hôte du connecteur [4]. Nous allons donner quelques précisions sur les AltLex et connecteurs implicites, avant d'aborder le sens les connecteurs et la détermination de leurs arguments.

4. Précisément, l'Arg2 d'un connecteur adverbial contient le VP dans lequel il apparaît, l'Arg2 d'une conjonction ou préposition contient le segment introduit par la conjonction ou préposition.

2.1 AltLex et connecteurs implicites

La seconde tâche identifiée ci-dessus a pour objectif de connecter une phrase (typographique) dans laquelle aucun connecteur explicite n'a été repéré dans le FDTB1 au reste du discours, le plus souvent à la phrase précédente par un principe d'adjacence. Néanmoins, il a été souligné dans divers travaux qu'une phrase sans connecteur explicite peut se voir relier à son contexte gauche par des items lexicaux n'appartenant pas à la catégorie des connecteurs de discours et qui ont été baptisés AltLex (Alternative Lexicalization) dans le PDTB. Illustrons sur des exemples : en (3)a, le connecteur explicite *parce que* lie les deux propositions avec un sens causal. En (3)c, le lecteur doit inférer que les deux phrases sont reliées par une relation causale : on parle alors de "connecteur implicite", noté ø. A l'intermédiaire, (3)b ne comporte pas de connecteur explicite mais le lecteur ne doit faire aucune inférence : le fait que le contenu de la proposition *Fred a mal dormi* explique le contenu de *Fred est de mauvaise humeur* est explicitement indiqué par la séquence *Ceci est dû au fait que* qui se voit attribuer le statut de AltLex. Soulignons que cette séquence est compositionnelle et qu'elle ne saurait en aucun cas faire partie d'un lexique de connecteurs.

(3) a. Fred est de mauvaise humeur parce qu' il a mal dormi.
b. Fred est de mauvaise humeur. Ceci est dû au fait qu' il a mal dormi.
c. Fred est de mauvaise humeur. ø Il a mal dormi.

Le PDTB décrit quelques cas d'AltLex pour l'anglais, les définit par le fait qu'"on ne peut pas leur ajouter de connecteur sans produire un effet de redondance". En ce qui concerne le français, c'est un vaste champ d'étude sur lequel aucune ligne n'a encore jamais été écrite sauf sur les "verbes de discours" (Danlos (2006)), mais il nous semble qu'une définition reposant sur une absence d'inférence par le lecteur soit préférable à une définition reposant sur un effet de redondance, la redondance n'étant pas exclue dans la langue et éventuellement non perçue[5]. Nous pouvons toutefois déjà signaler qu'il semble exister un continuum entre AltLex et connecteur de discours, continuum qui reflète un processus de grammaticalisation (une étude dans ce sens a été mené par (Rysová & Rysová (2014) sur le Tchèque). Ainsi, nous avons parfois hésité entre le statut de connecteur ou de AltLex. C'est le cas pour :

- les expressions *résultat, la preuve, remarque* qui ont été enregistrées dans LexConn sous la catégorie adverbiale, bien que ce ne soit ni des adverbes ni des SP. Finalement, nous avons choisi de les considérer comme des connecteurs,
- les anaphores temporelles : par exemple, en (4), *alors* est une anaphore de l'expression temporelle *de 1870 à 1914* située dans la phrase précédente. Cet adverbe n'introduit donc pas à proprement parler de relation de discours entre les deux propositions, (Bras (2008)). Il pourrait donc être considéré comme un AltLex mais nous l'avons annoté comme connecteur.

5. Ainsi, la requête sur Google "Cela a ensuite été suivi" avec deux marqueurs (redondants) de la relation de précédence temporelle, à savoir le connecteur *ensuite* et le verbe de discours *suivre*, ramène aux alentours de 22 800 résultats, comme le texte suivant qui n'est pas vraiment perçu comme redondant : *L'excitation a commencé vendredi après un très laconique annonce de quatre lignes par la FINMA. Cela a ensuite été suivi de certains reportages à la fois par ... ...*

(4) De 1870 à 1914, plus de cinq cents millions d'images en cinquante langues sortent des bâtiments du quai de Dogneville, à Epinal. L'imagerie compte alors jusqu'à cent cinquante salariés.

Lorsque la matrice d'une phrase typographique ne comporte aucun connecteur explicite et aucun AltLex, un connecteur implicite est positionné [6]. Les AltLex et les connecteurs implicites se voient attribuer un sens et deux arguments comme les connecteurs explicites.

2.2 Sens des connecteurs

Il n'est pas question ici de présenter et discuter les différents sens attribués aux connecteurs (explicites et implicites) et AltLex. Cependant nous souhaitons attirer l'attention sur le point suivant. Dans le FDTB1, certains éléments ont été repérés comme connecteurs de discours de façon discutable. On pense en particulier aux conjonctions de coordination *et* et *ou* : toutes les occurrences de ces conjonctions qui coordonnent des éléments phrastiques ou VP ont été projetées dans le FDTB1 et considérées comme connecteurs puisqu'elles ne sont l'objet d'aucune des tâches de désambiguïsation listées à la Section 1. C'est ainsi que *et* est considéré comme connecteur dans (1) et comme non-connecteur dans (2). Par le même principe, ces conjonctions sont considérées comme connecteurs dans les exemples suivants où elles coordonnent respectivement deux compléments phrastiques ou à l'infinitive.

(5) C'est un mensonge qui ne coûte pas cher, je ne sais pas si cela vient de lui ou si on le lui a suggéré.

(6) Tass affirme aussi que les dirigeants du front populaire local ont appelé à expulser tous les Arméniens vivant encore à Bakou, la capitale azerbaidjanaise, et à lancer une "marche sur le Haut-Karabakh".

Le rôle discursif de ces conjonctions en (5) et (6) est tout à fait douteux : elles semblent jouer uniquement un rôle syntaxique et sémantique au sein de la phrase. Néanmoins, il sera possible de leur attribuer des sens spécifiques, disons OR et AND, qui les distinguent d'autres emplois où ces conjonctions ont clairement un rôle discursif. Autrement dit, l'annotation complète du FDTB permettra de filtrer, grâce aux sens attribués aux connecteurs, des éléments annotés comme connecteurs dans le FDTB1 mais que certains chercheurs ne souhaitent pas considérer comme connecteurs.

2.3 Arguments des connecteurs

En ce qui concerne les arguments des connecteurs, rappelons que nous imposons la contrainte qu'un segment de discours élémentaire comporte un VP. Cette contrainte vient de raisons théoriques — e.g. éviter de considérer la préposition *à cause de* qui introduit un SN comme un connecteur de discours — et pratiques — e.g. limitation du travail d'annotation dans le FDTB1. Cette contrainte fait que *jusqu'à* en (7) n'est pas considéré comme connecteur car son Arg2 serait un SN, alors que *jusqu'à ce que* en (8) est considéré comme connecteur. De même,

6. Des connecteurs implicites peuvent être aussi positionnés à l'intérieur d'une phrase, mais nous laissons ce sujet de côté ici.

cependant n'est pas considéré comme connecteur en (9) car son Arg2 ne contient pas de VP.

(7) Début 1986, Mr Seydoux lance aux côtés de Mr Silvio Berlusconi la première version de la Cinq, qui vivra jusqu'à l'arrivée de Mr Jacques Chirac à Matignon.

(8) Début 1986, Mr Seydoux lance aux côtés de Mr Silvio Berlusconi la première version de la Cinq, qui vivra jusqu'à ce que Mr Jacques Chirac arrive à Matignon.

(9) Depuis 1986, le dernier grand hiver qu'ait connu la Suède, la mission urbaine a ouvert des auberges de nuit et même de jour moins cependant pour les femmes que pour les hommes.

Les arguments qui comportent un VP à temps fini incluent les relatives qui peuvent comporter des connecteurs adverbiaux, (10). Signalons que les relatives qui ne comportent pas de connecteurs de discours n'ont pas été prises en compte même si elles ont un rôle d'"eliciture" (variante de l'implicature) comme en (11) traduit de (Kehler *et al.* (2015)) où le contenu de la relative peut être perçu comme la cause du renvoi.

(10) La CFDT et la CFTV bénéficient dans l'ensemble de l'érosion de la CGT, qui reste pourtant bien au dessus de la moyenne nationale.

(11) Le directeur a renvoyé l'employé qui avait détourné de l'argent.

Les VP à temps non fini regroupent les infinitives, e.g. Arg2 de *sans* en (12), participes présents, e.g. Arg2 de *cependant* en (13) et participes passés, e.g. Arg1 de *pourtant* en (14).

(12) Il a expliqué qu'il apporterait son soutien à des candidats dans toutes les circonscriptions, sans se présenter en tant que parti.

(13) Luc a présenté sa candidature pour les municipales, ne cachant cependant pas son dégoût de la vie politique.

(14) Dégoûté par la vie politique, Luc a pourtant présenté sa candidature pour les municipales.

La seule exception à la contrainte sur les arguments concerne les phrases averbales[7] : ainsi en (15) *donc* est considéré comme un connecteur dont l'Arg2 est la phrase averbale *Une véritable tragédie pour lui.* Notons que si *donc* n'était pas considéré comme un connecteur dans le FDTB1, un connecteur implicite serait alors positionné pour relier les deux phrases typographiques lors de la deuxième tâche du FDTB, ce qui est inutilement compliqué.

(15) Luc a dû démissionner. Une véritable tragédie donc pour lui.

7. Rappelons qu'une phrase averbale est une phrase typographique — encadrées de deux ponctuations fortes — ne comportant pas de verbe matrice.

3 Présentation de LexConn

LexConn est un lexique qui répertorie les connecteurs de discours du français (Roze *et al.* (2012)). Ce sont des marqueurs lexicaux explicites utilisés pour exprimer les relations sémantiques ou pragmatiques entre segments de discours. LexConn (dans sa première version de 2012) en compte 325 qui sont listés avec leur catégorie syntaxique et la ou les relation(s) de discours qu'ils expriment. Les catégories grammaticales des connecteurs sont les suivantes : les conjonctions de coordination (cco) comme *et*, *ou* et *mais* ; les conjonctions de subordination (csu) comme *parce que*, *même si* et *tandis que* ; les prépositions (prep) comme *pour*, *afin de* et *avant de* ; les adverbiaux (adv), comme *néanmoins* qui est un adverbe simple et *en tout cas* qui est un groupe prépositionnel (SP) utilisé comme adverbial.

3.1 Les prépositions et les conjonctions

La compilation des prépositions et conjonctions de LexConn s'est faite à partir de ressources lexicales existantes. Il y a 65 prépositions dans LexConn, en mettant de côté *à* qui a un emploi discursif rare et vieilli (*À voler bas, au lieu de signaux d'amitié, on récolte des explosions de torpilles.*). Les prépositions sont toutes suivies d'une infinitive sauf *en* et *tout en* qui sont suivies d'un participe présent, en formant un gérondif, comme en (16) où le gérondif indique comment la porte d'entrée a été cassée[8].

(16) Paul a cassé la porté d'entrée, **en l'ouvrant violemment**.

Cinq prépositions (*pour* (+infinitive), *afin de*, *plutôt que de*, *jusqu'à* (+ infinitive) et *avant de*) peuvent avoir des emplois non-connecteurs, ce qui est l'objet de la Section 6 qui discutera aussi des conjonctions de subordination *pour que, afin que* et *avant que*. Sinon les autres conjonctions de subordination de LexConn ont systématiquement été annotées comme connecteurs dès qu'elles introduisent un complément phrastique — *parce que* n'est pas considéré comme connecteur dans *Paul n'est pas venu, parce que malade.* — et qu'elles sont bien employées comme conjonction de subordination — *quand* n'est pas une conjonction de subordination dans *Paul se demande quand Marie viendra*, voir Section 4.

De même, les conjonctions de coordination de LexConn ont été annotées comme connecteurs dès qu'elles coordonnent des éléments phrastiques ou des VP (voir Section 2).

Les prépositions et les conjonctions de subordination peuvent être modifiées par des adverbes (e.g. *probablement*) ou des incises (e.g. *paraît-il*) qui apparaissent à gauche ou à droite du connecteur ou même à l'intérieur de celui-ci quand c'est une forme composée :

(17) a. Paul s'est excusé probablement **pour que** Marie...
b. Paul s'est excusé **pour que** probablement Marie...
c. Paul s'est excusé **pour** probablement **que** Marie accepte de lui prêter de l'argent.

8. Les équivalents anglais des gérondifs (les formes *by V-ing*) n'ont pas été annotés comme connecteurs dans le PDTB.

Les modifieurs de connecteurs n'ont pas été répertoriés dans LexConn mais ils sont annotés dans le FDTB1 dans la mesure où ils sont intégrés dans l'empan du connecteur.

3.2 Les adverbiaux

La tâche la plus délicate de LexConn a été de déterminer parmi les adverbiaux (adverbes et syntagmes prépositionnels) ceux qui pouvaient jouer le rôle de connecteur, une telle liste n'existant pas auparavant.

Le premier critère pour qu'un adverbial soit connecteur est que sa phrase hôte ait un contexte gauche. Aucun discours, écrit ou oral, ne peut commencer par une phrase contenant un adverbial connecteur : par exemple, la phrase *Du coup, Paul est de mauvaise humeur* est interdite à l'initiale d'un discours, ce qui est un indice que *du coup* est connecteur. Comme ce critère n'est pas suffisant dans la mesure où la phrase *Le matin suivant, Paul est arrivé* est aussi interdite à l'initiale d'un discours bien que *le matin suivant* soit un AltLex de localisation temporelle et non un connecteur, d'autres critères ont été mis en avant lors du développement de LexConn. Nous allons en présenter cinq : le clivage, la substituabilité, la variabilité et la modification qui sont des critères sémantiques et le critère de cohérence qui relève plutôt du discours. Ces critères sont aussi utilisables dans le FDTB1 pour déterminer si une occurrence d'un adverbial de LexConn est employée comme connecteur ou pas et pour différencier les connecteurs des AltLex.

Critère de clivage

Un connecteur adverbial n'est pas intégré au contenu propositionnel de sa phrase hôte, ce qui se manifeste entre autres par le fait qu'il ne peut pas être clivé. Ainsi l'adverbial *à ce moment-là* peut être clivé en (18) ce qui indique que ce n'est pas un connecteur, tandis qu'il ne peut pas être clivé en (19) ce qui indique que c'est un connecteur.

(18) a. Il a commencé à pleuvoir. A ce moment-là , Pierre est arrivé.
b. Il a commencé à pleuvoir. C'est **à ce moment-là** que Pierre est arrivé.

(19) a. Tu as l'air de penser qu'elle n'est pas honnête. A ce moment-là , ne lui raconte rien.
b. Tu as l'air de penser qu'elle n'est pas honnête. #C'est **à ce moment-là** que tu ne lui racontes rien. (Roze *et al.* (2012))

Critère de substituabilité

Ce critère concerne les adverbiaux qui contiennent une pro-forme. Lorsque la pro-forme réfère à un segment de discours du contexte gauche, autrement dit lorsqu'elle est anaphorique, l'adverbial est considéré comme non-connecteur, voir *après ça* en (20). Sinon, il est considéré comme connecteur, voir (21) :

(20) a. Bruno est allé en Argentine. **Après ça**, il est allé au Pérou.
b. Bruno est allé en Argentine. **Après** [***qu'il est allé en Argentine***], il est allé au Pérou.

(21) a. Hier soir j'ai croisé Pierre dans un bar. **À part ça**, il nous dit tout le temps qu'il est fatigué.
b. #Hier soir j'ai croisé Pierre dans un bar. **À part [*qu'hier soir je l'ai croisé dans un bar*]**, il nous dit tout le temps qu'il est fatigué. (Roze *et al.* (2012))

Critère de compositionalité : variabilité et modification

Ce critère concerne les adverbiaux qui contiennent un nom. L'hypothèse est que lorsque ces adverbiaux sont employés comme connecteurs, ils ont subi un processus de grammaticalisation et perdu de leur compositionalité : d'une part le nom n'est plus variable en nombre, voir (22), d'autre part il n'est pas modifiable, voir (23) :

(22) Tu dis que tu ne la trouves pas honnête. **A ce moment-là/*ces moments-là**, ne lui fais pas confiance.

(23) Je ne sais plus s'il y avait vraiment de la neige, ce Noël-là. **En tout cas (*envisagé/possible)**, dans mon souvenir, je la vois tomber la nuit, à gros flocons, sur la route et les écuries. (Roze *et al.* (2012))

Critère de cohérence

Lorsque deux phrases adjacentes simplement juxtaposées sont incohérentes et que l'insertion d'un adverbial rend le discours cohérent, alors l'adverbial est connecteur, voir (24) :

(24) Ce serait vraiment utile pour nous d'aller à cette réunion.
a. **Ceci dit**, on peut s'en passer.
b. #On peut s'en passer. (Roze *et al.* (2012))

Les cinq critères que nous venons de rappeler ont servi à trier parmi l'ensemble des adverbiaux ceux qui pouvaient avoir un emploi de connecteur et donc figurer dans LexConn. Ils n'ont pas toujours suffi pour désambiguïser les occurrences de ces adverbiaux dans le FDTB1. C'est pourquoi la Section 4 présente au cas par cas une description des ambiguïtés des adverbiaux potentiellement connecteurs.

Terminons en indiquant que LexConn s'est voulu aussi exhaustif que possible mais que l'annotation effectuée pour le FDTB1 a mis en évidence 29 connecteurs non répertoriés dans LexConn (et nous en avons supprimé 3). Nous arrivons donc à un total de 353 connecteurs (sans tenir compte du fait qu'un connecteur peut avoir plusieurs sens). Une nouvelle version de LexConn est disponible à l'adresse https://gforge.inria.fr/frs/?group_id=6145. Cette version enregistre aussi 76 variantes des connecteurs présents dans le lexique, en particulier des connecteurs accompagnés d'un modifieur.

4 Ambiguïtés morpho-syntaxiques

Le premier aspect de la désambiguation dans le FDTB1 consiste, pour chaque occurrence d'item qui peut être connecteur, à décider si elle correspond morpho-syntaxiquement au connecteur recherché. Cet aspect de la désambiguïsation ne

présente pas d'aspect novateur, cependant, il est essentiel si on veut bénéficier d'une pré-annotation automatique des candidats connecteurs dans un corpus annoté pour la morpho-syntaxe.

Le premier cas d'ambiguïté morpho-syntaxique est celui des homonymes, c'est-à-dire d'un mot qui ne correspond pas au token visé parce qu'il relève d'une autre catégorie grammaticale, par exemple le mot *car* qui peut-être une conjonction de coordination répertoriée dans LexConn, ou un nom commun, auquel cas il n'est pas connecteur, voir (25).

(25) Le **car** a roulé très vite.

Le second cas correspond à celui d'une suite de mots qui a été répertoriée comme connecteur dans LexConn mais qui peut correspondre à d'autres catégories morpho-syntaxiques. Par exemple, la suite de mots *en fait* est répertoriée comme connecteur adverbial dans LexConn mais elle peut correspondre à un pronom (Pro) suivi d'un verbe (V), voir (26).

(26) La Grand-Place était piétonne. Le maire **en fait** un parking.

La Table 1 dresse la liste des éléments de LexConn qui présentent une ambiguité morpho-syntaxique. Cette liste se vaut aussi exhaustive que possible mais il se peut que certains cas aient été oubliés, en particulier pour les éléments de LexConn non présents dans le FDTB1.

TABLE 1 – Eléments de LexConn qui présentent une ambiguïté morpho-syntaxique

CONN	**Cat** de LexConn	**Autre cat/fonction**	Exemple non connecteur
à l'inverse	adv	SP adjectival	Nos positions sont **à l'inverse**.
alors que	csu	adv + C	Il savait alors qu'il deviendrait dentiste
à la place	adv	SP	Cette avenue mène **à la place** de la Bastille.
à la réflexion	adv	SP	Je repense **à la réflexion** qu'il m'a faite.
au fait	adv	SP	Je me raccroche **au fait** que c'est bientôt fini.
avant de	prep	adj + prep	Le passager avant de la voiture
bien que	csu	adv + C	Il sait **bien que** je suis là
bref	adv	adj	Son passage fut **bref**.
car	cco	N	Le **car** a roulé très vite.
du coup	adv	SP	Au moment **du coup** de tonnerre...
d'un coup	adv	SP	Il a suffi **d'un coup** pour l'assommer.
du reste	adv	SP	Occupe-toi de ça, je m'occupe **du reste**.
d'ailleurs	adv	SP locatif	Il n'est pas d'ici mais vient **d'ailleurs**.
en fait	adv	Pro + V	La place est piétonne. Le maire **en fait** un parking.
en gros	adv	SP adjectival	Il se lance dans la vente **en gros**.
en plus	adv	SP	Il est de plus **en plus** grand.
le temps de		SN	Il a pris **le temps de** s'excuser.
maintenant	adv	V-ant	Tout en **maintenant** votre tête en arrière...
or	cco	N	L'**or** a perdu de sa valeur.
par ailleurs	adv	SP locatif	Je ne l'ai pas vu ; il a dû passer **par ailleurs**.
parallèlement	adv	adv (à)	La ligne des montagnes court **parallèlement** à la Loire.
quand	csu	Pro	Il peut y aller n'importe **quand**.
		C	Il se demande **quand** Luc viendra
remarque	adv	N	Sa **remarque** était pertinente.
résultat	adv	N	Le **résultat** est très mauvais.
si	csu	C	Je me demande **si** elle viendra.
		adv	Elle est **si** grande.
soudain	adv	adj	Le choc a été **soudain**.
tant que	csu	adv + C	Ce n'est pas **tant qu'**il est grand...

5 Les adverbiaux entre grammaire et discours

Le deuxième aspect de la désambiguïsation dans le FDTB1 consiste à distinguer les adverbiaux de LexConn qui ont une fonction discursive entre deux segments du discours de ceux qui ont une fonction syntaxique et un rôle sémantique à l'intérieur de leur phrase hôte. Dans les termes de (Molinier & Lévrier (2000)), ceci s'approche de la distinction entre les adverbiaux 'de phrase" et les autres adverbiaux.

Cet aspect de la désambiguïsation s'appuie sur les critères généraux présentés à la Section 2.2. Nous n'avons pas trouvé d'autres critères permettant de savoir au cas par cas si un adverbial potentiellement connecteur est effectivement employé comme connecteur en contexte. Nous nous contentons donc de lister pour un certain nombre d'adverbiaux ambigus assez délicats un emploi comme connecteur — en donnant un aperçu de la ou les relations de discours

qu'il lexicalise, généralement grâce à un exemple issu du FDTB1 — et un emploi comme non connecteur — en essayant de préciser sa fonction à l'intérieur de sa phrase hôte. Le reste des adverbiaux ambigus est plus simplement listé dans la Table 2. Enfin, les adverbiaux qui sont toujours employés comme connecteurs (au moins dans le FTB) sont listés dans la Table 3.

Au contraire

Au contraire a été annoté comme connecteur lorsqu'il lexicalise un contraste, (27), entre sa phrase hôte et son contexte gauche :

(27) Selon cette enquête, 15 % se prononcent pour un arrêt rapide du programme nucléaire français, 22 % sont au contraire favorables à sa poursuite et à la construction de nouvelles centrales.

Au contraire n'est pas retenu comme connecteur, lorsqu'il est employé pour renforcer une assertion négative, (28) :

(28) La nouvelle diminution du taux d'escompte de la Banque du Japon n'a nullement déprimé la monnaie japonaise, au contraire .

Ainsi

Ainsi a été annoté comme connecteur lorsqu'il lexicalise une relation de résultat ou d'exemplification, comme en (29) sans inversion de l'ordre canonique sujet-verbe ou en (30) avec inversion (Molinier (2013); Karssenberg & Lahousse (2014)) :

(29) La Commission nationale des comptes de campagne et des financements politiques se limite à vérifier si les obligations comptables et financières sont remplies. Ainsi , il n'existe à ce jour aucun contrôle des dépenses des partis

(30) M. Hockey ne mâche pas ses mots. Ainsi a-t-il invité les pays émergents à «se sevrer de la morphine de l'argent facile et à engager des réformes».

Ainsi n'est pas connecteur lorsqu'il est utilisé comme anaphore de manière, (31), ou comme anaphore ou cataphore d'un discours rapporté, (32) :

(31) Luc s'est comporté ainsi parce qu'il était fatigué.

(32) M. Michel Charasse, ministre du budget, a ainsi déclaré au micro de RMC : "C'est une affaire privée, et je ne vois pas pourquoi les pouvoirs publics seraient impliqués là-dedans".

D'abord

D'abord a été annoté comme connecteur lorsqu'il introduit une relation de précédence entre l'événement dénoté par sa phrase hôte et un événement évoqué dans son contexte droit, comme en (33) où il fonctionne de façon appairée avec l'expression *après un temps de réflexion* :

(33) il vous dira d'abord qu'il ne se sent pas capable, mais si vous lui dites que le groupe l'a jugé digne de cette confiance, il acceptera après un temps de réflexion, et son action dépassera vos propres espérances.

D'abord n'a pas été retenu comme connecteur lorsqu'il sert à mettre un fait au premier plan, comme en (34) où *d'abord* est paraphrasable par *avant tout* :

(34) Manger bio est d'abord une question de bon sens.

De fait (effectivement)

De fait a été annoté comme connecteur dans un exemple comme (35) :

(35) La corruption n'est pas l'apanage des pays en voie de développement, bien au contraire.
De fait, sur les 427 affaires de corruption transnationales survenues dans le monde depuis février 1999, 128 ont été sanctionnées aux États-Unis.

De fait n'a pas été retenu comme connecteur lorsqu'il est antonyme de *en théorie* ou *en principe*.

(36) Après les dévaluations de ces dernières [la livre britannique et la lire italienne], le franc s'est trouvé, de fait, réévalué.

De même (de la même façon)

De même a été annoté comme connecteur lorsqu'il établit un parallèle entre sa phrase hôte et son contexte gauche, (37) :

(37) Une application stricte de la nouvelle loi fait craindre à certains une limitation de l'accès à l'information pour le grand public. De même, elle pourrait interdire la possibilité de consulter des sites de partage de vidéos comme YouTube.

De même n'a pas été retenu comme connecteur lorsqu'il est utilisé comme un adverbe de manière, (38), ou qu'il apparaît dans l'expression *il en est de même*, (39) :

(38) Si vous en usez bien, il en usera de même.

(39) Il en est de même, quoique dans une moindre mesure, pour les pays d'Amérique latine.

(Tout) d'un coup (tout à coup)

Tout d'un coup ou *d'un coup* sont connecteurs lorsqu'ils introduisent une rupture dans la narration entre l'événement énoncé dans leur phrase hôte et les événements ou la situation énoncés dans le contexte gauche, (40) :

(40) L'Europe a été longtemps une région de départ. Tout d'un coup, elle devient une terre non seulement d'immigration mais de peuplement.

D'un coup n'est pas connecteur lorsqu'il est utilisé comme un adverbe de manière et exprime le fait que la survenance de l'événement dénoté par la phrase est soudaine, (41) :

(41) Et si d'aventure on venait à suivre leurs dangereuses recommandations, on ruinerait d'un coup l'acquis de six années ...

Enfin (finalement)

Enfin est connecteur lorsqu'il introduit une conclusion à une énumération d'événements, comme en (42) :

(42) L'UE ferme ses marchés financiers à la Russie. Elle interdit de nouvelles ventes d'armes et matériels militaires à Moscou. Enfin, l'Union limite considérablement les exportations de technologies de pointe dans les domaines de l'exploitation pétrolière et gazière.

Enfin n'a pas été retenu comme connecteur lorsqu'il est utilisé avec une valeur affective pour exprimer la fin d'une longue attente, (43) :

(43) Le dessinateur Katsuhiro Otomo est enfin consacré à Angoulême.

Inversement

Inversement est annoté comme connecteur lorsqu'il lexicalise une relation de contraste, (44) :

(44) Les écoles d'arts attirent peu d'étudiants étrangers : ils ne sont que 15 % aux Beaux-Arts de Paris ou de Nîmes et pratiquement absents ailleurs. Inversement, les jeunes artistes français sont peu tentés de passer la frontière.

Inversement n'a pas été retenu comme connecteur lorsqu'il est employé dans l'expression figée *et inversement* ou *ou inversement* qui marque le fait que les deux termes d'une relation binaire sont interchangeables, (45). *Inversement* n'est pas non plus connecteur, lorsqu'il est modifieur d'un adjectif, (46).

(45) Qui a dit que les écrans détournaient les jeunes de l'imprimé ? Les gros succès en librairie font des cartons au box-office et inversement.

(46) Bien sûr, le taux d'équipement est inversement proportionnel à l'âge de la construction ...

Parallèlement (simultanément, réciproquement, également)

Parallèlement a été annoté comme connecteur lorsqu'il introduit une relation de concomitance entre deux événements d'une même narration, (47) :

(47) En 1983, il est élu vice-président du CRIF national, qu'il présidera de 1989 à 1995.
Parallèlement, il est élu en 1991 président du Consistoire israélite du Bas-Rhin.

Parallèlement n'a pas été considéré comme connecteur lorsqu'il est modifieur d'un prédicat pluriel qui décrit plusieurs événements qui ont lieu de façon concomitante, (48) :

(48) Les autorités bancaires allemande, américaine, britannique, singapourienne et suisse ont ouvert parallèlement des enquêtes sur ce scandale.

Plutôt

Plutôt est connecteur lorsqu'il lexicalise une relation de contraste entre sa phrase hôte et son contexte gauche :

(49) C'est absurde ! Il n'a trahi personne, il a plutôt été trahi par des membres de son parti

Plutôt n'a pas été retenu comme connecteur lorsqu'il est employé comme adverbe de manière associé à un prédicat gradable pour le nuancer, comme en (50) :

(50) Ces derniers temps, elle a plutôt embelli.

Plutôt est plus souvent utilisé, avec le même sens, suivi d'un adjectif *Il est* ***plutôt*** *heureux* ou d'un adverbe *On vit* ***plutôt*** *bien à Paris.*

Quand même (tout de même)

Quand même est connecteur lorsqu'il lexicalise une opposition :

(51) Ssangyong, 600 à 800 voitures en 1993, 1500 à 2000 en 1994. Une arrivée sur la pointe des pieds afin d'éviter les réactions de rejet.
Les voitures du pays du Matin calme pourraient quand même bien, d'ici quelques années, venir troubler le ciel des constructeurs français...

Quand même n'est pas connecteur lorsqu'il est employé pour exprimer une opinion subjective de l'auteur sur le contenu informatif de la phrase, (52) :

(52) Une lettre postée dans le 4ème arrondissement de Paris arrive dès le surlendemain dans le 6ème. C'est quand même formidable, ça.

Surtout

Surtout est connecteur lorsqu'il introduit une relation de continuation entre deux segments de discours, (53) :

(53) Le député des Yvelines n'a pas apprécié de n'avoir pas été écouté sur la stratégie à suivre pour organiser le retour de l'ancien président. Surtout, il ne digère pas de ne plus avoir autant d'influence qu'avant sur Nicolas Sarkozy.

Surtout n'a pas été annoté comme connecteur lorsqu'il introduit certes un contraste mais entre un élément de sa phrase hôte et une alternative qu'il faut inférer, (54) :

(54) En Allemagne, l'immigration est surtout d'origine européenne.

TABLE 2 – Autres adverbiaux de LexConn qui présentent une ambiguïté fonctionnelle

CONN	**Fonction**	Exemple non connecteur
à ce propos	à CE(ana) propos	La question qu'il a posée **à ce propos** est d'actualité.
autrement	adv manière	Il en a décidé **autrement**.
	adv quantifieur	Ceci nous pose une question **autrement** délicate.
au moins	adv quantifieur	Luc a **au moins** trois amis.
au total	adv quantifieur	Luc a **au total** trois amis.
aussitôt	modif SP temporel	Luc est parti **aussitôt** après le café.
comme ça	comme ça(ana)	Le problème ne se pose pas **comme ça**.
	assez ... comme ça	Tu as assez bu **comme ça**.
dans ce cas	dans CE(ana) cas	Ils ne sont que deux ou trois **dans ce cas**.
de plus	adv quantifieur	Luc a gagné 20 euros, Marie 4 **de plus**.
en gros	SP manière	Luc a expliqué **en gros** ses motivations.
en même temps	SP temporel	Tout le monde est arrivé **en même temps**.

TABLE 3 – Adverbiaux de LexConn sans ambiguïté fonctionnelle

– à l'inverse	– du moins	– malgré tout
– à la réflexion	– d'un autre côté	– néanmoins
– à preuve	– du reste	– par comparaison
– à vrai dire	– en attendant	– par conséquent
– au bout du compte	– en conséquence	– par contre
– après tout	– en définitive	– par exemple
– bref	– en d'autres termes	– peu importe
– ceci/cela dit	– en effet	– pour autant
– cependant	– en outre	– pour preuve
– corrélativement	– en réalité	– pourtant
– d'ailleurs	– en revanche	– premièrement
– dans tous les cas	– en somme	– quoi qu'il en soit
– de ce fait	– en tous les cas	– résultat
– de toute façon	– en tout état de cause	– somme toute
– de toute manière	– en vérité	– tout au moins
– dès lors	– faute de quoi	– tout compte fait
– du coup	– la preuve	– tout d'abord

6 Les prépositions entre grammaire et discours

Le troisième aspect de la désambiguïsation du FDTB1 consiste à distinguer les prépositions qui ont une fonction purement discursive de celles qui sont sous-catégorisées par un élément de leur phrase hôte, cet élément pouvant être un verbe, nom, adjectif ou adverbe. Ces cas de désambiguation concernent cinq prépositions qui introduisent des infinitives (notées VInf) : *pour*, *afin de*, *plutôt que de*, *jusqu'à* et *avant de*. Les prépositions *pour* et *jusqu'à* peuvent également introduire un SN mais rappelons (Section 1) que les occurrences de ces prépositions ont été filtrées automatiquement pour ne garder que celles introduisant une infinitive.

Cette désambiguation concerne également les conjonctions de subordination

reliées morphologiquement à trois de ces prépositions, à savoir *pour que*, *afin que* et *plutôt que*.

Pour (+ VInf)

Le cas fort complexe de la préposition *pour* suivie d'une infinitive a fait l'objet d'un long travail débouchant sur une publication, (Colinet *et al.* (2014)), qui est résumée ici dans ses grandes lignes [9].

La préposition *pour* peut être connecteur, avec une valeur finale, causale ou temporelle, comme illustré dans les trois exemples ci-dessous.

(55) Côté alliances, DEC, qui s'est associé à Olivetti pour développer notamment des machines Risc - un microprocesseur à jeu d'instructions réduit...

(56) L' an dernier, le correspondant du quotidien britannique Financial Times s'est fait expulser pour avoir fait état de "l'évaporation" des énormes bénéfices tirés des exportations de pétrole pendant la guerre du Golfe.

(57) De son côté, la construction de logements reprend effectivement, après une forte baisse en 1991, pour remonter à un rythme annuel de 1,3 million de mises en chantier contre 1 million l'année précédente.

La préposition *pour* peut également introduire un complément sous-catégorisé par un verbe (58), un nom (59), un adjectif (60) ou encore un adverbe (61) (l'élément sous-catégorisant est souligné dans ces exemples) :

(58) Dans ces conditions, on peut se demander pourquoi le gouvernement n'a pas profité de l'occasion pour trancher en expliquant dès maintenant que le ralentissement de la croissance exige un accroissement des prélèvements obligatoires.

(59) "Si nous savons compléter en 1993 le travail entrepris depuis deux ans, Olivetti a toutes les qualités pour profiter de la nouvelle phase de croissance", déclare -t-il

(60) Certains voient dans le projet une aubaine inespérée : 280000 tonnes de céréales — dont la moitié de blé et de maïs — seront nécessaires, chaque année, pour nourrir les poules

(61) Mais la quasi-disparition du déficit commercial avec notre voisin d'outre-Rhin — . . .— est trop rapide pour être durable.

Enfin, *pour* peut introduire une « relative sans mot QU » (Huddleston & Pullum (2002)), (62), et des emplois méta-discursifs, illustrés en (63).

(62) Un pont pour franchir l'Amazone a été construit en 1745.

(63) Pour conclure. Pour ne citer que lui. Pour le dire autrement.

Si les « relatives sans mot QU » et les *pour* introducteurs d'expressions métadiscursives sont faciles à identifier, la distinction entre *pour* connecteurs de discours et *pour* introduisant un argument sous-catégorisé n'est pas aisée. Il

9. Ce résumé est inspiré de celui fait dans (Sagot *et al.* (2014)).

s'agit en effet d'une instance particulièrement délicate du problème général de la distinction entre arguments et modifieurs.

Colinet *et al.* (2014) rappellent deux critères classiques pour distinguer les arguments des modifieurs (critères 1 et 2 ci-dessous), et en proposent d'autres qui s'appliquent spécifiquement au cas de la préposition *pour*. Certains de ces critères s'appliquent correctement à toutes les catégories prédicatives, d'autres ne concernent que la sous-catégorisation verbale.

1. Un argument est *sémantiquement obligatoire*, pas un ajout (on notera que le caractère sémantiquement obligatoire n'implique pas que sa réalisation en syntaxe soit obligatoire). On peut tester ce critère en tentant de supprimer l'argument en *pour* et en vérifiant l'acceptabilité et le changement éventuel de sens. Ainsi, dans *Luc a fait une bouillabaisse pour faire plaisir à Marie*, l'infinitive est effaçable sans changement d'acceptabilité. En revanche, dans *Luc s'est dépêché pour finir*, l'effacement de l'infinitive conduit à un énoncé dont l'interprétation nécessite de récupérer l'argument effacé dans le contexte discursif gauche.
2. Un ajout est plus facilement antéposable qu'un argument. Il s'agit d'une tendance et non d'un critère strict.
3. *pour* commute dans certains cas avec les prépositions *à*, *de* et *contre* qui introduisent un complément considéré comme sous-catégorisé dans les principaux lexiques du français.
4. En général, lorsque le connecteur de discours *pour* introduit une infinitive, cette infinitive est contrôlée par le sujet (le sujet de l'infinitive est coréférent au sujet de la principale). Si donc on est face à un cas de contrôle d'une infinitive en *pour* par un argument autre que le sujet, il y a de fortes chances qu'il s'agisse d'un cas de sous-catégorisation.
5. De manière générale, on ne peut rajouter la séquence « *et ce,* » avant *pour* que dans les cas où *pour* est connecteur (cf. *Luc a fait une bouillabaisse et ce, pour faire plaisir à Marie* vs. **Luc s'est dépêché et ce, pour finir*). Certains cas sont toutefois moins clairs (*?Luc a pris un taxi et ce, pour aller à l'aéroport*).
6. Les possibilités de réfutation diffèrent selon que *pour* introduit un argument ou un modifieur. Ainsi, si l'on répond « *c'est faux* » à une assertion comme *Luc s'est dépêché pour finir*, on affirme que Luc ne s'est pas dépêché. A rebours, si l'on répond « *c'est faux* » à une assertion telle que *Luc a fait une bouillabaisse pour faire plaisir à Marie*, on affirme soit que Luc n'a pas fait de bouillabaisse, soit qu'il en a faite une mais dans un autre but que celui de faire plaisir à Marie. Nous ne rentrerons pas dans les détails, mais d'autres questions sémantiques liées aux interrogatives et aux propriétés de factualité vont dans le même sens.
7. Si l'on dispose d'un exemple où deux arguments en *pour* sont rattachés au même verbe sans être coordonnés, alors le premier est susceptible d'être sous-catégorisé, et le second est nécessairement modifieur (autrement dit, le second *pour* est un connecteur de discours).

Grâce à ces critères, il a été possible d'annoter manuellement toutes les occurrences de *pour* introduisant une infinitive dans le FTB [10]. La conjonction

10. Sur les 1161 occurrences ainsi annotées, 558 introduisent des compléments sous-

pour que peut aussi être connecteur ou introduire un complément sous-catégorisé par un verbe (64), un nom (65), un adjectif (66) et un adverbe (67) :

(64) Marie va s'arranger pour que la babysitter garde les enfants mercredi soir.

(65) Ce candidat a toutes les qualités pour que les militants lui fassent confiance et l'élisent au premier tour.

(66) Un complément d'informations est nécessaire pour que je puisse accomplir cette tache.

(67) Il pleut trop pour que nous puissions faire une promenade.

La conjonction *pour que* a été annotée en suivant les critères mis au point pour *pour* et il en est de même pour la préposition *afin de* et la conjonction *afin que*.

Avant de, jusqu'à, plutôt que de (+ VInf)

Les prépositions *avant de*, *jusqu'à* et *plutôt que de* peuvent être sous-catégorisées respectivement par *attendre*, *aller* et *préférer* :

(68) Elle a attendu que la première couche sèche avant de/pour passer la deuxième.

(69) La première épure du projet de loi de finances pour 1993 va même jusqu'à prévoir leur suppression pure et simple ! [11]

(70) Il préfère démissionner (plutôt) que d' accepter ces nouvelles conditions de travail. [12]

Sinon, ces prépositions sont connecteurs de discours :

(71) Elle a étudié à Paris 8 pendant 5 ans avant de se faire embaucher à Paris 3.

(72) Il a téléphoné au centre de service après vente jusqu'à s'endormir sur son bureau.

(73) Marie va à la fac en vélo, plutôt que d' utiliser les transports en commun.

7 Conclusion

Les données chiffrées concernant la taille du FTB (en articles, phrases et mots) et le nombre de connecteurs annotés dans le FDTB1 avec leurs catégories sont données dans la Table 4. Les mêmes données concernant le sous-corpus de Sequoia (voir note 2) à savoir *l'Est Républicain* et le FrWiki sont données dans la Table 5.

catégorisés, 518 sont des connecteurs de discours, 52 introduisent des « relatives sans mot QU », 33 introduisent des expressions métadiscursives.

11. Le verbe *aller* sous-catégorise aussi des compléments de forme *en V-ant* : *L'affaire va en s'empirant.*

12. Quand le complément indirect de *préférer* est une phrase, il est introduit par *plutôt que* ou *à ce que* : Il préfère que Pierre le fasse plutôt/à ce que Marie s'en charge.

TABLE 4 – Taille du FTB et nombre de connecteurs dans le FDTB1

FTB		FDTB1	
articles	1005	adv	3221
phrases	18535	cco	3653
mots	535 000	csu	1949
		prép V-inf	1070
		prép V-ant	536
		TOTAL	10441

TABLE 5 – Données chiffrées pour *l'Est Républicain* (ER) et FrWiki

ER			
articles	39	adv	26
phrases	524	cco	7
mots	10 000	csu	18
		prép V-inf	7
		prép V-ant	1
		TOTAL	77

FrWiki			
articles	19	adv	32
phrases	996	cco	5
mots	22 000	csu	17
		prép V-inf	7
		prép V-ant	2
		TOTAL	85

Le seul autre corpus français concernant l'écrit [13] qui a été annoté pour le discours est le corpus Annodis (Péry-Woodley *et al.* (2011)). Ce corpus a reçu deux annotations : annotation en relations rhétoriques et annotation en structures multi-échelles. La première correspond à l'étude de l'organisation discursive qui est étudiée dans le FDTB, même si les approches sont différentes : l'annotation en relation rhétoriques d'Annodis s'inspire de la SDRT, Segmented Discourse Representation Theory, (Asher & Lascarides (2003)), tandis que, rappelons-le, l'annotation du FDTB1 et dans le futur du FDTB s'inspire du PDTB avec un focus sur les marques lexicales (connecteurs et AltLex) des relations rhétoriques.

Les données chiffrées concernant l'annotation en relations rhétoriques d'Annodis sont données dans la Table 6 [14]. Si l'on compare la taille des corpus en termes de mots, on s'aperçoit que le corpus Annodis est nettement plus petit que le corpus FDTB (environ 5%) ; il est en fait comparable en taille à la sous-partie du corpus Sequoia (ER et FrWiki) que nous avons annoté selon les principes du FDTB1. Les connecteurs de discours n'ayant pas joué un rôle central dans Annodis, on ne sait pas quel est leur nombre dans le corpus : on sait qu'il y a 3355 relations de discours (RD), et si on pose approximativement que 50%

13. Il existe un corpus oral français annoté discursivement (Crible & Zufferey (2015)).
14. Ces données sont extraites du site http://redac.univ-tlse2.fr/corpus/annodis/.

des relations de discours sont marquées par un connecteur explicite, cela donne 1778 connecteurs. Par contre, on connaît le nombre de segments de discours élémentaires (EDU, Elementary Discours Unit) ou complexes (CDU, Complex Discours Unit), voir note 3.

TABLE 6 – Données chiffrées pour Annodis

Annodis		Relations discursives	
articles	100	EDU	3188
mots	28 000	CDU	1395
		RD	3355

L'accord inter-annotateurs pour le FDTB1 a été évalué de deux façons :
- l'accord antre deux annotateurs experts (deux auteurs de cet article) sur un échantillon de 13 articles du FTB donne un kappa de 0,70 ;
- l'accord entre un annotateur expert (auteur de cet article) et un annotateur naïf [15] sur le corpus de L'Est Républicain donne un kappa de 0,63 ;
- l'accord entre un annotateur expert (auteur de cet article) et un annotateur naïf sur le corpus de FrWiki donne un kappa de 0,46.

Les corpus annotés sont librement disponibles sur le site https://gforge.inria.fr/frs/?group_id=6145. Sur les 353 éléments enregistrés dans LexConn (version 2), 241 ont au moins une occurrence dans le FDTB1, soit près de 70%. La table 7 donne la liste des 100 connecteurs les plus fréquents du FDTB1.

TABLE 7: Liste des 100 connecteurs les plus fréquents du FDTB1

Forme	Catégorie	Nombre	Forme	Catégorie	Nombre
et	cco	2194	maintenant	adv	29
mais	cco	948	de même	adv	29
npour	prep	519	déjà	adv	24
en	prep-V-ant	490	globalement	adv	23
si	csu	483	de plus	adv	23
aussi	adv	316	parallèlement	adv	22
ainsi	adv	296	sans que	csu	22
donc	adv	269	bref	adv	21
alors que	csu	235	à l'inverse	adv	21
car	cco	175	comme si	csu	20
comme	csu	168	au lieu de	prep	20
enfin	adv	159	à cet égard	adv	20
en effet	adv	152	du coup	adv	20
également	adv	152	en réalité	adv	19
sans	prep	148	encore	adv	19
pourtant	adv	131	plutôt	adv	18
cependant	adv	128	ou bien	cco	18
en revanche	adv	124	c'est pourquoi	csu	16
lorsque	csu	122	en particulier	adv	15
puisque	csu	121	résultat	adv	15

15. L'annotateur naïf est en fait un binôme d'étudiants du Master de Linguistique Informatique de L'Université Paris-Diderot après qu'ils se soient mis d'accord sur leurs annotations respectives.

quand	csu	116	dès lors	adv	14
ou	cco	112	en conséquence	adv	14
afin de	prep	110	dès que	csu	14
après	prep	108	avant que	csu	14
alors	adv	106	de fait	adv	14
tandis que	csu	93	depuis	adv	14
parce que	csu	93	sinon	adv	13
même si	csu	91	effectivement	adv	13
puis	cco	85	d'une part	adv	13
avant de	prep	79	depuis que	csu	12
d'ailleurs	adv	79	tant que	csu	12
or	cco	77	tout de même	adv	12
d'autre part	adv	70	après que	csu	11
surtout	adv	66	en attendant	adv	11
d'abord	adv	62	auparavant	adv	10
par exemple	adv	61	de toute façon	adv	10
en fait	adv	60	à condition que	csu	10
en outre	adv	58	alors même que	csu	10
ensuite	adv	53	tout d'abord	adv	10
tout en	prep-V-ant	46	malheureusement	adv	9
par ailleurs	adv	45	en même temps que	csu	9
pour autant	adv	44	en même temps	adv	9
finalement	adv	44	au total	adv	9
bien que	csu	43	quitte à	prep	9
néanmoins	adv	40	dès lors que	csu	9
au moment où	csu	39	plutôt que de	prep	8
au contraire	adv	37	dans le but de	prep	8
en tout cas	adv	32	faute de	prep	8
soit	cco	31	afin que	csu	8
d'autant que	csu	29	en vue de	prep	8

Remerciements

Ce travail a bénéficié de crédits de l'IUF et de l'axe 5 du Labex EFL.

Bibliographie

Abeillé, A., Clément, Lionel, & Toussenel, F. 2003. Building a treebank for French. *In :* Abeillé, A. (ed), *Treebanks*. Dordretch : Kluwer Academic Publishers.

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

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

Candito, Marie, & Seddah, Djamé. 2012 (June). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *In : TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*.

Colinet, Margot, Danlos, Laurence, Dargnat, Mathilde, & Winterstein, Grégoire. 2014. Emplois de la préposition pour suivie d'une infinitive : description, critères formels et annotation en corpus. *In : Actes du Congrès Mondial de Linguistique Franccaise (CMLF, 2014)*.

Crible, Ludivine, & Zufferey, Sandrine. 2015. Assessing the validity of annotations guidelines : DRDs across languages and modalities. *In : Proceedings of the First Action Conference of TextLink*.

Danlos, Laurence. 2006. Discourse verbs and discourse periphrastic links. *In : Proceedings of the second workshop on Constraints in Discourse (CID 2006)*.

Danlos, Laurence. 2012. Méthodologie pour le FDTB (French Discourse Tree Bank). *Page 2 p. of : La linguistique de corpus à l'heure de la confrontation entre concepts, techniques et applications*.

Danlos, Laurence, Antolinos-Bassos, Diégo, Braud, Chloé, & Roze, Charlotte. 2012. Vers le FDTB : French Discourse Tree Bank. *In : Actes de TALN 2012*.

Degand, Liesbeth, & Fagard, Benjamin. 2011. Alors between Discourse and Grammar. The role of syntactic position. *Functions of Language*, **18(1)**, 29–56.

Huddleston, R., & Pullum, G. 2002. *The Cambridge Grammar of the English Language*. Cambridge : Cambridge University Press.

Karssenberg, Lena, & Lahousse, Karen. 2014. Ainsi en tête de phrase + inversion : une analyse de corpus. *SHS Web of Conferences*, **8**, 2413–2427.

Kehler, Andrew, Cohen, Jonathan, & Rohde, Hannah. 2015. Conversational Eliciture. *In : Proceedings of the First Action Conference of TextLink*.

Levy, Roger, & Andrew, Galen. 2006. Tregex and Tsurgeon : tools for querying and manipulating tree data structures. *In : Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Molinier, Christian. 2013. Ainsi : Deux emplois complémentaires d'un adverbe type. *Linguisticae Investigationes*, **36-2**, 311–327.

Molinier, Christian, & Lévrier, Francis. 2000. *Grammaire des adverbes*. Genève : Droz.

Prasad, Rashmi, Dinesh, Nikhil, Lee, Alan, Miltsakaki, Eleni, Robaldo, Livio, Joshi, Aravind, & Webber, Bonnie. 2008. The Penn Discourse Treebank 2.0. *In : Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M., & Asher, N. 2011. La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Revue TAL*, **52(3)**, 71–101.

Roze, Charlotte, Danlos, Laurence, & Muller, Philippe. 2012. LEXCONN : a French Lexicon of Discourse Connectives. *Revue Discours*.

Rysová, M., & Rysová, K. 2014. The Centre and Periphery of Discourse Connectives. *Pages 452–459 of : Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation*.

Sagot, Benoît, Danlos, Laurence, & Colinet, Margot. 2014. Sous-catégorisation en pour et syntaxe lexicale. *In : Traitement Automatique du Langage Naturel 2014*.