FDTB1: Repérage des connecteurs de discours dans un corpus français

Laurence Danlos, Margot Colinet, Jacques Steinlin

FDTB1: Repérage des connecteurs de discours dans un corpus français

Laurence Danlos

Université Paris-Diderot, Alpage et IUF

Laurence.danlos@inria.fr

Margot Colinet

Alpage

margotcolinet@gmail.com

Jacques Steinlin

Alpage

Jacques.steinlin@gmail.com

Abstract

This paper presents the identification 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, but for example are used as an adverb of manner or a preposition introducing a subcategorized complement. More than 10K connectives have been identified.

Keywords:discourse connectives, discourse annotation, grammar and discourse.

Résumé

Cet article présente le repérage 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 identifiés.

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 manuellement 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 manuellement 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 *grosso modo* 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, entre autres du fait que de nombreux items lexicaux (e.g. *et*, *engros*, *ainsi*, *alors*) sont ambigus entre un emploi comme connecteur de discours et un emploi non discursif. A titre d'illustration, la conjonction de coordination *et* est connecteur en [1a] et non-connecteur en [1b]. De même, l'adverbial *en gros* est connecteur en [2a] et non-connecteur en [2b].

[1a] Fred a fini d'écrire son article et il est parti en vacances.

[1b] Fred et Marie sont de très bons amis.

[2a] Fred s'est cassé le bras et a attrapé la grippe. En gros, il ne va pas bien du tout.

[2b] Ce film traite en gros du réchauffement climatique.

La détermination du statut discursif de *et* dans les exemples en [1] est triviale, mais ceci est loin d'être toujours le cas, comme le montre la littérature sur *ainsi* (Molinier, 2013; Karssenberg, Lahousse, 2014) ou *alors* (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. Il n'est pas clair de savoir comment la fréquence des connecteurs anglais a été déterminée vu l'ambiguïté dont nous venons de parler. Seule une étude telle que celle menée dans le FDTB1 permet de déterminer la fréquence des connecteurs et d'identifier les 100 connecteurs français les plus fréquents (au moins dans un corpus journalistique) : ceux-ci sont présentés dans l'Annexe C.

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

Ce travail repose donc crucialement sur la notion de connecteur de discours qui est définie de manière fonctionnelle : les connecteurs de discours sont des items lexicaux qui permettent d'exprimer explicitement les relations discursives (sémantiques ou rhétoriques) entre deux segments de discours, « élémentaires » ou « complexes »[2]. Les connecteurs de discours du français ont été répertoriés dans LexConn (Roze et al., 2012), un lexique qui recense de la manière la plus exhaustive possible les connecteurs avec leur catégorie syntaxique et la ou les relation(s) de discours qu'ils lexicalisent. Les catégories syntaxiques sont : conjonction de coordination, conjonction de subordination, préposition (introduisant un VP à l'infinitif ou au participe présent) et adverbial (catégorie qui regroupe principalement des adverbes simples et des syntagmes prépositionnels).

Le travail effectué dans le FDTB1 s'appuie sur LexConn tant sur le plan théorique que méthodologique. Sur le plan théorique, les principes qui ont guidé l'élaboration de LexConn ont tous été retenus dans le FDTB1. 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 occurrences d'éléments de LexConn qui n'ont pas portée sur un VP dans le corpus ont été éliminées automatiquement. A titre d'illustration, seules les occurrences de la préposition *pour* introduisant un VP à l'infinitif ont été projetées sur le FTB, en excluant celles introduisant un SN[3].

Sur le plan méthodologique, nous avons projeté automatiquement sur le FTB les éléments de LexConn respectant le principe ci-dessus, puis effectué des tâches de désambiguïsation pour savoir si ces occurrences étaient effectivement employées comme connecteurs. Les tâches de désambiguïsation sont les suivantes :

- désambiguïsation morpho-syntaxique (Section 3), par exemple pour les homonymes comme *bref* qui peut êtreun adverbe connecteur ou un adjectif,
- désambiguïsation entre grammaire et discours (Section 4) pour les adverbiaux comme *ainsi* et *alors* quipeuvent avoir un emploi comme connecteur et un emploi d'ajout à l'intérieur de leur phrase hôte,
- désambiguïsation entre grammaire et discours (Section 5) pour les prépositions et conjonctions de subordination comme *pour* et *pour que* qui peuvent avoir un emploi comme connecteur et un emploi d'introducteur de complément sous-catégorisé par un élément (verbal, nominal, adjectival ou adverbial) de la phrase où ils apparaissent.

Le corpus FDTB1 est librement disponible à l'adresse https://gforge.inria.fr/frs/?group_id=6145. Les résultats quantitatifs de l'annotation

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

[3] Un tel filtrage bénéficie de l'annotation morpho-syntaxique du FTB et s'effectue automatiquement avec l'outil Tregex (Levy, Andrew, 2006).

sont donnés à la Section 6. Avant d'expliquer les tâches de désambiguïsation, nous allons présenter LexConn et préciser la notion de connecteur de discours qui est au cœur du FDTB1.

2 LexConn et la notion de connecteur de discours

LexConn (dans sa première version de 2012) compte 325 connecteurs de discours qui sont listés avec leur catégorie syntaxique et la ou les relation(s) de discours qu'ils expriment et qui sont illustrées par des exemples (principalement issus de Frantext, http://www.frantext.fr). Rappelons les catégories grammaticales des connecteurs : 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 (prép) 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.

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 [3] où le gérondif indique comment la porte d'entrée a été cassée[4].

[3] Fred a cassé la porte d'entrée, en l'ouvrant violemment.

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, [4]. 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.

[4a] Fred s'est excusé probablement pour que Marie accepte de lui prêter de l'argent.

[4b] Fred s'est excusé pour que probablement Marie accepte de lui prêter de l'argent.

[4c] Fred s'est excusé pour probablement que Marie accepte de lui prêter de l'argent.

La tâche la plus délicate de LexConn a été de déterminer parmi l'ensemble des 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

[4] Pour un gérondif, nous considérons que le connecteur est *en ...-ant*. Les équivalents anglais des gérondifs (les formes *by V-ing*) n'ont pas été annotés comme connecteurs dans le PDTB.

connecteur : par exemple, la phrase *Du coup, Fred 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 lendemain matin, Fred est arrivé* est aussi interdite à l'initiale d'un discours bien que *le lendemain matin* ne soit pas connecteur, d'autres critères ont été mis en avant lors du développement de LexConn. Nous allons en présenter quatre : clivage, substituabilité, compositionalité qui sont des critères sémantiques et le critère de cohérence qui relève plutôt du discours.

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 *le lendemain matin* peut être clivé en [5] ce qui indique que ce n'est pas un connecteur, tandis que *par exemple* ne peut pas être clivé en [6] ce qui indique que c'est un connecteur.

[5a] Marie est tombée malade. Le lendemain matin, Fred est arrivé.

[5b] Marie est tombée malade. C'est le lendemain matin que Fred est arrivé.

[6a] Manet admirait chez son illustre prédécesseur [Velasquez] le recours à des coloris très vifs, ... L'influence de Vélasquez se retrouve par exemple dans Le Joueur de fifre, . . . [Wikipedia]

[6b] #Manet admirait chez son illustre prédécesseur le recours à des coloris très vifs, ... C'est par exemple que l'influence de Vélasquez se retrouve dans Le Joueur de fifre, ...

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 [7]. Sinon, il est considéré comme connecteur, voir *à part ça* en[8].

[7a] Fred est allé en Argentine. Après ça, il est allé au Pérou.

[7b] Fred est allé en Argentine. Après qu'il est allé en Argentine, il est allé au Pérou.

[8a] Hier soir j'ai croisé Fred dans un bar. À part ça , il nous dit tout le temps qu'il est fatigué.

[8b] #Hier soir j'ai croisé Fred 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 figement. Ce critère concerne principalement les adverbiaux de type SP 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, ce qui se manifeste d'une part par le fait que le nom n'est plus variable en nombre, [9a] d'autre part qu'il n'est pas modifiable, [9b].

[9a] Fred n'est pas allé au cinéma. A la place (*aux places), il a fait du jogging.

[9b] Fred n'est pas allé au cinéma. A la place (*précise), il a fait du jogging.

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, [10].

[10] Ce serait vraiment utile pour nous d'aller à cette réunion.

[10a] Ceci dit, on peut s'en passer.

[10b] #On peut s'en passer. (Roze et al., 2012)

La notion de connecteur de discours explicite s'oppose à celles de connecteur implicite et de « AltLex » : lorsqu'une une phrase (typographique) ne contient aucun connecteur explicite, il est souvent considéré qu'elle est reliée à son contexte gauche par un connecteur implicite (voir note 1); toutefois, il a été souligné dans divers travaux qu'une phrase sans connecteur explicite peut se voir relier à son contexte gauche par une relation discursive lexicalisée par des items lexicaux n'appartenant pas à la catégorie des connecteurs de discours, qui ont été baptisés AltLex (Alternative Lexicalization) dans le PDTB. Illustrons sur des exemples : en [11a] le connecteur explicite *parce que* lie les deux propositions avec un sens causal. En [11c], le lecteur doit inférer que les deux phrases sont reliées par une relation causale : on doit positionner un connecteur implicite, noté ø. A l'intermédiaire, [11b] 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 d'AltLex .

[11a] Fred est de mauvaise humeur parce qu'il a mal dormi.

[11b] Fred est de mauvaise humeur. Ceci est dû au fait qu'il a mal dormi.

[11c] Fred est de mauvaise humeur. ø Il a mal dormi.

Le PDTB décrit quelques cas d'AltLex pour l'anglais et les définit par le fait qu'on ne peut pas leur ajouter de connecteur sans produire un effet de redondance. Pour le français, c'est un vaste champ d'étude non exploré (à l'exception des « 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 de la langue et éventuellement non perçue[5].

[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

La séquence *Ceci est dû au fait que* en [11b] est compositionnelle et à ce titre ne saurait en aucun cas être considérée comme un connecteur de discours : c'est clairement un AltLex. Toutefois, la distinction entre connecteur de discours et AltLex peut être plus subtile. A titre d'illustration, considérons l'adverbial *à ce moment-là*. En [12a] cet adverbial est compositionnel (avec un sens de concomitance temporelle) car le déterminant *ce ...là* est anaphorique comme le montre la paraphrase en *au moment où il a commencé à pleuvoir*. A l'inverse, cet adverbial est non compositionnel en [13a] où *ce ...là* est non anaphorique. Ceci indique que seul à ce moment- là en [13a] peut prétendre au statut de connecteur (avec un sens de conséquence). Ce statut est confirmé par deux autres critères : d'abord *à ce moment-là* en [13a] ne peut pas être clivé — [13b] est inacceptable contrairement à [12b] —, ensuite, *moment* en [13a] ne peut pas être modifié — [13c] est inacceptable contrairement à [12c].

> [12a] Il a commencé à pleuvoir. A ce moment-là, Pierre est arrivé.
>
> [12b] Il a commencé à pleuvoir. C'est à ce moment-là que Pierre est arrivé.
>
> [12c] Il a commencé à pleuvoir. A ce moment-là précis, Pierre est arrivé.
>
> [13a] Tu as l'air de penser qu'elle n'est pas honnête. A ce moment-là, tu devrais ne rien lui raconter.
>
> [13b] #Tu as l'air de penser qu'elle n'est pas honnête. C'est à ce moment-là que tu devrais ne rien lui raconter.
>
> [13c] #Tu as l'air de penser qu'elle n'est pas honnête. A ce moment-là précis, tu devrais ne rien lui raconter. (Roze et al., 2012)

Tous les critères convergent donc pour indiquer que *à ce moment-là* en [12] est un AltLex tandis que c'est un connecteur en [13a] Toutefois, la situation n'est pas toujours aussi tranchée : 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 menée par (Rysová, Rysová, 2014) sur le tchèque). Ainsi, nous avons parfois hésité entre le statut de connecteur ou 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 [14] *alors* est une anaphore de l'expression temporelle *de 1870 à 1914* située dans la phrase

---

texte suivant qui n'est pas 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 ...*

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.

[14] De 1870 à 1914, plus de cinq cents millions d'images en cinquante langues sortent des bâtiments duquai de Dogneville, à Epinal. L'imagerie compte alors jusqu'à cent cinquante salariés.

En ce qui concerne les arguments des connecteurs, rappelons que la contrainte qu'un segment de discours élémentaire comporte un VP est appliquée tant pour LexConn que pour FDTB1. Cette contrainte vient de raisons théoriques — éviter de considérer la préposition *à cause de* qui n'introduit que des SN comme un connecteur de discours — et pratiques — limitation du travail d'annotation dans le FDTB. Cette contrainte fait que la préposition *jusqu'à* en [15a] n'est pas considérée comme connecteur car elle introduit un SN, alors qu'elle est considérée comme connecteur en [15b] ; de même, *cependant* n'est pas considéré comme connecteur en [15c] car sa portée ne contient pas de VP. Cette contrainte débouche donc sur des décisions peu heureuses mais il sera toujours possible d'étendre le FDTB en s'en affranchissant. Une telle extension sera plus facilement réalisable le jour où on disposera d'une annotation des noms du corpus comme événementiels ou pas : ainsi le nom *arrivée* a une lecture événementielle dans un exemple comme *Fred est resté jusqu'à l'arrivée des secours*, et une lecture non événementielle dans des exemples comme *Le cheval a bien franchi la ligne d'arrivée* et *On se retrouve devant l'arrivée de la course*.

[15a] Fred a couru jusqu'à épuisement

[15b] Fred a couru jusqu'à en être épuisé.

[15c] 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.

3 Ambiguïtés morpho-syntaxiques

Le premier aspect de la désambiguïsation dans le FDTB1 consiste, pour chaque occurrence d'item qui peut être connecteur, à décider si elle correspond morpho-syntaxiquement à la catégorie du connecteur recherché. Le premier cas d'ambiguïté est celui des homonymes, par exemple le mot *car* qui peut-être une conjonction de coordination (répertoriée dans LexConn) ou un nom commun (*Le car est stationné sur la place de l'Etoile*). Le second cas correspond à 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 dans LexConn comme adverbial (de type syntagme prépositionnel), [16a] mais elle peut correspondre à un pronom suivi d'un verbe comme en [16b].

[16a] Fred avait l'air sûr de lui. En fait, il était mort de trouille.

[16b] La Grand-Place était piétonne. Le maire en fait un parking.

Ces deux cas d'ambiguïté peuvent être levés automatiquement grâce à l'annotation morpho-syntaxique du corpus initial. L'Annexe A donne la liste des éléments de LexConn qui présentent une ambiguïté morpho-syntaxique.

4 Les adverbiaux entre grammaire et discours

Le deuxième aspect de la désambiguïsation consiste à distinguer les occurrences des adverbiaux de LexConn qui ont une fonction discursive de ceux qui ont un rôle sémantique à l'intérieur de leur phrase hôte (avec la fonction syntaxique d'ajout et plus rarement de complément). Dans les termes de (Molinier, Lévrier, 2000), ceci s'approche de la distinction entre « adverbe de phrase » et autre adverbe. Cette désambiguïsation s'appuie sur les critères utilisés dans LexConn (rappelés à la Section 2). Au cas par cas, pour aider à déterminer si un adverbial potentiellement connecteur est effectivement employé comme connecteur en contexte, il a paru nécessaire de décrire l'emploi comme connecteur en donnant un aperçu de la ou les relation(s) de discours lexicalisée(s), et l'emploi comme non connecteur en précisant le rôle sémantique à l'intérieur de la phrase hôte. Ce travail prolonge à large échelle celui qui a été effectué par des linguistes sur quelques connecteurs adverbiaux; il est illustré ci-dessous pour *au contraire* (pas étudié dans la littérature à notre connaissance) et *ainsi* (largement étudié dans la littérature).

Au contraire a été annoté comme connecteur lorsqu'il lexicalise un contraste, [17a] ou une sorte de reformulation du contexte gauche perçu comme une litote, [17b]; 38 occurrences. *Au contraire* n'est pas retenu comme connecteur lorsqu'il renforce une assertion négative, [18]; 8 occurrences.

[17a]. 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.

[17b] Qu'il y ait aujourd'hui, ou qu'il y ait encore après le prochain comité directeur, plusieurs textes d'orientation en présence n'est pas en soi nuisible. Cela peut être au contraire une preuve de la vitalité du seul parti véritablement démocratique en France [...]

[18] La nouvelle diminution du taux d'escompte de la Banque du Japon n'a nullement déprimé la monnaie japonaise, au contraire. [6]

Ainsi a été annoté comme connecteur lorsqu'il lexicalise une relation de résultat ou d'exemplification, comme en [19a] sans inversion de l'ordre canonique sujet-verbe ou en [19b] avec inversion (Molinier, 2013; Karssenberg, Lahousse, 2014) ;

[6] On pourrait considérer que dans cet exemple *au contraire* est un connecteur avec un second argument vide. Mais il ne saurait être question d'autoriser un connecteur à avoir un argument vide.

291 occurrences. *Ainsi* n'est pas connecteur lorsqu'il est utilisé comme anaphore de manière [20a] ou comme anaphore ou cataphore d'un discours rapporté, [20b] ; 32 occurrences.

> [19a] La Commission nationale [...] se limite à vérifier si les obligations comptables et financières sontremplies. Ainsi , il n'existe à ce jour aucun contrôle des dépenses des partis .
>
> [19b] 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 ».

> [20a] Fred a été grossier avec Marie. Il s'est comporté ainsi parce qu'il était fatigué.
>
> [20b] 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 ».

L'Annexe B donne les différents emplois d'une centaine d'adverbiaux qui sont ambigus entre grammaire et discours. Cette annexe dresse aussi la liste des adverbiaux de LexConn qui sont toujours employés comme connecteurs (au moins dans notre corpus) : ils sont une cinquantaine.

5 Les prépositions et conjonctions de subordination entre grammaire et discours

Le troisième aspect de la désambiguïsation du FDTB1 consiste à distinguer les occurrences des prépositions et conjonctions de subordination qui ont une fonction discursive de celles qui introduisent un complément sous-catégorisé par un élément verbal, nominal, adjectival ou adverbial. Cette désambiguïsation concerne cinq prépositions qui introduisent des infinitives — *pour, afin de, plutôt que de, jusqu'à* et *avant de* — et trois conjonctions de subordination reliées morphologiquement à trois de ces prépositions, à savoir *pour que, afin que* et *plutôt que*.

Le cas le plus complexe et le plus fréquent est celui de la préposition *pour* suivie d'une infinitive. Cette préposition peut être connecteur, avec une valeur finale, causale ou temporelle, [21].

> [21a] Côté alliances, DEC, qui s'est associé à Olivetti pour développer notamment des machines Risc...
>
> [21b] L'an dernier, le correspondant du quotidien britannique Financial Times s'est fait expulser pouravoir fait état de «l'évaporation» des énormes bénéfices tirés des exportations de pétrole pendant la guerre du Golfe.
>
> [21c] 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.

Cette préposition peut également introduire un complément sous-catégorisé par un verbe [22a], un nom [22b], un adjectif [22c], ou encore un adverbe [22d] (l'élément sous-catégorisant est souligné dans ces exemples).

[22a] Le gouvernement n'a pas profité de l'occasion pour trancher.

[22b] Olivetti a toutes les qualités pour profiter de la nouvelle phase de croissance.

[22c] 280000 tonnes de céréales seront nécessaires, chaque année, pour nourrir les poules.

[22d]. Ceci est trop rapide pour être durable.

Enfin, pour peut introduire une « relative sans mot qu » (Huddleston, Pullum, 2002), [23a], et introduire des expressions méta-discursives, [23b].

[23a] Un pont pour franchir l'Amazone a été construit en 1745.

[23b] 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 connecteur de discours et pour introduisant un argument sous-catégorisé n'est pas aisée. Il s'agit en effet d'une instance particulièrement délicate du problème général de la distinction entre arguments et modifieurs, pour laquelle une batterie de critères a été mise au point (Colinet et al., 2014). Ces critères ont permis d'annoter manuellement les 1161 occurrences de *pour* introduisant une infinitive dans le FTB : 518 sont des connecteurs de discours (44%), 558 introduisent des compléments sous-catégorisés, 52 introduisent des relatives sans mot qu, et 33 introduisent des expressions métadiscursives. Ce travail a aussi permis de compléter les lexiques syntaxiques dans lesquels la préposition *pour* est largement ignorée comme introducteur de complément sous-catégorisé (Sagot et al., 2014).

La conjonction *pour que* peut aussi être connecteur ou introduire un complément sous-catégorisé par un verbe [24a] un nom [24b] un adjectif [24c] et un adverbe [24d]. Elle 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*.

[24a] Marie va s'arranger pour que la babysitter garde les enfants mercredi soir.

[24b] Ce candidat a toutes les qualités pour que les militants l'élisent au premier tour.

[24c] Un complément d'informations est nécessaire pour que je puisse accomplir cette tâche.

[24d] Il pleut trop pour que nous puissions faire une promenade.

Les prépositions *avant de, jusqu'à* et *plutôt que de* peuvent introduire des compléments sous-catégorisés respectivement par *attendre*, *aller* et *préférer*, [25]. Sinon, ces prépositions sont connecteurs de discours, [26].

[25a] Elle a attendu que la première couche sèche avant de/pour passer la deuxième.

[25b] La première épure du projet de loi de finances pour 1993 va même jusqu'à prévoir leur suppression pure et simple ![7]

[25c] Il préfère démissionner plutôt que d'accepter ces nouvelles conditions de travail.[8]

[26a] Elle a étudié à Paris 8 pendant 5 ans avant de se faire embaucher à Paris 3.

[26b] Il a téléphoné au centre de service après vente jusqu'à s'endormir sur son bureau.

[26c] Marie va à la fac en vélo, plutôt que d'utiliser les transports en commun.

6 Conclusion et perspectives futures

Les données chiffrées concernant la taille du FTB (en articles, phrases et mots) et le nombre de connecteurs repérés dans le FDTB1 avec leurs catégories sont données dans la Table 1. 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 (ER)[9]. Les données chiffrées concernant le sous-corpus de Sequoia sont données dans la Table 2.

Insert Figure-Danlos-1 here

Insert Figure-Danlos-2 here

Cette annotation a mis en évidence une trentaine de connecteurs non répertoriés dans LexConn dont une nouvelle version (comptant 353 éléments) est disponible sur le site du FDTB1 (https://gforge.inria.fr/ frs/?group_id=6145). Près de 70% des éléments de LexConn ont au moins une occurrence dans le FDTB1. L'Annexe C donne la liste des 100 connecteurs les plus fréquents repérés dans le FDTB1.

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

[8] 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 que/à ce que Marie s'en charge.*

[9] 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.

L'accord inter-annotateurs pour le FDTB1 a été évalué de la façon suivante :

- l'accord entre 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 [10] 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.

Le seul autre corpus français concernant l'écrit[11] 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 FDTB s'inspire du PDTB avec un focus sur les marques lexicales (connecteurs et AltLex) des relations discursives.

Les données chiffrées concernant l'annotation en relations rhétoriques d'Annodis sont données dans la Table 3[12]. 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 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 connaît juste le nombre de segments de discours élémentaires (EDU, Elementary Discours Unit), de segments complexes (CDU, Complex Discours Unit, voir note 3) et de relations de discours (RD)[13].

Insert Figure-Danlos-3 here

Le FDTB1 est donc le premier corpus écrit où les connecteurs du discours du français sont repérés systématiquement. Ce corpus peut être utilisé par les linguistes intéressés par les connecteurs. Il peut aussi être utilisé pour développer des méthodes d'apprentissage afin de repérer automatiquement les connecteurs dans un

[10] 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.

[11] Il existe un corpus oral français annoté discursivement (Crible & Zufferey, 2015).

[12] Ces données sont extraites du site http://redac.univ-tlse2.fr/corpus/annodis/.

[13] Si l'on pose approximativement — mais en se basant sur des travaux dans d'autres langues dont l'anglais — que 50% des relations de discours sont marquées par un connecteur explicite ou un AltLex, on peut spéculer sur le fait que le corpus Annodis comporte aux alentours de 1600 connecteurs.

autre corpus, et ce d'autant plus aisément qu'il repose sur une annotation morpho-syntaxique.

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

1. annotation du sens et des arguments des connecteurs explicites repérés dans le FDTB1,

2. identification des AltLex et des connecteurs implicites (éléments définis à la Section 2),

3. annotation du sens et des arguments des éléments identifiés à l'étape 2. La première et la troisième tâche seront effectuées dans l'esprit du PDTB, avec quelques modifications mineures concernant la hiérarchie des sens de connecteurs et l'annotation de leurs arguments (Danlos et al., 2012).

La première tâche, qui débouchera sur le FDTB2, est en cours avec une étude approfondie des résultats quantitatifs et qualitatifs de l'accord inter-annotateurs.

Remerciements

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

Références

ABEILLE A., CLEMENT L. , TOUSSENEL F. (2003). *Building a treebank for French*. In A. Abeillé, Ed., Treebanks. Dordretch : Kluwer Academic Publishers.

ASHER N., LASCARIDES A. (2003). *Logics of Conversation*. Cambridge : Cambridge University Press.

BRAS M. (2008). *Entre relations temporelles et relations de discours*. Université de Toulouse le Mirail : HDR.

CANDITO M., SEDDAH D. (2012). *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, Grenoble, France.

COLINET M., DANLOS L., DARGNAT M., WINTERSTEIN G. (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), Berlin, Allemagne.

CRIBLE L., ZUFFEREY S. (2015). *Assessing the validity of annotations guidelines: Drds across languages and modalities*. In Proceedings of the First Action Conference of TextLink, Louvain-La Neuve, Belgique.

DANLOS L. (2006). *Discourse verbs and discourse periphrastic links*. In Proceedings of the second workshop on Constraints in Discourse (CID 2006), Maynooth, Ireland.

DANLOS L., ANTOLINOS-BASSOS D., BRAUD C., ROZE C. (2012). *Vers le FDTB : French Discourse Tree Bank.* In Actes de TALN 2012, Grenoble, France.

DEGAND L., FAGARD B. (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 L., LAHOUSSE K. (2014). *Ainsi en tête de phrase + inversion : une analyse de corpus.* SHS Web of Conferences, 8, 2413–2427.

LEVY R., ANDREW G. (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), Gènes, Italie.

MOLINIER C. (2013). *Ainsi : Deux emplois complémentaires d'un adverbe type*. Linguisticae Investigationes, 36-2, 311–327.

MOLINIER C.,LEVRIER F. (2000). Grammaire des adverbes. Genève : Droz.

PRASAD R., DINESH N., LEE A., MILTSAKAKI E., ROBALDO L., JOSHI A. , WEBBER B. (2008). *The Penn Discourse Treebank 2.0.* In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrackech, Maroc.

PERY-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 C., DANLOS L., MULLER P. (2012). *LexConn : a French Lexicon of Discourse connectives*. Revue Discours.

RYSOVA M., RYSOVA K. (2014). *The centre and periphery of discourse connectives*. In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, p. 452–459, Phuket, Thailand.

SAGOT B., DANLOS L., COLINET M. (2014). *Sous-catégorisation en pour et syntaxe lexicale*. In Traitement Automatique du Langage Naturel 2014, Marseille, France.

A Eléménts de LexConn présentant une ambiguïté morpho-syntaxique

La Table 4 dresse la liste des éléments de LexConn qui présentent une ambiguïté morpho-syntaxique. Cette liste se veut 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 corpus.

Rappelons que la catégorie adverbiale (adv) est utilisée dans LexConn pour les adverbes et les syntagmes prépositionnels. Elle est aussi utilisée pour des noms nus comme *résultat* ou *remarque*. Le symbole C représente un complémenteur.

Insert Figure-Danlos-4 here

B Adverbiaux ambigus ou non-ambigus entre grammaire et discours

Cette annexe élabore sur l'ambiguïté des adverbiaux entre grammaire et discours, sujet esquissé à la Section 4. Dans un premier temps, les adverbiaux pour lesquels il est délicat de cerner tous les emplois (discursifs ou non) sont décrits sur le modèle de la description pour *au contraire* ou *ainsi* présentée à la Section 4. Dans un deuxième temps, les adverbiaux qui présentent un emploi non discursif facilement distinguable de l'emploi connecteur sont répertoriés dans la Table 5. Enfin, la Table 6 dresse la liste d'une cinquantaine d'adverbiaux de LexConn qui sont toujours employés comme connecteurs (dans notre corpus).

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 [27a] où il fonctionne de façon appariée avec l'expression après un temps de réflexion. *D'abord* n'a pas été retenu comme connecteur lorsqu'il sert à mettre un fait au premier plan, comme en [27b] où d'abord est paraphrasable par avant tout.

> [27a] 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.
>
> [27b] Manger bio est d'abord une question de bon sens.

De fait (effectivement) a été annoté comme connecteur dans un exemple comme [28a] De fait n'a pas été retenu comme connecteur lorsqu'il est antonyme de en théorie ou en principe, [28b].

> [28a] 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.
>
> [28b] Après les dévaluations de ces dernières [les livres britannique et italienne], le franc s'est trouvé, de fait, réévalué.

De même (de la même façon) a été annoté comme connecteur lorsqu'il établit un parallèle entre sa phrase hôte et son contexte gauche, [29]. *De même* n'a pas été retenu comme connecteur lorsqu'il est utilisé comme un adverbe de manière, [30a], ou qu'il apparaît dans l'expression il en est de même, [30b].

> [29] 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.
>
> [30a] Si vous en usez bien, il en usera de même.
>
> [30b] Il en est de même, quoique dans une moindre mesure, pour les pays d'Amérique latine.

(Tout) d'un coup (tout à 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, [31a]. *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, [31b].

> [31a] L'Europe a été longtemps une région de départ. Tout d'un coup, elle devient une terre non seulement d'immigration mais de peuplement.
>
> [31b] Et si d'aventure on venait à suivre leurs dangereuses recommandations, on ruinerait d'un coup l'acquis de six années . . .

Enfin (finalement) est connecteur lorsqu'il introduit une conclusion à une énumération d'événements, comme en [32a]. *Enfin* n'a pas été retenu comme connecteur lorsqu'il est utilisé avec une valeur affective pour exprimer la fin d'une longue attente, [32b].

> [32a] 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.
>
> [32b] Le dessinateur Katsuhiro Otomo est enfin consacré à Angoulême.

Inversement est annoté comme connecteur lorsqu'il lexicalise une relation de contraste, [33]. *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, [34a]. *Inversement* n'est pas non plus connecteur, lorsqu'il est modifieur d'un adjectif, [34b].

> [33] 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.
>
> [34a] 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.
>
> [34b] Bien sûr, le taux d'équipement est inversement proportionnel à l'âge de la construction . . .

Parallèlement (simultanément, réciproquement, également) a été annoté comme connecteur lorsqu'il introduit une relation de concomitance entre deux événements d'une même narration, [35a]. *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, [35b].

> [35a]. 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.

[35b]. Les autorités bancaires allemande, américaine, britannique, singapourienne et suisse ont ouvert parallèlement des enquêtes sur ce scandale.

Plutôt est connecteur lorsqu'il lexicalise une relation de contraste entre sa phrase hôte et son contexte gauche, [36a]. *Plutôt* n'a pas été retenu comme connecteur lorsqu'il est employé comme adverbe modifiant un prédicat gradable pour le nuancer, ce prédicat pouvant être un verbe comme en [36b], un adjectif (*Il est plutôt heureux*) ou un adverbe (*On vit plutôt bien à Paris*).

[36a]. C'est absurde! Il n'a trahi personne, il a plutôt été trahi par des membres de son parti

[36b]. Ces derniers temps, elle a plutôt embelli.

Quand même (tout de même) est connecteur lorsqu'il lexicalise une opposition, [37a]. *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, [37b].

[37a] Sangyong, 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...

[37b] Une lettre postée dans le 4ème arrondissement de Paris arrive dès le surlendemain dans le 6ème. C'est quand même formidable, ça !

Surtout est connecteur lorsqu'il introduit une relation de continuation entre deux segments de discours, [38a]. *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 deviner, [38b].

[38a] 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.

[38b] En Allemagne, l'immigration est surtout d'origine européenne.

La Table 5 donne la liste d'autres adverbiaux qui sont ambigus entre grammaire et discours.

Insert Figure-Danlos-5 here

La Table 6 dresse la liste d'une cinquantaine d'adverbiaux de LexConn qui sont toujours employés comme connecteurs dans notre corpus.

Insert Figure-Danlos-6 here

C Les 100 connecteurs les plus fréquents du FTB

Insert Figure-Danlos-7 here