Analyse discursive et informations de factivité

Laurence Danlos

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Analyse discursive et informations de factivité

Laurence Danlos
ALPAGE, Université Paris Diderot (Paris 7), 175 rue du Chevaleret, 750013 Paris
Laurence.Danlos@linguist.jussieu.fr

Résumé. Les annotations discursives proposées dans le cadre de théories discursives comme RST (Rhetorical Structure Theory) ou SDRT (Segmented Dicourse Representation Theory) ont comme point fort de construire une structure discursive globale liant toutes les informations données dans un texte. Les annotations discursives proposées dans le PDTB (Penn Discourse Tree Bank) ont comme point fort d'identifier la "source" de chaque information du texte — répondant ainsi à la question qui a dit ou pense quoi ? Nous proposons une approche unifiée pour les annotations discursives alliant les points forts de ces deux courants de recherche. Cette approche unifiée repose crucialement sur des information de factivité, telles que celles qui sont annotées dans le corpus (anglais) FactBank.

Abstract. Discursive annotations proposed in theories of discourse such as RST (Rhetorical Structure Theory) or SDRT (Segmented Representation Theory Dicourse) have the advatange of building a global discourse structure linking all the information in a text. Discursive annotations proposed in PDTB (Penn Discourse Tree Bank) have the advatange of identifying the "source" of each information — thereby answering to questions such as who says or thinks what ? We propose a unified approach for discursive annotations combining the strengths of these two streams of research. This unified approach relies crucially on factivity information, as encoded in the English corpus FactBank.

Mots-clés : Discours, Analyse discursive, Factivité (véracité), Interface syntaxe-sémantique, RST, SDRT, PDTB, FactBank.

Keywords: Discourse, Discursive analysis, Factuality (vericity), Syntax-semantic interface, RST, SDRT, PDTB, FactBank.

1 Introduction

L'analyse discursive d'un texte s'effectue généralement en deux étapes : la première consiste à segmenter le texte en "unités de discours élémentaires" (EDU, Elementary Discourse Unit), la seconde consiste à construire la "structure du discours ", cette structure reposant sur les "relations de discours" ("relations rhétoriques") qui relient deux segments de discours en spécifiant le rôle d'un segment par rapport à l'autre, spécifiant par là-même l'intention communicative de l'auteur du texte. Un segment de discours est soit une EDU soit un segment complexe groupant plusieurs EDU avec récursivement leur structure discursive. C'est cette approche de l'analyse discursive qui est adoptée dans les deux principales théories du discours, RST (Rhetorical Structure Theory, (Mann & Thompson, 1988; Taboada & Mann, 2006)) et SDRT (Segmented Dicourse Representation Theory, (Asher, 1993; Asher & Lascarides, 2003)) pour lesquelles des corpus ont été annotés manuellement, RST-corpus pour RST en anglais (Carlson *et al.*, 2003), et ANNODIS pour SDRT en français (Péry Woodley *et al.*, 2009).

Parallèlement à ces travaux, des applications récentes du TAL comme la détection d'opinion ou les systèmes de question/réponse ont fait surgir le besoin de savoir qui pense quoi ou qui a dit quoi. Ceci nécessite en premier lieu de pouvoir identifier la "source" d'une information se trouvant dans un texte : est-elle attribuée à l'auteur du texte (le "locuteur") ou à une autre personne mentionnée dans le texte ? De plus, il faut pouvoir déterminer si une information concernant un événement (une éventualité) présente cet événement comme correspondant à une situation du monde ou comme une simple possibilité ou hypothèse ; en termes techniques, il faut pouvoir déterminer la "factivité des événements". Ces deux aspects sont intrinsèquement liés dans la mesure où la factivité d'un événement peut être évaluée différemment, par exemple, par le locuteur et par une source autre que le locuteur. Ainsi, dans *Fred a dit que Jane était la plus belle*, la source de l'information "Jane est la plus belle"

est Fred, la source de l'information que Fred a émis des propos est le locuteur. L'information subjective "Jane est la plus belle" est prise en charge par Fred qui est prêt à défendre son point de vue tandis qu'elle n'est pas prise en charge par le locuteur (voir la notion de "commitment" (Hamblin, 1970)). Ce courant de recherche a donné lieu à des corpus annotés pour la factivité, principalement FactBank pour l'anglais (Saurí, 2008; Saurí & Pustejovsky, 2009) (Section 3).

Le PDTB (Penn Discourse Tree Bank, (PDTB Group, 2008)) est un corpus annoté pour l'anglais qui allie tant des informations sur la structure discursive du texte que des informations sur la source et le degré de factivité des éventualités présentées dans le texte. C'est donc un effort notable pour aller vers une compréhension profonde d'un texte. Le travail présenté dans cet article s'inscrit dans la lignée du PDTB : il vise à jeter les bases théoriques d'un manuel pour annoter tant la structure discursive que les informations de factivité (relativement à une source) d'un texte français. Il se départit néanmoins du PDTB sur les points suivants :

– Le PDTB ne repose pas sur une théorie du discours mais sur un formalisme D-LTAG (Discourse Lexicalized Tree Adjoining Grammar, (Webber, 2004; Forbes-Riley *et al.*, 2006) qui étend un formalisme d'analyse syntaxique (LTAG) au discours. L'objectif du PDTB n'est pas d'annoter la structure discursive globale d'un texte. L'objectif principal est d'annoter pour chaque connecteur de discours explicite (lexicalement marqué) et pour certains (mais pas tous) connecteurs implicites (phonologiquement vides) leurs arguments (Section 2). A l'inverse, notre travail s'inscrit dans la perspective d'annoter une analyse discursive complète d'un texte comme en RST ou SDRT.
– Dans le PDTB, les informations de factivité sont annotées après que les arguments des connecteurs de discours ont été identifiés. Nous montrerons à l'inverse que les informations de factivité doivent être annotées en premier car elles sont primordiales pour déterminer les arguments des connecteurs (Section 4.1).
– Dans le PDTB, les informations sur les sources d'un connecteur de discours et de ses arguments sont aussi annotées après que les arguments de la relation de discours ont été identifiés et avant les informations de factivité. Nous montrerons à nouveau que les informations de factivité doivent être annotées en premier car elles sont primordiales pour déterminer la source des relations de discours et par là-même la source de leurs arguments (Section 4.2).

Pour présenter les différentes positions théoriques que nous défendons, nous nous concentrons dans cet article sur des discours dont au moins une phrase (la première ou la seconde) est de forme *N0$_{hum}$ V W que P = : Fred a dit à Marie que Zoé allait venir pour Noël*, c'est-à-dire une phrase complexe contenant un segment *N0$_{hum}$ V W* dont la tête est un verbe *V* qui sous-catégorise une complétive *P*. Le symbole *N0* désigne le sujet de *V* qui est considéré ici comme humain, *W* désigne un ensemble éventuellement vide de compléments (en incluant les clitiques) ou d'ajouts (en incluant les particules négatives).

La Section 2 discute de la relation de discours $Attribution$ qui est utilisée en RST et SDRT et que nous adoptons pour relier *N0$_{hum}$ V W* à la complétive *P* dans une phrase de forme *N0$_{hum}$ V W que P*. La relation $Attribution$ n'étant pas utilisée dans le PDTB, nous présentons dans cette section les choix alternatifs effectués dans le PDTB. La Section 3 présente les informations de factivité telles qu'elle son annotées dans le corpus FactBank et que nous adoptons. La Section 4 met en avant les positions théoriques que nous défendons en examinant les analyses discursives des discours dont la première phrase est de forme *N0$_{hum}$ V W que P* (Section 4.1) et de ceux dont la seconde phrase est de cette forme (Section 4.2). La Section 5 montre qu'il n'est pas nécessaire d'annoter la source des arguments des relations de discours : ces informations peuvent être déduites des annotations sur la source des relations de discours (lorsqu'on a recours à la relation $Attribution$). La Section 6 conclut en présentant des perspectives pour une annotation de corpus (français) suivant les principes théoriques défendus dans cet article.

2 Relation de discours $Attribution$

Considérons le discours (1) qui enchaîne deux phrases à complétive. Les deux complétives (soulignées) sont liées par la relation $Résultat$ marquée par le connecteur adverbial *du coup*. Par conséquent, il faut que ces deux complétives soient considérées comme des EDU.

(1) Fred a dit à Marie que Zoé allait venir pour Noël. Ensuite, Marc a ajouté que du coup Sue ne viendrait pas.

Il y a donc consensus pour considérer que, dans un discours indirect de forme *N0$_{hum}$ V W que P*, la complétive *P* forme un EDU. Il en est de même pour les discours directs (*"Zoé va venir pour Noël", a dit Fred à Marie*) où

la citation forme un EDU. Par contre, les positions divergent sur la question de savoir si le segment attributif — la séquence *N0*$_{hum}$ *V W* dans un discours indirect ou l'incise de citation dans un discours direct — forme un EDU ou non. Examinons les différentes positions.

– En RST, il est posé que le segment attributif forme un EDU lié au segment attribué par la relation $Attribution$ (Wolf & Gibson, 2006; Redeker & Egg, 2006). Cette relation de discours n'est pas standard puisque le segment attributif ne forme pas un segment autonome (syntaxiquement et sémantiquement) et que $Attribution$ ne joue aucun rôle rhétorique.
– La SDRT adopte la relation $Attribution$ comme la RST mais avec une différence de taille : le segment attributif est considéré comme complet (Hunter *et al.*, 2006). En effet, il est posé qu'il contient une variable existentiellement quantifiée correspondant à l'argument phrastique du verbe du segment attributif, cette variable étant une anaphore liée par (le contenu sémantique de) le segment attribué. Sans entrer dans des détails trop techniques, retenons que cette position revient à analyser un discours indirect tel que *Fred a dit à Marie que Zoé allait venir* comme *Fred l'a dit à Marie, que Zoé allait venir* et un discours direct tel que *"Zoé va venir pour Noël", a dit Fred* comme *"Zoé va venir pour Noël", Fred l'a dit*.
La différence d'utilisation de $Attribution$ en RST et SDRT est illustrée sur l'exemple (2a) avec segmentation en EDU : (2b) présente l'analyse discursive en RST, (2c) celle en SDRT. Ces deux analyses s'accordent sur le point suivant : le premier argument de $Résultat$ n'est pas le segment 2 ; ceci est exclu car la cause du segment 3 *il l'a énervée* construit autour du verbe causatif *énerver* ne peut être qu'un acte de Fred (Pustejovsky, 1995; Danlos, 2000). Les analyses (2b) et (2c) ne diffèrent que par le contenu sémantique qui est éventuellement donné au premier argument de $Attribution$: non existant en RST qui pose que le premier argument de $Résultat$ est le segment complexe formé par l'ensemble segment attributif et segment attribué, noté $[1, 2]$; en revanche, SDRT considère que le premier argument de $Résultat$ est le segment 1 dont le contenu sémantique est considéré comme équivalent à *Fred l'a dit à Marie*.

(2) a. (Fred a dit à Marie)$_1$ que (Zoé allait venir pour Noël)$_2$. (Il l'a énervée)$_3$.
b. $Attribution(1, 2) \wedge Résultat([1, 2], 3)$
c. $Attribution(1, 2) \wedge Résultat(1, 3)$

Par contre, les deux théories proposent la même analyse pour (3a), voir (3b).

(3) a. (Fred est très énervé en ce moment)$_1$. (Jane dit)$_2$ qu'(il est gravement malade)$_3$.
b. $Attribution(2, 3) \wedge Explication(1, 3)$

– Dans le PDTB, la relation $Attribution$ n'est pas utilisée car ce n'est pas une relation de discours standard (Prasad *et al.*, 2006). Les annotations effectuées dans le PDTB sont détaillées ci-dessous, mais, en faisant abstraction du format particulier utilisé, elles sont présentées en (4) et (5) pour (2a) et (3a) respectivement. En (4), les arguments de $Résultat$ correspondent aux deux phrases. En (5), le second argument de $Explication$ est le segment 2 (qui comporte un trait spécifiant son segment attributif *Jane dit que*, cf ci-dessous).

(4) a. (Fred a dit à Marie que Zoé allait venir pour Noël)$_1$. (Il l'a énervée)$_2$.
b. $Résultat(1, 2)$

(5) a. (Fred est très énervé en ce moment)$_1$. Jane dit qu'(il est gravement malade)$_2$.
b. $Explication(1, 2)$

Les analyses proposées en RST et dans le PDTB peuvent être considérées comme équivalentes : nous laissons au lecteur le soin de se convaincre que les analyses (2b) et (4b) ou (3b) et (5b) reviennent fondamentalement au même, ne différant que par l'utilisation ou non de $Attribution$. Par contre, la position prise en SDRT est radicalement différente : les analyses (2b) à la RST et (2c) à la SDRT divergent sur le premier argument de $Résultat$. Nous écartons la position de SDRT pour deux raisons : la première est que l'analyse proposée en (2c) est contre-intuitive et peut donc amener à des erreurs d'annotation ; la seconde est qu'elle n'est pas linguistiquement justifiée pour des discours directs [1] dont l'incise de citation a pour tête un verbe comme *mentir* qui ne permet pas de discours indirect, (6a). Il est donc linguistiquement injustifié d'analyser (6b) comme **"Zoé va venir pour Noël", Fred le mentit* car la citation en (6b) ne correspond pas à un argument du verbe *mentir*. Celui-ci — comme plus de 500 autres verbes — demande une entrée lexicale spéciale pour prendre en compte son emploi comme tête d'une incise de citation (Danlos *et al.*, 2010).

(6) a. *Fred a menti que Zoé allait venir pour Noël.
b. "Zoé va venir pour Noël", mentit Fred.

1. Rappelons que les positions prises en RST, SDRT et dans le PDTB s'appliquent tant aux discours indirects qu'aux discours directs.

Entre les positions quasiment équivalentes prises en RST et dans le PDTB, nous préférons celle prise en RST qui a recours à $Attribution$. En effet, le recours à $Attribution$ permet de construire la structure discursive globale d'un texte en intégrant tout fragment de texte, y compris tout segment attributif [2]. En résumé, nous adoptons la position prise en RST et donc des analyses comme (2b) et (3b) [3].

La relation $Attribution$ est utilisée en RST (ou SDRT) pour relier le segment attributif *N0*$_{hum}$ *V W* et la complétive lorsque *V* est un verbe de discours rapporté comme *dire* mais aussi lorsque *V* est un verbe d'attitude propositionnelle comme *craindre, croire, douter* ou *réaliser*. De ce fait, $Attribution$ ne prend pas en compte les différences sémantiques entre tous ces verbes (leur factivité, par exemple) ni d'ailleurs le fait que le verbe soit sous la portée d'une négation. Il faut donc avoir recours à d'autres mécanismes pour prendre en compte ces différentes propriétés du segment attributif qui influent sur l'analyse discursive. De tels mécanismes seront présentés à la Section 3.

Auparavant, examinons la solution adoptée dans le PDTB où la relation $Attribution$ n'est pas utilisée. Pour chaque connecteur de discours explicite, sont annotés non seulement ses arguments mais aussi la source de la relation de discours marquée par le connecteur et la source de chacun de ses arguments. Ainsi, pour l'exemple (7a) de (Prasad *et al.*, 2006) où le connecteur *while* est souligné, ses argument Arg1 et Arg2 sont repérés par les segments de texte respectivement en italiques et en gras. Le segment attributif, *purchasing agent said*, placé dans une boîte, ne fait partie d'aucun de ces arguments. Ces annotations sont complétées par le tableau en (7b) qui indique la valeur du trait [Source] pour les sources de la relation REL (marquée par *while* et identifiée comme étant $Contraste$), de Arg1 et de Arg2. La valeur "Wr" est utilisée pour l'auteur ("writer") du texte, "Inh" indique que la valeur de [Source] est héritée de celle de REL, "Ot" ("other") est utilisée pour un (ou des) individu(s) autre(s) que l'auteur (il s'agit des purchasing agents pour Arg2). Comme la valeur du trait [Source] de Arg2 est "Ot", cet argument possède un trait qui spécifie son segment attributif, ici *purchasing agent said*. Le tableau comporte d'autres informations relatives à la factivité et la polarité, les traits [Type], [Polarity] et [Determinacy], mais nous ne les détaillerons pas ici, non pas parce que nous les considérons comme non pertinentes mais parce que, d'une part, nous préférons celles mises au point dans FactBank qui sont plus sophistiquées (Section 3) et que, d'autre part, nous pensons que les informations de factivité doivent être déterminées **avant** d'établir quels sont les arguments d'un connecteur de discours (Section 4.1) et quelle est sa source (Section 4.2).

(7) a. *Factory orders and construction outlays were largely flat in December* while purchasing agents said **manufacturing shrank further in October**.

b.

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

Le PDTB apporte donc un grand soin à des annotations permettant de déterminer qui a dit quoi, ce qui est effectivement une question importante pour de nombreuses applications du TAL. Nous nous inspirons de leur approche pour cet aspect de l'annotation discursive. Par contre, dans le PDTB, seuls les arguments de certains connecteurs implicites (phonologiquement vides) sont annotés. Ainsi sont prises en compte les relations de discours qui doivent être inférées entre deux phrases adjacentes juxtaposées à l'intérieur d'un même paragraphe, mais pas celles entre deux phrases adjacentes séparées par une marque de paragraphe (PDTB Group, 2008). Il n'y a donc pas annotation de la structure discursive globale du texte, alors que c'est l'ambition affichée dans le RST-corpus (Carlson *et al.*, 2003) et ANNODIS (Péry Woodley *et al.*, 2009) [4]. De plus, rien ne garantit qu'on puisse déduire la structure discursive globale d'un paragraphe à partir des annotations faites sur les arguments des connecteurs (explicites ou implicites) figurant à l'intérieur de ce paragraphe. Le problème ne vient pas des segments attributifs qui peuvent

2. Signalons toutefois que le recours à $Attribution$ pose un léger problème technique lorsque des compléments ou ajouts du segment attributif apparaissent après la complétive, voir *Fred a prévenu Marie que Zoé allait venir pour Noël par fax et par e-mail*. De tels exemples seront ignorés dans cet article.

3. La relation $Attribution$ est considérée par tous les auteurs qui l'utilisent comme satellite-nucleus (dans les termes de la RST) ou subordonnante (dans les termes de la SDRT). Néanmoins, en RST, les auteurs ne sont pas d'accord sur le nucleus de cette relation subordonnante : (Wolf & Gibson, 2006) considèrent que c'est le segment attributif tandis que (Redeker & Egg, 2006) considèrent que c'est le segment attribué. En SDRT, il est posé que c'est parfois l'un parfois l'autre, selon par exemple l'emploi intensionnel versus évidentiel de *dire*, voir (2) versus (3). Cette position de la SDRT nous paraît tout à fait justifiée, mais, faute de place, nous ne discuterons pas de la question du nucleus de $Attribution$ dans cet article.

4. L'annotation effectuée dans ANNODIS pour le français est donc nettement plus ambitieuse que celle effectuée dans le PDTB. Mais elle ne concerne que 4000 relations de discours contre plus de 40 000 dans le PDTB.

être facilement intégrés à l'analyse discursive en faisant appel à la relation $Attribution$ telle qu'utilisée en RST. Il vient de ce qu'un fragment de texte qui n'est intégré dans aucun argument d'aucun connecteur du paragraphe est tout bonnement et simplement ignoré. En d'autres termes, l'approche du PDTB — identifier les connecteurs explicites et certains connecteurs implicites puis annoter leurs arguments — est orthogonale à l'approche incrémentale préconisée en RST ou SDRT — identifier un nouveau segment de discours et l'intégrer à l'analyse discursive déjà construite. Nous nous inspirons de cette approche incrémentale, qui s'inscrit dans le courant de la sémantique dynamique, pour cet aspect de l'analyse discursive visant à construire la structure discursive globale d'un texte.

En conclusion, comme préconisé en RST (ou SDRT), il semble justifier d'utiliser la relation $Attribution$ tout en ayant recours à d'autres mécanismes pour indiquer les propriétés sémantiques de son premier argument, un segment attributif. Néanmoins, en RST ou SDRT, la question de savoir quelles sont les sources d'une relation de discours et de ses arguments est à tort passée sous silence. Nous considérons qu'il doit être inscrit dans les objectifs de l'analyse discursive de déterminer quelles sont ces sources. Plus précisément, nous considérons que l'analyse discursive doit indiquer la source de chaque relation de discours ; pour indiquer qu'une relation de discours R est attribuée à la source s_j, nous utilisons la notation R_{s_j} dans l'analyse discursive. Pour les arguments d'une relation de discours, nous montrerons à la Section 5 qu'il n'est pas nécessaire d'annoter leur source car celle-ci peut être déduite des informations contenues dans la structure discursive, à savoir les sources des relations de discours et la présence de relation(s) $Attribution$.

Une des difficultés rencontrées dans l'analyse discursive pour des exemples mettant en jeu $Attribution$ est de savoir si c'est le segment attribué ou l'ensemble segment attributif et segment attribué qui est argument d'une relation de discours donnée, voir le contraste entre (2b) et (3b). C'est à cette difficulté que nous allons nous attaquer, entre autres, dans la Section 4. Auparavant, présentions les informations de factivité annotées dans le corpus FactBank.

3 Présentation de FactBank

FactBank est un corpus (anglais) annoté pour la factivité événementielle. Pour chaque événement (éventualité) e_i d'une phrase, les informations de factivité sont données relativement à une source donnée, soit l'auteur soit un (ou des) individu(s) à qui est attribuée les informations concernant e_i. Par exemple, dans une construction à complétive de forme *N0*$_{hum}$ *V W que P*, les informations de factivité concernant l'événement décrit dans la complétive *P* sont données par rapport à l'auteur et au référent de *N0* (si différent de l'auteur). D'une manière générale, une information de factivité est de la forme $f(e_i, s_j) = x$, où s_j désigne l'auteur ou la source des informations concernant e_i [5]. La valeur x d'une information de factivité est une paire $Mod(x)Pol(x)$ contenant une valeur de modalité et une valeur de polarité. Les valeurs de modalité sont au nombre de quatre : certain (CT), probable (PR), possible (PS), non-spécifié (U), avec une relation d'ordre entre ces valeurs. Les valeurs de polarité sont au nombre de trois : positive (+), négative (-) et non-spécifié (u) [6]. A titre d'illustration, pour la phrase en (8) avec sa segmentation en deux EDU, les informations de factivité à la FactBank sont données en (i). L'événement noté e_i est l'événement décrit dans le segment i sans les éventuelles informations de modalité ou de polarité négative ; par exemple, pour le segment 2, e_2 correspond à *Fred ira à Dax*. Pour e_1, la seule source pertinente est l'auteur qui affirme que cet événement est vrai (l'auteur a asserté la phrase donc le segment 1), d'où $f(e_1, auteur) = CT+$. Pour e_2, les informations de factivité sont évaluées relativement à l'auteur et à Jane. Concernant l'auteur, la sémantique du verbe *penser* implique que l'auteur ne s'engage pas [7], ce qui veut dire qu'il ne se prononce pas ou ne veut pas se prononcer sur le contenu propositionnel de *P*, d'où $f(e_2, auteur) = Uu$. Concernant Jane, *penser* implique que Jane juge le contenu propositionnel de *P* comme possible (PS) ; comme *P* en (8) est sous une polarité négative, $f(e_2, Jane) = PS-$.

(8) (Jane pense)$_1$ que (Fred n'ira pas à Dax)$_2$.
(i) $f(e_1, auteur) = CT+ \wedge f(e_2, auteur) = Uu \wedge f(e_2, Jane) = PS-$

5. En fait, il peut y avoir plusieurs sources quand, par exemple, il y a enchâssement d'une complétive dans une complétive, voir *Luc pense que Zoé a dit que P*. Néanmoins, nous laissons de côté pour l'instant ces cas complexes.

6. Les paires $Mod(x)Pol(x)$ sont cependant au nombre de 8 et non de 12 car certaines combinaisons de modalité et polarité ne font pas sens, par exemple $U+$ ou $U-$

7. Nous utilisons *s'engager* comme traduction de *commit* communément utilisé dans la littérature anglophone.

FactBank n'est pas concerné par les relations de discours et par l'analyse discursive : les annotations de factivité sont effectuées phrase par phrase, sans tenir compte du contexte discursif ni des connaissances du monde, même celles concernant "l'autorité morale" d'une source d'information. Ces annotations proviennent uniquement de connaissances en sémantique lexicale, en particulier des propriétés sémantiques des verbes à complétive, et de la présence de marqueurs de modalité (épistémique) ou de polarité, l'interaction entre ces différents facteurs ayant été modélisée. Nous allons montrer que dans une approche de sémantique dynamique, la mise à jour de l'analyse discursive par un nouveau segment de discours doit se dérouler en trois étapes :

(i) détermination des informations de factivité événementielle dans le nouveau segment sur la seule base de connaissances linguistiques,

(ii) mise à jour de la structure discursive en s'appuyant d'une part sur ces informations de factivité événementielle d'autre part sur d'autres connaissances linguistiques, des connaissances pragmatiques et des connaissances du monde ; cette mise à jour doit intégrer l'identification de la source des relations de discours introduites dans la structure discursive,

(iii) révision et/ou complétion des informations de factivité événementielle en fonction de l'analyse discursive.

FactBank n'étant pas concerné par les relations discursive, celles-ci ne sont pas annotées d'informations de factivité. Néanmoins, il est nécessaire de disposer de telles annotations, qui existent d'ailleurs dans le PDTB (Section 2). Nous ajoutons donc dans les tâches de l'annotation de l'analyse discursive l'annotation des informations de factivité concernant les relations de discours (tâche qui doit être effectuée à l'étape ii).

4 Analyse discursive de discours comportant $Attribution$

Nous allons examiner l'analyse discursive de discours comprenant une phrase à complétive de forme *N0*$_{hum}$ *V W que P*, mettant donc en jeu une relation $Attribution$. Dans un premier temps, nous discutons du cas où la phrase à complétive apparaît à l'initiale du discours et est suivie d'une phrase simple (i.e. sans relation $Attribution$) introduite par un connecteur de discours *Conn* éventuellement précédé d'un signe de ponctuation (*Ponct*), soit des discours de forme *N0*$_{hum}$ *V W que P (Ponct) Conn P'*. Nous montrerons que les informations de factivité sont primordiales pour déterminer l'argument gauche de *Conn*. Dans un second temps, nous discutons du cas où la phrase à complétive apparaît après une autre phrase (avec ou sans relation $Attribution$), soit des discours de forme *P. (Conn) N0*$_{hum}$ *V W que (Conn) P'*. Nous montrerons que les informations de factivité sont primordiales pour identifier la source de la relation de discours marquée par *Conn*.

Avant d'entrer dans le vif du sujet, présentons nos conventions de segmentation en EDU, qui s'harmonisent avec celles du PDTB. Un connecteur de discours, qu'il soit de type adverbial ou conjonction, n'est pas considéré comme faisant partie d'une EDU. Cette convention n'est pas celle adoptée dans le projet ANNODIS (Péry Woodley *et al.*, 2009) où un connecteur est intégré à l'EDU correspondant à sa phrase hôte. Ce choix effectué dans ANNODIS préjuge de la portée sémantique des connecteurs. Ainsi, pour le discours en (9a) — dont il sera longuement question à la Section 4.2 —, la segmentation d'ANNODIS donnée en (9b) préjuge à tort que *ensuite* a portée sémantique sur *Jane croit*, contrairement à la nôtre donnée en (9c).

(9) a. Fred ira à Dax pour Noël. Ensuite, Jane croit qu'il ira à Pau.
b. (Fred ira à Dax pour Noël)$_1$. (Ensuite, Jane croit)$_2$ (qu'il ira à Pau)$_3$.
c. (Fred ira à Dax pour Noël)$_1$. Ensuite, (Jane croit)$_2$ qu'(il ira à Pau)$_3$.

Lorsqu'un connecteur adverbial ne se trouve pas en tête de l'EDU qui correspond à sa phrase hôte, comme dans l'exemple (10a) où *ensuite* se trouve au milieu du noyau verbal de sa phrase hôte, il est possible d'avoir recours à une "forme normalisée de discours" (Danlos, 2009) qui fait abstraction de la position de l'adverbial — tout en gardant une trace de cette position car il existe de cas où elle induit une différence de sens (Bras, 2008). La segmentation de (10a) est alors celle donnée en (10b).

(10) a. Fred ira à Dax pour Noël. Il ira ensuite à Pau.
b. (Fred ira à Dax pour Noël)$_1$. ensuite^{interne} (Il ira à Pau)$_2$.

4.1 $Attribution$ dans la première phrase (*N0$_{hum}$ V W que P (Ponct) Conn P'*)

Pour ne pas introduire de bruit dans l'analyse des données, nous allons nous concentrer principalement sur des discours *Fred V W qu'il détestait les grévistes (Ponct) Conn il est syndicaliste*, en nous contentant de faire varier *V W*, soit les propriétés du segment attributif, et de faire varier *Conn* entre la conjonction de subordination *alors que* et l'adverbial *pourtant*, ces deux connecteurs marquant la relation $Contraste$. Nous allons montrer qu'il existe trois interprétations pour ces discours de forme *N0$_{hum}$ V W que P (Ponct) Conn P'*.

Dans la première interprétation, observée uniquement avec *Conn = alors que*, soit une conjonction de subordination et non un connecteur adverbial, le segment attribué est *P alors que P'*, ce qui se traduit en discours direct par *"P alors que P'", V N0 W*, voir (11b). La citation de (11b) ou le discours rapporté de (11a) met en jeu une "violation d'attente", à savoir l'attente $syndicaliste(x) > \neg détesterLesGrévistes(x)$. C'est cette violation d'attente qui légitime la relation $Contraste$ marquée lexicalement par *alors que* et dont la source est Fred. Avec cette interprétation, l'analyse discursive de (11a) est celle donnée en (11c). Les informations de factivité sont données en (11d). La relation $Contraste$ posée par Fred doit être évaluée par rapport à l'auteur : la sémantique de *dire* implique que l'auteur ne s'engage pas sur les propos de Fred, soit $f(Contraste_{Fred}(2,3), auteur) = Uu$, ce qui implique *a priori* $f(e_2, auteur) = Uu \wedge f(e_3, auteur) = Uu$. Par contre, $f(Contraste_{Fred}(2,3), Fred) = CT+$ implique *a priori* $f(e_2, Fred) = CT+ \wedge f(e_3, Fred) = CT+$.

(11) a. (Fred a dit)$_1$ qu'(il détestait les grévistes)$_2$ alors qu'(il est syndicaliste)$_3$.
b. "Je déteste les grévistes alors que je suis syndicaliste", a dit Fred.
c. $Attribution_{auteur}(1, [2,3]) \wedge Contraste_{Fred}(2,3)$
d. $f(Contraste_{Fred}(2,3), auteur) = Uu \wedge f(e_2, auteur) = Uu \wedge f(e_3, auteur) = Uu$
$f(Contraste_{Fred}(2,3), Fred) = CT+ \wedge f(e_2, Fred) = CT+ \wedge f(e_3, Fred) = CT+$.

Nous écartons désormais cette interprétation, en ne retenant pour les exemples suivants que celles mettant en jeu $Attribution_{auteur}(1,2)$, interprétations privilégiées avec *Conn = pourtant*. Il reste alors deux possibilités pour l'argument gauche de $Contraste$: $Contraste_{auteur}(2,3)$ ou $Contraste_{auteur}([1,2],3)$ avec $Attribution(1,2)$. Nous allons montrer que le choix entre l'une ou l'autre de ces possibilités dépend des informations de factivité venant du segment attributif *N0 V W*.

Commençons par examiner des exemples où e_2 et e_3 mettent en jeu une violation d'attente, e.g. (Fred déteste les grévistes)$_2$ et (Fred est syndicaliste)$_3$. Considérons (12a) construit autour du verbe *réaliser* qualifié de "factif" dans la littérature. Les informations à la FactBank sont données en (12b). L'auteur croit en la véracité de $e2$ et $e3$ qui mettent en jeu la violation d'attente $syndicaliste(x) > \neg détesterLesGrévistes(x)$. Il peut donc poser la relation $Contraste$ entre les segments 2 et 3, ce qui débouche sur l'analyse donnée en (12c). L'intention communicative de l'auteur est de montrer que Fred est incohérent avec lui-même.

(12) a. (Fred a réalisé)$_1$ qu'(il détestait les grévistes)$_2$. Pourtant (il est syndicaliste)$_3$.
b. $f(e_1, auteur) = CT+$
$f(e_2, Fred) = CT+ \wedge f(e_2, auteur) = CT+$
$f(e_3, auteur) = CT+$
c. $Attribution_{auteur}(1,2) \wedge Contraste_{auteur}(2,3)$

Passons maintenant à (13a) construit autour de *prétendre* qui est un "factif négatif", voir les informations de factivité concernant e_2 en (13b) où $f(e_2, auteur) = CT-$ indique que d'après l'auteur ce n'est certainement pas le cas que Fred déteste les grévistes. L'auteur ne peut donc pas poser qu'il y a violation de l'attente $syndicaliste(x) > \neg détesterLesGrévistes(x)$. Il ne peut donc poser qu'un contraste entre le fait que Fred a prétendu qu'il détestait les grévistes et e3, ce qui débouche sur l'analyse donnée en (13c). L'intention communicative de l'auteur est de montrer que Fred a menti.

(13) a. (Fred a prétendu)$_1$ qu'(il détestait les grévistes)$_2$. Pourtant (il est syndicaliste)$_3$.
b. $f(e_2, Fred) = CT+ \wedge f(e_2, auteur) = CT-$
c. $Attribution_{auteur}(1,2) \wedge Contraste_{auteur}([1,2],3)$

Finalement, considérons (14a) construit autour de *dire* avec les informations de factivité pour $e2$ données en (14b) : l'auteur ne se prononce pas sur la véracité de $e2$. Le discours (14a) ne suffit pas à lui-seul pour déterminer l'opinion

de l'auteur, qui peut d'ailleurs être hésitant comme en témoigne le fait que (14a) peut être prolongé par une phrase qui indique explicitement son hésitation, (14c). Nous considérons donc (14a) comme ambigu, ambiguïté traduite dans l'analyse discursive en (14d) comportant une disjonction. Cette disjonction se traduit par les informations de factivité données en (14e).

(14) a. (Fred a dit)$_1$ qu'(il détestait les grévistes)$_2$. Pourtant (il est syndicaliste)$_3$.
b. $f(e_2, Fred) = CT+ \wedge f(e_2, auteur) = Uu$
c. Fred a dit qu'il détestait les grévistes. Pourtant il est syndicaliste. Soit il est incohérent avec lui-même soit il a menti.
d. $Attribution_{auteur}(1,2) \wedge (Contraste_{auteur}([1,2],3) \vee Contraste_{auteur}(2,3))$
e. $f(Contraste_{auteur}([1,2],3), auteur) = PS+ \wedge f(Contraste_{auteur}(2,3)), auteur) = PS+$

Tournons-nous rapidement vers des exemples où le segment 2 est le contraire de 3, e.g. (il pleuvait)$_2$ et (il ne pleuvait pas)$_3$. Comme un individu ne peut pas croire simultanément en une chose et son contraire [8], on a $POL(f(e_2, s_j)) = -POL(f(e_3, s_j))$ avec $-u = u$. Cette règle explique l'incohérence de *#Fred a réalisé qu'il pleuvait. Pourtant, il ne pleuvait pas.* construit avec le verbe factif *réaliser*. Cette règle explique aussi que le discours en (15a) construit avec *dire* est non ambigu, contrairement au discours en (14a) construit aussi avec *dire*. En effet, l'auteur ne pouvant poser $Contraste(2,3)$, il pose $Contraste([1,2],3)$, voir l'analyse en (15b) qui indique que l'intention communicative de l'auteur est de montrer que Fred a menti.

(15) a. (Fred a dit)$_1$ qu'(il pleuvait)$_2$. Pourtant (il ne pleuvait pas)$_3$.
b. $Attribution_{auteur}(1,2) \wedge Contraste_{auteur}([1,2],3)$

Cette discussion sur les analyses discursives des discours de forme *(N0 V W)*$_1$ *que (P)*$_2$ *(Ponct) Conn (P')*$_3$ amène les remarques suivantes respectivement sur le PDTB et SDRT :

1) PDTB : Tout annotateur "naïf" se reposant sur son intuition immédiate se jetterait probablement sur la violation d'attente $syndicaliste(x) > \neg détesterLesGrévistes(x)$ pour poser systématiquement $Contraste(2,3)$ dans les exemples (12)-(14). Or cette analyse est fausse pour (13) et (14). Ce point est à relier au suivant : dans le PDTB, les arguments Arg1 et Arg2 d'une relation de discours REL sont annotés avant les informations de factivité (qui sont ajoutées sur REL, Arg1 et Arg2, voir le tableau en (7b)). Les exemples (12)-(14) montrent qu'à l'inverse il faut annoter les informations de factivité **avant** d'annoter les arguments d'une relation de discours donnée.

2) SDRT : En SDRT, la relation $Contraste$ est considérée comme "véridique", ce qui est défini dans (Asher & Lascarides, 2003) comme une relation impliquant la véracité (du contenu propositionnel) de ces deux arguments. Néanmoins, les discours mis en avant par ces auteurs ne sont que des discours simples dans la mesure où ils ne mettent jamais en jeu la relation $Attribution$. En d'autres termes, dans ces discours, tout est attribué à l'auteur, les relations de discours et leurs arguments, ce qui permet d'ignorer le fait que les informations de véracité (factivité) doivent être évaluées relativement à plusieurs sources. Il faut donc réviser la notion de relation de discours véridique en prenant en compte l'évaluation des informations de factivité relativement à plusieurs sources (Danlos, 2011).

4.2 $Attribution$ dans la seconde phrase (*P. (Conn) N0*$_{hum}$ *V W que (Conn) P'.*)

Pour ne pas introduire de bruit dans l'analyse des données, nous allons nous concentrer uniquement sur des discours avec $Conn = ensuite$, ce connecteur marquant la relation $Narration$ de succession temporelle entre deux événements. Considérons d'abord des phrases *Ensuite, N0 V W que P* (avec une relation $Attribution$) qui sont énoncées dans un contexte gauche comprenant aussi une relation $Attribution$, par exemple (16a) (le contexte gauche, i.e. *P*, est mis en italiques). Dans cet exemple, *ensuite* a portée sur *elle a cru* : c'est un ajout syntaxique et sémantique sur le noyau verbal, qui peut être déplacé à l'intérieur de celui-ci sans changement de sens, (16b), mais qui ne peut pas être déplacé à l'intérieur la complétive sans induire un changement de sens important, (16c).

(16) a. *Jane a (d'abord) cru que Fred irait à Dax pour Noël.* Ensuite, elle a cru qu'il irait à Pau.
b. = *Jane a (d'abord) cru que Fred irait à Dax pour Noël.* Elle a ensuite cru qu'il irait à Pau.

8. Citons (Prabhakaran *et al.*, 2010) : "We cannot both believe something and not believe it : *#John won't be here, but nevertheless I think he may be here.*"

c. $\neq$ *Jane a (d'abord) cru que Fred irait à Dax pour Noël.* Elle a cru qu'ensuite il irait à Pau.

Le discours (16a) décrit la succession temporelle de croyances de Jane. En suivant sa segmentation donnée en (17a), il doit donc être analysé avec $Narration([1,2],[3,4])$. La relation $Narration$ est posée par l'auteur, tout comme les deux relations $Attribution$. Au total, l'analyse discursive de (16a) est celle présentée en (17b).

(17) a. *(Jane a (d'abord) cru)*$_1$ *que (Fred irait à Dax pour Noël)*$_2$. Ensuite, (elle a cru)$_3$ qu'(il irait à Pau)$_4$.
b. $Attribution_{auteur}(1,2) \wedge Attribution_{auteur}(3,4) \wedge Narration([1,2],[3,4])$

Considérons maintenant des phrases *Ensuite, N0 V W que P* qui sont énoncées dans un contexte gauche ne comprenant pas de relation $Attribution$, par exemple (18a). Dans cet exemple, le déplacement de *ensuite* à droite de *croit* débouche sur une incohérence, (18b), mais le déplacement de *ensuite* à l'intérieur de la complétive n'induit pas de différence de sens majeur, (18c).

(18) a. *Fred ira à Dax pour Noël.* Ensuite, Jane croit qu'il ira à Pau.
b. # *Fred ira à Dax pour Noël.* Jane croit ensuite qu'il ira à Pau.
c. $=$ *Fred ira à Dax pour Noël.* Jane croit qu'ensuite il ira à Pau.

L'analyse syntaxique de (18a) pose problème pour l'interface syntaxe-sémantique mais nous n'aborderons pas ce sujet ici (Dinesh *et al.*, 2005). En (18a) et (18c), l'événement du segment 1 (le voyage de Fred à Dax) est présenté comme précédant dans le temps l'événement du segment 3 (le voyage de Fred à Pau), on a donc $Narration(1,3)$. Il reste à examiner quelle est la source de $Narration$. Commençons par un exemple comme (18c) où *ensuite* se situe dans la complétive. Une source, que ce soit l'auteur ou non, ne peut poser de relation de succession temporelle entre deux événements que si cette personne est au courant des deux événements en jeu, ou tout du moins si elle pense qu'ils se sont passés (ou vont se passer) certainement, probablement ou peut-être [9]. Examinons les conséquences de cette affirmation pour (18c), répété en (19a) avec segmentation en EDU et dont les informations de factivité concernant e_3 sont données en (19b). L'auteur ne peut pas poser une relation de succession temporelle entre e_1 et e_3 car il ne s'engage pas sur le statut factuel de e_3. En revanche, Jane peut poser une telle relation car elle juge e_3 probable. L'analyse discursive est donc celle donnée en (19c) où la source de la relation $Narration$ est Jane. L'auteur ne prend pas en charge cette relation, (19d). Il peut d'ailleurs la mettre en question dans une troisième phrase, voir (19e), qui amène à une révision de la valeur de factivité donnée en (19d). Par ailleurs, les informations de factivité événementielle ne disent rien sur $f(e_1, Jane)$ puisque la source de 1 est l'auteur, mais l'analyse discursive avec $Narration_{Jane}(1,3)$ amène à poser $f(e_1, Jane) = CT+$ (ou à la rigueur $PR+$ou $PS+$). Ces données montrent que les informations de factivité événementielle doivent être révisées ou complétées après l'analyse discursive.

(19) a. *(Fred ira à Dax pour Noël)*$_1$. (Jane croit)$_2$ qu'ensuite (il ira à Pau)$_3$.
b. $f(e_3, Jane) = PR+ \wedge f(e_3, auteur) = Uu$
c. $Attribution_{auteur}(2,3) \wedge Narration_{Jane}(1,3)$
d. $f(Narration_{Jane}(1,3), auteur) = Uu$
e. *Fred ira à Dax voir sa mère.* Jane croit qu'ensuite il ira à Pau voir son père. *Mais je pense qu'il ira voir son père avant d'aller voir sa mère.*

Passons à (20a) dont les informations factuelles pour e_3 sont données en (20b). La mère de Fred n'étant pas au courant de l'événement e_3, elle ne peut pas poser de succession temporelle entre $e1$ et $e3$. La relation de succession temporelle ne peut donc être posée que par l'auteur, voir l'analyse en (20c).

(20) a. (Fred ira à Dax pour Noël voir sa mère)$_1$. (Celle-ci ne sait pas)$_2$ qu'ensuite (il ira à Pau voir son père)$_3$.
b. $f(e_3, MèredeFred) = Uu \wedge f(e_3, auteur) = CT+$
c. $Attribution_{auteur}(2,3) \wedge Narration_{auteur}(1,3)$

L'exemple (20a) contredit une idée qu'on aurait pu avoir *a priori*, à savoir qu'un adverbial qui se situe dans une complétive *P* introduite par *N0 V W* a pour source le référent de *N0* [10]. Le contraste entre (19c) et (20c) montre que

9. Par souci de simplification, nous ignorons dans cette discussion la succession temporelle d'événements dont un au moins est sous la portée d'une polarité négative.

10. (Bonami & Godard, 2008) mettent aussi en avant des exemples où un adverbe évaluatif tel que *bizarrement* se situe dans une complétive sans avoir pour source le référent de *N0*.

les informations de factivité sont nécessaires pour déterminer la source des relations de discours (en plus d'être nécessaires pour déterminer leurs arguments, comme montré à la section précédente).

Examinons (21a) où la seconde phrase comporte une coordination. Son analyse discursive, construite par analogie avec celle de (19a), est donnée en (21a). Celle-ci est un contre-exemple à la contrainte de la frontière droite posée en SDRT (Asher & Lascarides, 2003). Rappelons que ces auteurs étudient seulement des discours ne comportant que des assertions de l'auteur et que donc toutes les relations de discours sont attribuées à l'auteur. La contrainte de la frontière droite a pour effet d'interdire $R(\alpha, \beta) \wedge R'(\alpha, \gamma)$ lorsque R est coordonnante. Si cette interdiction est probablement valable lorsque seules sont considérées des assertions de l'auteur, soit R_{auteur} et R'_{auteur}, elle ne l'est plus lorsqu'on considère des sources autres que l'auteur : voir l'analyse en (21b) avec $Narration_{Jane}(1,3) \wedge Narration_{Zoé}(1,5)$ où $Narration$ est sans conteste coordonnante. Il est donc nécessaire d'étudier la validité de la contrainte de la frontière droite dans des discours mettant en jeu la relation $Attribution$.

(21) a. *(Fred ira à Dax pour Noël)*$_1$. (Jane croit)$_2$ qu'ensuite (il ira à Pau)$_3$ et (Zoé croit)$_4$ qu'(ensuite) (il ira à Bayonne)$_5$.
b. $Attribution_{auteur}(2,3) \wedge Narration_{Jane}(1,3) \wedge Attribution_{auteur}(4,5) \wedge Narration_{Zoé}(1,5)$

Tournons-nous maintenant vers la source de *ensuite* dans des exemples de forme *Ensuite, N0 V W que P* en considérant les trois discours obtenus à partir de (19a)-(21a) en déplaçant *ensuite* en tête de la phrase, voir (22). En (22a), qui répète (18a), la source de *ensuite* n'est pas claire : on peut hésiter entre l'auteur ou Jane [11]. En (22b), la source de *ensuite* est l'auteur comme en (20a). Enfin, en (22c), la source de *ensuite* est l'auteur alors que c'est Jane ou Zoé en (21a).

(22) a. *Fred ira à Dax pour Noël.* Ensuite, Jane croit qu'il ira à Pau.
b. *Fred ira à Dax pour Noël.* Ensuite, Jane ne sait pas qu'il ira à Pau.
c. *Fred ira à Dax pour Noël.* Ensuite, Jane croit qu'il ira à Pau et Zoé croit qu'il ira à Bayonne.

En conclusion, la position du connecteur, en tête de la phrase complexe ou en tête de la complétive, n'induit pas de différence de sens majeur à l'exception du fait qu'elle peut éventuellement influer sur la source du connecteur (et donc sur la source de la relation de discours qu'il exprime).

5 Source des arguments d'une relation de discours

Montrons brièvement qu'une structure discursive faisant appel à $Attribution$ et composée d'une conjonction de formules $R_{s_j}(\alpha, \beta)$ où les sources s_j des relations R sont annotées suffit à déterminer la source des arguments α et β de R_{s_j}, quelle que soit R, et par là-même récursivement la source de chaque EDU du texte.

1) $R = Attribution$. Le premier argument α est alors forcément une EDU. En effet, $Attribution$ amène à décomposer un enchâssement de plusieurs complétives : (23a) reçoit la segmentation en (23b) et l'analyse discursive en (23c).

(23) a. Fred pense que Zoé a dit qu'il neigeait.
b. (Fred pense)$_1$ que (Zoé a dit)$_2$ qu'(il neigeait)$_3$.
c. $Attribution_{auteur}(1, [2,3]) \wedge Attribution_{Fred}(2,3)$

$Attribution_{s_j}(\alpha, \beta)$ indique que la source de α est s_j. Le second argument β peut être une EDU ou un segment complexe. Lorsque β est une EDU, sa source est déterminée par le segment attributif α. Si celui-ci est de forme *N0 V W* alors la source de β est le référent de *N0*. Lorsque β est un segment complexe, la structure discursive de ce segment permet récursivement de déterminer la source de ses EDU. Ainsi, en (23a), la source de 1 est l'auteur, celle de 2 Fred et celle de 3 Zoé.

2) $R \neq Attribution$. Si les arguments α et β de R sont des EDU, alors soit ils apparaissent l'un et/ou l'autre dans la structure discursive comme argument d'une relation $Attribution$, par exemple $Attribution_{s_k}(\gamma, \beta)$, et l'étape précédente s'applique alors pour déterminer leur source, soit ce n'est pas le cas et leur source est alors l'auteur. Si les arguments α et β ne sont pas des EDU, le mécanisme s'applique récursivement.

11. Ce type d'hésitation sur la source d'un connecteur a amené les auteurs du PDTB à faire un choix par défaut, à savoir l'auteur.

6 Conclusion et perspectives

En nous inspirant de théories sur le discours (RST et SDRT) d'une part, et du corpus annoté PDTB d'autre part, nous avons mis en évidence que le calcul ou l'annotation de l'analyse discursive d'un texte devait produire une structure hiérarchique connexe où tous les éléments d'information devaient être reliés — ce qui n'est pas le cas dans le PDTB — et où les sources des différentes relations de la structure discursive devaient être calculées ou annotées — ce qui n'est pas le cas en RST ou SDRT. En corollaire, nous avons montré qu'une structure discursive qui utilise la relation $Attribution$ et qui définit la source de chaque relation de discours est suffisante pour déterminer la source de chaque EDU du texte (Section 5).

Nous avons montré de plus que les informations de factivité à la FactBank étaient primordiales non seulement pour identifier les arguments d'une relation de discours donnée (Section 4.1) mais aussi pour identifier sa source (Section 4.2), et donc, d'après notre corollaire, identifier la source de chaque EDU du texte. A rebours, nous avons montré que les informations de factivité à la FactBank devaient être révisées ou complétées après le calcul ou l'annotation de la structure discursive.

Dans nos perspectives de recherche, nous incluons :

1) L'extension du travail présenté ici à d'autres constructions que les discours indirects de forme *N0*$_{hum}$ *V W que P* discutées dans cet article. Nous avons déjà amorcé l'étude des discours directs dans (Danlos *et al.*, 2010). Il reste les constructions à infinitive (avec des "verbes à montée" ou "des verbes à contrôle"), les constructions impersonnelles et les constructions mettant en jeu un "verbe de discours" dont le sujet réfère à un événement (*Ceci a précédé/expliqué/prouvé que P*) (Danlos, 2006). Pour chacune de ces constructions, qui présentent un événement enchâssé sous un autre, il faut en premier lieu déterminer quelle est la segmentation en EDU — en se posant la question de savoir si cette segmentation peut être obtenue à partir d'une analyse syntaxique profonde. Il faut ensuite déterminer les différentes analyses discursives possibles, en s'appuyant sur les informations de factivité comme nous l'avons fait ici pour les discours indirects.

2) La réalisation d'un corpus annoté discursivement pour le français qui combine les avantages de RST, de SDRT et du PDTB en suivant les conclusions de cet article. Pour cette tâche, nous pouvons bénéficier du corpus French TimeBank (Bittar, 2010) qui peut servir de première étape pour fournir un corpus French FactBank — le corpus anglais FactBank a été réalisé comme une seconde couche d'annotations sur TimeBank (Saurí & Pustejovsky, 2009). Les informations de factivité du French FactBank recevant ultimement une autre couche d'annotations concernant la structure discursive. Cette tâche est de longue haleine et ne saurait être confiée à des annotateurs "naifs", mais est-il besoin de rappeler que seuls des corpus annotés proprement permettront aux techniques d'apprentissage (supervisées) de fournir des résultats dignes d'intérêt ?

Remerciements

Je remercie chaleureusement Owen Rambow avec qui j'ai eu des discussions tout à fait éclairantes sur différents thèmes abordés dans cet article, Sylvain Kahane et Philippe Muller pour leur relecture attentive, et les reviewers anonymes de TALN.

Références

ASHER N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht : Kluwer.

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

BITTAR A. (2010). *Building a TimeBank for French : A Reference Corpus Annotated According to the ISO-TimeML Standard*. PhD thesis, Université Paris Diderot (Paris 7).

BONAMI O. & GODARD D. (2008). Lexical semantics and pragmatics of evaluative adverbs. In L. MCNALLY & C. KENNEDY, Eds., *Adverbs and Adjectives : Syntax, Semantics, and Discourse*, p. 274–304. Oxford University Press.

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

CARLSON L., MARCU D. & OKUROWSKI M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. VAN KUPPEVELT & R. SMITH, Eds., *Current Directions in Discourse and Dialogue*, p. 85–112. Kluwer Academic Publishers.

DANLOS L. (2000). Event coreference in causal discourses. In P. BOUILLON & F. BUSA, Eds., *The Language of Word Meaning*, p. 216–241. Cambridge University Press.

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

DANLOS L. (2009). D-STAG : un formalisme d'analyse automatique de discours basé sur les TAG synchrones. *Revue TAL*, **50**(1), 111–143.

DANLOS L. (2011). Factivity information and veridicality of discourse relations. In *Proceedings of the Constraints in Discourse workshop CID 2011*, Agay-Roches rouges, France.

DANLOS L., SAGOT B. & STERN R. (2010). Analyse discursive des incises de citations. In *Actes du Second Colloque Mondial de Linguistique Française"*, New-Orleans, USA.

DINESH N., LEE A., MILTSAKAKI E., PRASAD R., JOSHI A. & WEBBER B. (2005). Attribution and the (non)alignment of syntactic and discourse arguments of connectives. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II : Pie in the Sky*, p. 29–36, Ann Arbor, Michigan : Association for Computational Linguistics.

FORBES-RILEY K., WEBBER B. & JOSHI A. (2006). Computing discourse semantics : The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, **23**(1).

HAMBLIN C. L. (1970). *Fallacies*. London : Methuen.

HUNTER J., ASHER N., REESE B. & DENIS P. (2006). Evidentiality and intensionality : Two uses of reportative constructions in discourse. In *Proceedings of the Constraints in Discourse Workshop (CID'06)*, Maynoth, Ireland.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical Structure Theory : Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

PDTB GROUP (2008). *The Penn Discourse Treebank 2.0 Annotation Manual*. Rapport interne, Institute for Research in Cognitive Science, University of Philadelphia.

PÉRY WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO DAC L.-M., LE DRAOULEC A., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ COURET M., VIEU L. & WIDLÖCHER A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *Proceedings of TALN 2009*, p. 190–196, Senlis, France.

PRABHAKARAN V., RAMBOW O. & DIAB M. (2010). Automatic committed belief tagging. In *Proceedings of COLING 2010*, Pékin, Chine.

PRASAD R., DINESH N., LEE A., JOSHI A. & WEBBER B. (2006). Attribution and its annotation in the Penn Discourse Treebank. *Revue TAL*, **47**(2).

PUSTEJOVSKY J. (1995). *The generative Lexicon*. Cambridge : The MIT Press.

REDEKER G. & EGG M. (2006). Says who ? on the treatment of speech attributions in discourse structure. In *Proceedings of the Constraints in Discourse Workshop (CID'06)*, Maynoth, Ireland.

SAURÍ R. (2008). *A Factuality Profiler for Eventualities in Text*. PhD thesis, Brandeis University.

SAURÍ R. & PUSTEJOVSKY J. (2009). FactBank : A corpus annotated with event factuality. *Language Resources and Evaluation*, **43**, 227–268.

TABOADA M. & MANN W. (2006). Rhetorical Structure Theory : Looking back and moving ahead. *Discourse Studies*, **8**(3), 423–459.

WEBBER B. (2004). D-LTAG : extending lexicalized TAG to discourse. *Cognitive Science*, **28**(5), 751–779.

WOLF F. & GIBSON E. (2006). *Coherence in Natural Language : Data Structures and Applications*. London : The MIT Press.