La ressource EXPLICADIS, un corpus annoté spécifiquement pour l'étude des relations de discours causales

Caroline Atallah

La ressource EXPLICADIS, un corpus annoté spécifiquement pour l'étude des relations de discours causales

Caroline Atallah[1]
(1) CLLE-ERSS (UMR 5263), Université de Toulouse
caroline.atallah@univ-tlse2.fr

Résumé. Dans le but de proposer une caractérisation des relations de discours liées à la causalité, nous avons été amenée à constituer et annoter notre propre corpus d'étude : la ressource EXPLICADIS (EXPlication et Argumentation en DIScours). Cette ressource a été construite dans la continuité d'une ressource déjà disponible, le corpus ANNODIS. Proposant une annotation plus précise des relations causales sur un ensemble de textes diversifiés en genres textuels, EXPLICADIS est le premier corpus de ce type constitué spécifiquement pour l'étude des relations de discours causales.

Abstract.

A corpus specifically annotated for causal discourse relations studies : the EXPLICADIS resource.
In order to offer a characterization of causal discourse relations, we created and annotated our own corpus : EXPLICADIS (EXPlanation and Argumentation in DIScourse). This corpus was built in the continuity of a readily available corpus, the ANNODIS corpus. Providing a more precise annotation of causal relations in a set of texts that are representative of multiple textual genres, EXPLICADIS is the first corpus of its kind built specifically for causal discourse relations studies.

Mots-clés : annotation de corpus, discours, relations causales.

Keywords: corpus annotation, discourse, causal relations.

1 Introduction

Dans cet article, nous présentons une nouvelle ressource annotée spécifiquement pour l'étude des relations de discours causales. Cette ressource a été constituée dans le cadre du projet qui lui a donné son nom, le projet EXPLICADIS (EXPLICation et Argumentation en DIScours) [1].

Il nous semble important de distinguer deux types d'objectifs que peut viser la constitution d'un nouveau corpus annoté : celui de fournir des données directement exploitables pour un apprentissage automatique ou celui de tester des hypothèses et de faire émerger des données de nouveaux phénomènes à étudier. Le premier objectif implique que les recherches soient suffisamment avancées sur le sujet et que le modèle théorique utilisé soit stabilisé. L'annotation du corpus vise alors à constituer ce que l'on appelle un *corpus de référence*. Le second objectif, quant à lui, s'inscrit dans une perspective principalement *expérimentale*, il s'agit d'apprendre à partir des données et non de faire apprendre [2].

Bien que plusieurs projets antérieurs d'annotation discursive, tels que le RST TreeBank (Carlson *et al.*, 2001) ou le Penn Discourse TreeBank (PDTB, Miltsakaki *et al.*, 2004), affichent clairement l'objectif de constituer des corpus de référence, il nous semble, qu'au vu de l'état actuel des recherches sur le discours, l'annotation discursive ne peut, pour l'instant, que s'inscrire dans la seconde perspective.

Le corpus EXPLICADIS a été construit suite à l'exploitation du corpus ANNODIS [3] (ANNotation DIScursive de corpus, Péry-Woodley *et al.*, 2009, 2011; Afantenos *et al.*, 2012) pour l'étude spécifique des relations de discours causales. Même si, pour les raisons évoquées plus haut, un corpus tel que ANNODIS ne peut, à notre sens, être qualifié de *corpus de référence*, ce n'est pas pour autant qu'il ne présente pas d'intérêt. Bien au contraire, ce type de ressource ouvre de nombreuses

1. Ce projet a été co-financé par le PRES toulousain et la région Midi-Pyrénées (2010-2013).

2. Bien entendu, les données peuvent être exploitées pour des techniques d'apprentissage automatique, mais il faut alors être conscients que les résultats ne peuvent au mieux qu'être approximatifs.

3. La ressource ANNODIS est disponible sur le site REDAC (Ressources Développées à CLLE-ERSS : `http://redac.univtlse2.fr/`), sous licence « Creative Commons ».

possibilités d'exploitations, ou plus précisément d'*explorations*. Nous parlerons pour cela de *corpus exploratoire*, plutôt que de *corpus de référence*.

Dans cet article, nous défendons, à travers la présentation du corpus EXPLICADIS, l'intérêt de constituer mais aussi d'exploiter un corpus de type *exploratoire*. Dans les sections qui suivent, nous présenterons les différentes étapes de notre démarche – exploitation du corpus ANNODIS, puis constitution du corpus EXPLICADIS et, enfin, exploitation de ce dernier – et montrerons comment chacune d'entre elles a contribué à faire avancer les recherches sur les relations de discours causales, recherches qui seront encore amenées à évoluer sur la base de cette ressource prochainement disponible en ligne.

2 Exploitation du corpus ANNODIS pour l'étude des relations causales

Dans le cadre de nos recherches, nous nous intéressons aux relations de discours liées à la causalité et cherchons à caractériser celles-ci, sur la base de données attestées, dans le cadre théorique proposé par la SDRT (*Segmented Discourse Representation Theory*, Asher & Lascarides, 2003). Pour parvenir à un tel objectif, il était nécessaire de nous appuyer sur un corpus annoté en relations de discours, et notamment en relations causales. Le corpus ANNODIS, premier corpus de textes en français enrichis d'annotations discursives, offrait une base pertinente pour l'étude des relations de discours. Nos premières analyses sur les relations causales ont ainsi été menées à partir des annotations réalisées dans le cadre de ce projet. Nous présenterons brièvement cette ressource, puis nous verrons pourquoi l'exploitation de ces données nous a rapidement amenée à envisager la constitution d'un nouveau corpus, un corpus enrichi cette fois spécifiquement pour l'étude des relations causales.

Le projet ANNODIS a été constitué selon deux approches. L'une d'entre elles, dite *approche ascendante*, a abouti à la constitution d'un corpus enrichi avec des relations discursives, corpus sur lequel nos premières analyses se sont appuyées. Il s'agissait de construire une représentation de la structure du discours en liant des unités discursives entre elles par des relations rhétoriques. Pour ce faire, des textes ont d'abord été segmentés en unités de discours élémentaires. Puis, les segments constitués ont été liés entre eux par des relations de discours.

La segmentation ainsi que l'annotation des relations discursives ont été réalisées selon les recommandations fournies par un manuel rédigé spécifiquement pour le projet (Muller *et al.*, 2012). Avant de trouver sa forme définitive, ce manuel a été testé par des annotateurs dits *exploratoires* (notés par la suite "annotateurs A et B"), puis modifié suite à cette première phase d'annotation. Le guide finalisé, de nouveaux textes, au nombre de 42, ont été segmentés, puis triplement annotés par de nouveaux annotateurs. Deux annotations, dites *naïves*, ont été réalisées par des étudiants ne possédant pas de connaissances particulières sur les théories du discours (notés "annotateurs 1, 2 et 3"). Puis, les membres du projet ont eux-mêmes contribué à l'annotation, fournissant une troisième annotation, dite *experte*, de ces 42 textes. En parallèle, les annotateurs A et B ont poursuivi leurs annotations pour fournir au final un ensemble supplémentaire de 44 textes annotés. Les annotateurs experts ont également procédé à une nouvelle annotation de ces 44 textes. La ressource finale comporte ainsi 86 textes segmentés et au moins doublement annotés en relations de discours.

Parmi les dix-sept relations proposées dans le manuel d'annotation, figurent quatre relations causales : *Explication*, *Résultat*, *Explication** et *Résultat**, codées respectivement `explanation`, `result`, `explanation*` et `result*`. Afin de guider les annotateurs, le manuel propose, pour chaque relation, une définition, des exemples, ainsi que, parfois, une liste de marqueurs potentiels. Nous reprenons ci-dessous les éléments principaux concernant les relations causales.

Explication (`explanation`)
- — Définition : La relation d'Explication lie deux segments dont le second (celui qui est attaché) explique le premier (la cible) de façon explicite ou non (indépendamment de l'ordre de présentation). Le premier argument de la relation est le segment expliqué, et le deuxième la cause supposée. Si l'effet est attaché à la cause et non l'inverse, on a la relation de Résultat.
- — Exemple : [L'équipe a perdu lamentablement hier.]_1 [Elle avait trop de blessés.]_2 `explanation(1,2)`
- — Marqueurs possibles : *car, parce que, à cause de, du fait de, par la faute de, grâce à, si... c'est parce que..., depuis* (si causalité évidente)

Résultat (`result`)
- — Définition : La relation Résultat caractérise des liens entre deux segments introduisant deux éventualités (événements ou états) dont la 2ème résulte de la première.
- — Exemple : [Nicholas avait bu trop de vin]_1 [et a donc dû rentrer chez lui en métro.]_2 `result(1,2)`
- — Marqueurs possibles : *du coup, donc, par conséquent, en conséquence, par suite, à la suite de quoi*

Explication* (`explanation*`) et ***Résultat**** (`result*`)

— Définition : Dans certains cas, les effets sémantiques du lien rhétorique qui s'établit entre deux segments ne portent pas sur les événements décrits dans les segments, mais sur les actes de parole eux-mêmes.

— Exemples :
[Ferme la porte,]_1 [il fait froid]_2 `explanation*(1,2)`
[Il fait froid,]_1 [ferme la porte.]_2 `result*(1,2)`

Pour nos analyses, nous nous sommes focalisée, parmi les relations annotées dans le cadre du projet ANNODIS, sur ces quatre relations qui correspondent aux relations causales définies par Asher & Lascarides (2003) dans le cadre de la SDRT.

L'analyse de ces dernières nous a permis de faire différents constats. Tout d'abord, la très faible représentation dans le corpus, voire l'absence, des relations d'*Explication** et de *Résultat** dans ANNODIS (table 1) a retenu notre attention.

Nombre de relations annotées	Annot. A (44 textes)	Annot. B (43 textes)	Annot. 1 (28 textes)	Annot. 2 (27 textes)	Annot. 3 (26 textes)	Experts (86 textes)	Total
Explication	39	63	62	38	39	120	361
Résultat	48	97	58	45	28	162	438
*Explication**	7	6	8	0	0	0	21
*Résultat**	0	0	0	0	0	0	0
Total relations annotées	1390	1426	1060	1110	1114	3353	9453

TABLE 1 – Nombre de relations annotées dans la ressource ANNODIS par chaque annotateur

D'autre part, nous nous sommes rendu compte que les relations causales qui avaient été annotées faisaient l'objet d'un accord inter-annotateurs très faible (voir Atallah, 2014). Dans le but de comprendre ces désaccords, ainsi que la très faible représentation des relations d'*Explication** et de *Résultat**, nous avons fait le choix de nous confronter nous-même à la tâche d'annotation.

Cette expérience, et plus particulièrement l'analyse des quelques occurrences des relations étiquetées *Explication**, nous a rapidement amenée à nous interroger sur la pertinence de la gamme de relations causales envisagées en SDRT et reprises dans ANNODIS. Observons les segments suivants extraits du corpus :

[Arturo a de la chance,]_38 [il arrive en Chine]_39 [au moment de la fête de la nouvelle année.]_40

La relation `explanation*(38,[39-40])` a été annotée et pourtant la relation en jeu ne correspond pas à la relation d'*Explication** telle que définie par la SDRT et reprise dans ANNODIS. En effet, elle ne partage pas grand chose en commun avec celle qui s'établit dans l'exemple cité dans le manuel (*Ferme la porte. Il fait froid.*). Par ailleurs, nous avons jugé que cette relation ne pouvait pas non plus être considérée comme une simple relation d'*Explication* : *Arturo a de la chance* peut être perçu comme un fait subjectif, alors que les arguments des relations d'*Explication* correspondent à des descriptions objectives d'éventualités.

En nous penchant sur les données d'ANNODIS, nous avons relevé d'autres relations pouvant être rapprochées de celle en jeu dans l'exemple que nous avons rapporté. Nous en avons conclu que la gamme de relations envisagée pour l'annotation, et par là-même celle définie en SDRT, ne permettait pas de rendre compte de toutes les relations observables dans les textes. Nous avons alors cherché à réorganiser les relations causales annotées dans le corpus selon la nature du lien en jeu, jusqu'à parvenir à une classification plus pertinente pour rendre compte de la diversité des données. La mise au point d'une nouvelle typologie des relations causales a motivé la constitution du corpus EXPLICADIS annoté sur cette base.

3 Constitution d'une ressource spécifique pour l'étude des relations causales

Si nos premières analyses sur les relations causales ont été menées à partir des annotations réalisées dans le cadre du projet ANNODIS et notamment des situations de désaccords inter-annotateurs, il nous a rapidement semblé nécessaire de procéder à la constitution d'un nouveau corpus enrichi spécifiquement pour l'étude des relations causales. Ce corpus devait être annoté à l'aide d'un jeu d'étiquettes plus important que celui utilisé dans ANNODIS afin d'établir des distinctions plus fines entre les relations causales et de répondre ainsi à certaines difficultés rencontrées lors de la campagne d'annotation précédente. Ce raisonnement suit l'hypothèse posée et vérifiée par Prévot *et al.* (2009) selon laquelle l'introduction d'une gamme de relations plus riche permettrait de rendre l'annotation plus précise et donc moins confuse.

Nous avons ainsi complété la liste des relations retenues lors du projet ANNODIS à l'aide de nouveaux types de relations

causales. Nous avons défini cette nouvelle gamme de relations de discours causales dans (Atallah, 2014). On y retrouve les relations d'*Explication* et de *Résultat*, caractérisées comme des *relations inter-événementielles*, relations dont les effets sémantiques portent sur le contenu. Les relations d'*Explication** et de *Résultat** définies précédemment ont été renommées *relations pragmatiques* dans le but de les distinguer d'un nouveau type de relations causales : les *relations causales épistémiques*. Ces relations introduites plus tôt par Sweetser (1990) ont été caractérisées plus précisément dans le cadre de la SDRT (Atallah, 2014). Enfin, un type de relations causales épistémiques particulier a été distingué des autres sous la dénomination de *relations inférentielles*, suite au rapprochement effectué entre ces relations et celles étudiées par Bras *et al.* (2009).

Nous reprenons ci-dessous cette liste de relations causales, liste qui a servi de base pour l'annotation du corpus EXPLICADIS. Chacune de ces relations est associée à une étiquette et illustrée à l'aide d'un exemple extrait du corpus.

Explication (`explanation`)
- Définition : L'éventualité décrite dans le 2nd segment est la cause de l'éventualité décrite dans le 1er segment.
- Exemple : [L'armée est déçue,]_12 [il n'y a aucun viol, aucun pillage, aucun meurtre.]_13 `explanation(12,13)`

Résultat (`result`)
- Définition : L'éventualité décrite dans le 1er segment est la cause de l'éventualité décrite dans le 2nd segment.
- Exemple : [Arturo est un petit corbeau]_11 [qui s'ennuie.]_12 [Il décide d'aller visiter le monde.]_13 `result(12,13)`

Explication_épistémique ($\texttt{explanation}_{ep}$)
- Définition : Le locuteur rapporte ses croyances dans le 1er segment et justifie celles-ci dans le 2nd segment en exposant ses connaissances.
- Exemple : [Ce phénomène semble se confirmer à Mariana,]_37 [où on peut observer deux voies parallèles à la sortie sud de la ville.]_38 $\texttt{explanation}_{ep}\texttt{(37,38)}$

Résultat_épistémique ($\texttt{result}_{ep}$)
- Définition : Les connaissances exposées par le locuteur dans le 1er segment l'entraînent à croire certains faits rapportés dans le 2nd segment.
- Exemple : [Or la psychomécanique répond à ces deux types d'exigences.]_24 [Il serait donc intéressant de regarder si les outils théoriques qu'elle a développés permettent de rendre compte de certaines observations faites par la neuropsychologie.]_25 $\texttt{result}_{ep}\texttt{(24,25)}$

Explication_inférentielle ($\texttt{explanation}_{inf}$)
- Définition : Les deux segments reliés décrivent des connaissances du locuteur. Les faits connus décrits dans le 1er segment découlent logiquement de ceux décrits dans le 2nd segment.
- Exemple : [BITNET était différent d'Internet]_7 [parce que c'était un réseau point-à-point de type « stocké puis transmis ».]_8 $\texttt{explanation}_{inf}\texttt{(7,8)}$

Résultat_inférentiel ($\texttt{result}_{inf}$)
- Définition : Les deux segments reliés décrivent des connaissances du locuteur. Les faits connus décrits dans le 2nd segment découlent logiquement de ceux décrits dans le 1er segment.
- Exemple : [La première exposition avicole de Belfort date de 1922.]_4 [Cela fait donc plus de trois-quarts de siècle que la digne société du même nom encourage, dans la région, les éleveurs amateurs.]_5 $\texttt{result}_{inf}\texttt{(4,5)}$

Explication_pragmatique ($\texttt{explanation}_{prag}$)
- Définition : L'éventualité décrite dans le 2nd segment justifie l'acte de langage accompli lors de l'énonciation du 1er segment.
- Exemple : [Mais que ces derniers se rassurent,]_25 [il y aura encore deux autres tours]_26 [pour se rattraper.]_27 $\texttt{explanation}_{prag}\texttt{(25,[26,27])}$

Résultat_pragmatique ($\texttt{result}_{prag}$)
- Définition : L'éventualité décrite dans le 1er segment justifie l'acte de langage accompli lors de l'énonciation du 2nd segment.
- Exemple : [Suzanne Sequin n'est plus.]_1 [...] [Nos condoléances.]_35 $\texttt{result}_{prag}\texttt{(1,35)}$

Ce nouveau jeu de relations permet de rendre compte d'une dimension de la causalité non traitée dans ANNODIS. En effet, celui-ci intègre, en plus de la causalité inter-événementielle, des relations qui relèvent de l'argumentation.

En plus des relations que nous venons de présenter, nous avons annoté certains indices linguistiques qui nous semblaient pertinents pour l'étude des relations de discours causales. Ces indices sont de deux types.

Le premier type d'indices correspond à des indices que nous avons associés à l'expression de la causalité. Ces indices (ou faisceaux d'indices) pouvaient correspondre à des connecteurs (*car*, *parce que*, *donc*, *alors*...), mais aussi à des structures

syntaxiques particulières (apposition, participe présent, participe passé...) ou à des marques typographiques (deux points, guillemets, parenthèses...).

Le second type d'indices annoté concerne plus spécifiquement les relations causales épistémiques. Nous avons remarqué que ces relations s'accompagnaient souvent de la présence d'éléments exprimant la modalité. Nous avons donc repéré ces indices lorsque ceux-ci étaient présents. Parmi ceux-ci, on trouve, entre autres, des adverbes, comme *probablement*, des verbes modaux, comme *pouvoir*, mais aussi des terminaisons de conditionnel.

4 Présentation de la ressource EXPLICADIS et exploitations

Sur la base des éléments que nous venons de définir, nous avons pu procéder nous-même à l'annotation du corpus EXPLICADIS. Ce corpus a été constitué en trois grandes étapes : une phase exploratoire, suivie de deux phases successives de constitution, puis d'élargissement du corpus.

La première étape a permis de mettre au point la typologie de relations causales présentée plus haut. Pour cela, nous avons ré-annoté l'ensemble des textes annotés lors de la campagne naïve d'annotation d'ANNODIS (42 textes) en nous concentrant sur les relations causales. Dans un premier temps, nous avons cherché à repérer toutes les relations causales présentes en nous appuyant sur les textes segmentés issus d'ANNODIS. Puis, ce n'est que dans un second temps que nous avons pris connaissance des annotations réalisées dans le cadre du projet précédent. Cette démarche avait pour but d'éviter que nos annotations soient trop influencées par celles qui étaient déjà disponibles. Elle nous a permis d'ajouter des relations causales qui n'avaient été repérées par aucun annotateur, mais surtout de nous rendre compte des difficultés posées par l'annotation, étant donnée la gamme restreinte de relations causales considérée dans ANNODIS. Au cours de cette phase exploratoire, nous avons ainsi pu affiner la liste des relations causales nécessaires pour résoudre au mieux les désaccords entre les annotateurs d'ANNODIS.

Une fois les objets à annoter bien définis, nous avons pu procéder à la ré-annotation de l'ensemble des textes annotés lors de la campagne naïve (42 textes précédents) mais aussi exploratoire (44 textes supplémentaires) d'ANNODIS. Ce premier élargissement du corpus nous a permis d'obtenir un corpus plus grand que nous avons nommé "Corpus_86".

Pour l'ensemble des 86 textes, nous avons confronté nos annotations avec les annotations antérieures. Ainsi, à chaque fois qu'au moins un annotateur avait identifié une relation causale entre deux segments, nous avons proposé notre propre annotation, que la relation en jeu soit causale ou non. Nous avons ainsi ré-annoté 533 relations [4] sur l'ensemble du Corpus_86. Le fait de devoir proposer une relation entre des arguments nous obligeait à réfléchir aux motivations qui nous poussaient à retenir ou non l'annotation d'une relation causale.

Pour qu'un corpus soit le plus représentatif possible, il faut veiller à ce qu'il associe deux caractéristiques (Habert, 2000) : il doit être de taille suffisante et il doit pouvoir rendre compte de la diversité des usages langagiers. Afin de répondre à la première exigence, nous avons, comme indiqué précédemment, élargi notre tout premier corpus de 42 à 86 textes, obtenant ainsi un corpus dont la taille peut être jugée satisfaisante pour mener des analyses quantitatives (27 547 mots).

En ce qui concerne la seconde caractéristique, le Corpus_86 présentait certaines limites. En effet, celui-ci est essentiellement constitué d'extraits de textes issus de brèves de presse (textes à dominante narrative issus de *Est-Républicain* : NEWS) et d'articles encyclopédiques (textes à dominante expositive issus de *Wikipédia* : WIK). Seuls cinq textes à dominante argumentative ont été annotés : deux textes issus d'articles scientifiques de linguistique (LING) et trois de rapports scientifiques concernant la géopolitique (GEOP). Nous avons donc envisagé d'intégrer à notre corpus de nouveaux textes argumentatifs afin d'obtenir une meilleure représentativité des genres textuels au sein de notre corpus. Par ailleurs, cette intégration se voulait pertinente au vu de notre objet d'étude – la *causalité* – et des liens étroits que celui-ci entretient avec l'argumentation.

Notre corpus d'étude, dans sa version finale, comprend, en plus des 86 textes initiaux, 31 extraits de textes supplémentaires. Nous avons sélectionné ces textes parmi ceux qui ont été exploités lors du projet ANNODIS, dans le cadre d'une autre approche. Ceux-ci n'ayant pas été traités par l'approche *ascendante* du projet, nous avons dû procéder à leur segmentation en unités de discours élémentaires avant de les annoter en relations causales.

La table 2 présente l'ensemble de notre corpus d'étude. Les 31 textes supplémentaires y sont représentés sous l'étiquette de "Corpus_31". Nous avons souhaité faire en sorte que les textes issus des sous-corpus LING et GEOP constituent un ensemble de textes argumentatifs comparable, en termes quantitatifs (*cf.* nombre de mots), à l'ensemble des textes narratifs

4. Ces chiffres ne tiennent pas compte par ailleurs des relations causales que nous avons ajoutées et qui n'avaient été repérées par aucun annotateur.

	Corpus_86		Corpus_31		Total		
Sous-corpus	textes	mots	textes	mots	textes	mots	
NEWS	39	9 768	3	846	42	10 614	
WIK	42	15 983	0	0	42	15 983	
LING	2	586	19	6 691	21	7 277	12 506
GEOP	3	1 210	9	4 019	12	5 229	
Total	86	27 547	31	11 556	117	39 103	

TABLE 2 – Répartition des textes au sein d'EXPLICADIS en fonction des sources dont ils sont issus

issus de NEWS, ainsi qu'à l'ensemble des textes expositifs issus de WIK.

Sur l'ensemble de ces textes, 319 relations causales ont été repérées et annotées à l'aide des huit étiquettes présentées précédemment. La table 3 montre que, même si les relations portant sur le contenu propositionnel (*Explication* et *Résultat*) sont majoritaires dans le corpus, celles-ci ne représentent qu'un peu plus de la moitié des relations causales que nous avons relevées (environ 53 %). La présence des autres relations causales est loin d'être négligeable. Sachant que les relations pragmatiques représentent moins de 1 % de l'ensemble des relations annotées [5], la nécessité d'intégrer les relations causales épistémiques et inférentielles dans le cadre de la SDRT est confirmée par la réalité des données.

Nombre de relations annotées	Corpus_86	Corpus_31	Total
Explication	77	27	104
Résultat	55	10	65
Explication_épistémique	35	33	68
Résultat_épistémique	9	15	24
Explication_inférentielle	8	4	12
Résultat_inférentiel	26	17	43
Explication_pragmatique	1	1	2
Résultat_pragmatique	1	0	1
Total relations causales annotées	212	107	319

TABLE 3 – Nombre de relations causales annotées dans la ressource EXPLICADIS

Si la constitution de la ressource EXPLICADIS nous a permis d'aboutir à une meilleure caractérisation des relations de discours causales, elle nous a également autorisée à mener, par la suite, des analyses diversifiées. Nous avons pu notamment nous intéresser à la variation relative au genre textuel et mettre en évidence certaines corrélations entre différents paramètres : type de relation causale, choix rhétorique, genre textuel (voir Atallah, 2014).

Cette ressource devrait, par ailleurs, permettre à d'autres utilisateurs de l'exploiter pour leurs propres besoins. Par sa taille, la diversité des textes qui y sont représentés et les annotations proposées, ce corpus devrait constituer une base pertinente et originale pour l'étude des relations causales. En ce qui concerne la diversité des textes, nous notons qu'une telle ressource n'a, à notre connaissance, jamais été conçue. En effet, les projets antérieurs d'annotation discursive ont fait le choix de rester sur la construction de corpus homogènes, et plus spécifiquement de corpus constitués exclusivement de textes journalistiques (textes à dominante narrative) [6].

Par ailleurs, la gamme de relations retenues pour l'annotation constitue un véritable atout pour mener des études descriptives. D'une part, les annotations proposées tiennent compte de distinctions assez fines et, d'autre part, il s'agit du premier corpus annoté pour le français qui s'appuie sur une vision intégrative de la causalité, considérant non seulement sa dimension événementielle, dimension habituellement traitée, mais également sa dimension argumentative.

Si ce corpus présente un format idéal pour adopter une approche onomasiologique face aux données, c'est-à-dire une approche qui part de la relation elle-même et non de ses marqueurs potentiels, il permettra également, grâce à une projection des indices annotés d'envisager des études selon une perspective sémasiologique (qui part des indices).

Afin de permettre à d'autres utilisateurs de l'exploiter pour leurs propres besoins, la ressource EXPLICADIS sera disponible en ligne, aux côtés du corpus ANNODIS. Tout comme pour ce dernier, les annotations retenues ne peuvent être considérées

5. Cette très faible proportion s'explique par les types de textes retenus dans le corpus. Un corpus rendant compte de la langue parlée, ou impliquant plus généralement des interactions, serait plus approprié pour l'étude de ce dernier type de relations.

6. Par exemple, les textes proposés par le RST TreeBank et par le PDTB sont extraits du *Wall Street Journal*. Quant au French Discourse Tree Bank (FDTB, Danlos *et al.*, 2012), il est prévu que ce corpus soit constitué de textes tirés du journal *Le Monde*.

comme des informations à valeur certaine et stabilisée, elles resteront un reflet de notre propre point de vue sur la causalité. En cela, EXPLICADIS ne se veut pas *corpus de référence* pour l'étude de la causalité, mais bien *corpus exploratoire*. Il pourra ainsi servir de point de départ pour des analyses ultérieures sur la causalité, non comme un objet figé, mais comme un objet que nous invitons à faire évoluer.

Il faut par ailleurs noter que ce corpus n'a fait l'objet que d'une simple annotation. Il serait intéressant de confronter nos annotations à celles d'autres annotateurs. Cela permettrait de tester d'une part la pertinence du nouveau jeu de relations causales que nous avons défini et d'autre part l'hypothèse selon laquelle un jeu de relations plus précis mène à une annotation moins confuse (Prévot *et al.*, 2009). Plus généralement, l'analyse des accords entre annotateurs permettrait de valider notre contribution à l'étude des relations causales, et l'analyse des désaccords de mettre en évidence les améliorations encore nécessaires. Tout comme pour le corpus ANNODIS, les désaccords inter-annotateurs qui pourront être relevés sur le corpus EXPLICADIS ne devront pas être traités comme des erreurs, mais, au contraire, ils devront être considérés et étudiés avec la plus grande attention. Si Habert (2004) recommande au linguiste d'accepter de travailler avec des données imparfaites, nous pensons que le linguiste doit aussi apprendre à tirer de ces imperfections – ici, des désaccords inter-annotateurs – de nouvelles informations qui pourront servir ses recherches.

Références

AFANTENOS S. D., ASHER N., BENAMARA F., BRAS M., FABRE C., HO-DAC L.-M., LE DRAOULEC A., MULLER P., PÉRY-WOODLEY M.-P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ-COURET M. & VIEU L. (2012). An empirical resource for discovering cognitive principles of discourse organization : the ANNODIS corpus. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, p. 2727–2734, Istanbul, Turkey.

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

ATALLAH C. (2014). *Analyse de relations de discours causales en corpus : étude empirique et caractérisation théorique*. Thèse de Doctorat, Université de Toulouse, Toulouse.

BRAS M., LE DRAOULEC A. & ASHER N. (2009). A Formal Analysis of the French Temporal Connective *alors*. In BEHRENS & C. FABRICIUS-HANSEN, Eds., *Structuring information in discourse : the explicit/implicit dimension.*, volume 1, p. 149–170. Oslo Studies in Language.

CARLSON L., MARCU D. & OKUROWSKI M. E. (2001). Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, volume 16, p. 1–10, Stroudsburg, PA, USA : Association for Computational Linguistics.

DANLOS L., ANTOLINOS-BASSO D., BRAUD C. & ROZE C. (2012). Vers le FDTB : French Discourse Tree Bank. In *TALN 2012 : 19ème conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, p. 471–478, Grenoble, France.

HABERT B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? In M. BILGER, Ed., *Linguistique sur corpus. Etudes et réflexions*, volume 31 of *Cahiers de l'université de Perpignan*, p. 11–58. Perpignan : Presses Universitaures de Perpignan.

HABERT B. (2004). Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs. *Revue française de linguistique appliquée*, **IX**(1), 5–24.

MILTSAKAKI E., PRASAD R., JOSHI A. & WEBBER B. (2004). The Penn Discourse Treebank. In *In Proceedings of LREC 2004*, Lisbon, Portugal.

MULLER P., VERGEZ-COURET M., PRÉVOT L., ASHER N., BENAMARA F., BRAS M., LE DRAOULEC A. & VIEU L. (2012). Manuel d'annotation en relations de discours du projet ANNODIS. *Carnets de grammaire*, **21**.

PRÉVOT L., VIEU L. & ASHER N. (2009). Une formalisation plus précise pour une annotation moins confuse : la relation d'élaboration d'entité. *Journal of French Language Studies*, **19**(2), 207–228.

PÉRY-WOODLEY M.-P., AFANTENOS S. D., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *TAL*, **52**(3), 71–101.

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 *Actes de TALN 2009*, Senlis, France.

SWEETSER E. E. (1990). *From Etymology to Pragmatics : Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge : Cambridge University Press.