Exploitation d'un corpus annoté pour l'analyse des relations causales

Caroline Atallah

Exploitation d'un corpus annoté pour l'analyse des relations causales

Caroline Atallah

CLLE-ERSS, CNRS & Université de Toulouse, 31058 Toulouse Cedex 9

caroline.atallah@univ-tlse2.fr

RÉSUMÉ

Notre étude vise à proposer, à partir de l'observation d'énoncés attestés, une description des relations causales dans le cadre d'une théorie représentationnelle du discours, la SDRT. L'exploitation d'un corpus de textes enrichis d'annotations discursives nous a permis de confronter la théorie à la réalité des données. Constatant que la SDRT ne rendait pas compte de la diversité des relations causales présentes dans les textes, nous proposons d'enrichir le modèle à partir de nos observations. Pour la suite de nos analyses, nous envisageons d'élargir notre corpus de façon à le rendre plus représentatif.

ABSTRACT

Exploring an annotated corpus for the analysis of causal relations

This study aims at offering, from the observation of attested data, a description of causal relations in the framework of a representational theory of discourse called SDRT. Exploring a corpus of texts, which are annotated at discourse-level, allowed us to confront the theory with real data. Upon realising that SDRT did not reflect the diversity of causal relations that could be observed in texts, we offer to enrich the theoretical model on the basis of our observations. For further analysis, we are planning on expanding our corpus in order to make it more representative.

MOTS-CLES : *discours, corpus annoté, relations causales, SDRT, genre textuel.*
KEYWORDS : *discourse, annotated corpus, causal relations, SDRT, textual genre.*

1 Introduction

La causalité a fait l'objet de nombreuses études, et ce, depuis l'Antiquité et les travaux sur la rhétorique d'Aristote. A travers les siècles, la notion de *cause* est restée au cœur des préoccupations des philosophes (Hume, 1748, Russel, 1912, Lewis, 1973, Kistler, 2004) et il est encore aujourd'hui difficile de s'accorder sur la définition qui peut lui être attribuée : « nobody has provided a general definition of CAUSE, though causality has been the topic of centuries of debate. » (Hovy et Maier, 1993).

L'ambition de notre étude n'est pas de proposer une définition de la notion-même de *cause*. Nous nous intéressons tout d'abord à la langue et à ses réalisations, avec l'objectif de parvenir à une description des relations dans lesquelles la cause intervient.

En linguistique, la causalité est étudiée à des niveaux différents. Certains auteurs s'attachent à proposer une description syntaxico-sémantique de la causalité, envisageant celle-ci comme une relation liant une cause à son effet (Nazarenko, 2000, Gross, 2009). D'autres, en s'intéressant à l'argumentation, rendent compte d'un autre niveau dans lequel la causalité peut s'établir (Ducrot et Anscombre, 1983, Plantin, 1990).

Par ailleurs, la plupart des travaux portant sur la causalité se concentrent principalement sur l'étude des connecteurs permettant d'inférer un lien causal : *à cause de*, *alors*, *de ce fait*, *du coup*, *donc*, *etc.* (Jayez et Rossari, 2001), ou encore des verbes causaux tels que *provoquer*, *causer*, *occasionner* (Gross, 2009). Ces travaux s'appuient souvent sur des exemples construits ou utilisent des exemples attestés à des fins d'illustration.

Notre approche de la causalité se distingue de celles que nous venons de décrire. L'originalité

réside dans le fait que nous nous appuyons sur un corpus annoté spécifiquement pour l'étude des relations causales.

Nous considérons ce corpus comme le point de départ de notre étude. Selon la distinction établie par (Tognini-Bonelli, 2001), nous préférons à une approche « corpus-based », une approche « corpus-driven »[1].

D'autre part, la présence des annotations nous permet d'aborder la causalité sous un autre angle : c'est la relation de causalité elle-même et non ses marqueurs potentiels qui constitue le point de départ de nos analyses.

A partir de l'observation d'énoncés attestés, notre étude, qui s'inscrit au sein du projet EXPLICADIS[2], vise à proposer une description des relations causales dans le cadre d'une théorie du discours particulière : la SDRT. Nous souhaitons enrichir la théorie afin que celle-ci puisse rendre compte de la réalité des données.

Cet article s'articule en quatre sections. Nous présenterons tout d'abord le modèle théorique de la SDRT et le traitement qu'il propose pour les relations de discours causales. Nous décrirons ensuite notre corpus d'étude, corpus que nous avons constitué à partir d'un corpus annoté déjà existant. Puis, nous exposerons comment l'exploitation de ces données nous a permis d'envisager un enrichissement de la théorie. Enfin, nous ouvrirons une discussion sur la représentativité du corpus dans le but de définir des critères en vue d'un élargissement de notre corpus pour des analyses ultérieures.

2 Le modèle théorique de la SDRT et son traitement des relations causales

La SDRT, *Segmented Discourse Representation Theory* (Lascarides et Asher, 1993, Asher et Lascarides, 2003), est une théorie représentationnelle du discours, développée dans la continuité de la DRT, *Discourse Representation Theory*, de (Kamp et Reyle, 1993) et des théories sur la cohérence du discours (Hobbs, 1985, Mann et Thompson, 1988).

Nous présenterons brièvement cette théorie, puis nous nous attacherons aux traitements qu'elle propose pour les relations causales.

2.1 Présentation de la SDRT

Afin de rendre compte de la cohérence du discours, la SDRT propose de représenter le discours comme un ensemble de segments du discours liés entre eux par des relations de discours.

Cette représentation se construit selon une démarche ascendante. Il s'agit de :

- segmenter le discours en unités minimales ;
- définir les relations liant les segments entre eux, les segments reliés forment alors de nouvelles unités, dites *complexes* ;
- définir les relations liant les segments complexes à d'autres segments.

Selon la représentation proposée par la SDRT, chaque constituant, noté K_{π}, est désigné par une étiquette, notée π. Les constituants correspondent aux représentations du contenu propositionnel des segments, alors que les étiquettes correspondent aux actes de langage.

[1] Ce type d'approche consiste à s'appuyer sur l'observation de données attestées pour parvenir à une caractérisation théorique du phénomène étudié.

[2] Le projet EXPLICADIS, EXPLICation et Argumentation en DIScours, co-financé par le PRES toulousain et la région Midi-Pyrénées (2010-2013) implique les deux laboratoires toulousains CLLE-ERSS et IRIT.

Les relations s'établissent entre les étiquettes des constituants. On notera *R(α,β)* une relation liant les étiquettes α et β. La SDRT ne propose pas de liste finie de relations. Cependant, elle propose des outils permettant de caractériser chaque relation à l'aide de règles formulées dans le langage de la Glue Logic.

Parmi les relations envisagées par la SDRT, nous trouvons deux types de relations causales : la relation d'Explication et la relation de Résultat. Nous allons nous pencher à présent sur les caractéristiques propres à ces relations, telles que décrites dans (Asher et Lascarides, 2003).

2.2 Les relations causales en SDRT

Selon les règles énoncées en SDRT, les relations causales Explication et Résultat ont pour effet sémantique de lier les éventualités (événements ou états) de deux segments, notées e_α et e_β, par un lien causal :

Explication_Conséquence $\Phi_{\text{Explication}(\alpha,\beta)} \Rightarrow \text{cause}\ (e_\beta, e_\alpha)$

Résultat_Conséquence $\Phi_{\text{Résultat}(\alpha,\beta)} \Rightarrow \text{cause}\ (e_\alpha, e_\beta)$

Les effets sémantiques de ces deux relations se différencient au niveau de l'ordre des arguments, la relation d'Explication (1) présente l'effet en premier alors que la relation de Résultat (2) présente une cause puis son effet :

(1) Max est tombé. John l'a poussé.[3]

(2) John a poussé Max. Il est tombé.

La SDRT distingue un niveau supplémentaire de relations. Ces relations, notées *Explication** et *Résultat**, se réalisent à niveau illocutoire. Il s'agit de relations d'ordre pragmatique. La relation d'Explication* (3) lie un premier segment correspondant à un acte de langage à un second segment contenant la justification de l'énonciation de cet acte. La relation de Résultat* (4) présente ces segments dans un ordre contraire :

(3) Ferme la fenêtre. J'ai froid.

(4) J'ai froid. Ferme la fenêtre.

3 Une approche empirique de la causalité

La SDRT ne proposant à l'heure actuelle qu'une description succincte des relations causales, nous avons décidé de nous confronter à la réalité des données. Nous considérons le corpus comme point de départ pour nos analyses et procédons en deux temps : 1. observation des données, 2. enrichissement de la théorie à partir des observations.

La ressource ANNODIS[4] a été construite dans le but de permettre ce type d'exploitation. Nous décrirons dans cette section la méthodologie suivie pour la construction du corpus annoté issu du projet ANNODIS ainsi que le traitement qui a été fait des relations causales. Nous expliquerons ensuite comment nous avons exploité ces données pour procéder à l'analyse des relations de discours causales.

[3] Les exemples (1) à (4) sont empruntés à (Asher et Lascarides, 2003).

[4] Le projet ANNODIS (ANNOtation DIScursive de corpus), financé par l'ANR (2007-2010), a réuni des chercheurs des laboratoires CLLE-ERSS (Toulouse), IRIT (Toulouse) et GREYC (Caen).

3.1 ANNODIS, un corpus annoté au niveau discursif

Le projet ANNODIS (Péry-Woodley et *al.*, 2009, Péry-Woodley et *al.*, 2012) a donné naissance au premier corpus de textes en français enrichis d'annotations discursives. Ce corpus a été constitué selon deux approches.

La première approche, dite *approche macro*, s'est intéressée aux structures discursives de haut-niveau en proposant une annotation des marqueurs textuels relatifs à ces structures.

La deuxième approche, dite *approche ascendante*, s'est attachée à construire une représentation de la structure du discours en liant les unités discursives entre elles par des relations rhétoriques.

Dans le cadre de notre étude, nous nous concentrerons sur la seconde approche qui a abouti à l'élaboration d'un corpus enrichi avec des relations discursives.

3.1.1 Constitution du corpus selon une approche ascendante

L'élaboration du corpus annoté a été réalisée en deux temps. Des textes ont d'abord été segmentés en Unités de Discours Elémentaires. La segmentation de ces textes a fait état d'un accord entre les annotateurs. Les segments constitués ont ensuite été liés entre eux par des relations de discours. Lorsque cela était pertinent, de nouvelles relations ont été annotées liant des segments complexes à d'autres segments. L'ensemble des textes segmentés et les relations associées constituent le corpus annoté.

La segmentation ainsi que l'annotation des relations discursives ont été réalisées selon les recommandations fournies par un manuel rédigé spécifiquement pour le projet. Avant de trouver sa forme définitive, ce manuel a été testé et modifié lors d'une première phase d'annotation dite *exploratoire*.

Le guide finalisé, la campagne d'annotation a pu débuter sur de nouveaux textes. Trois annotateurs *naïfs*[5] ont procédé à une double annotation de chaque texte après s'être mis d'accord sur la segmentation. Ces mêmes textes ont ensuite été annotés par des *experts*, de façon à obtenir un ensemble de textes triplement annotés.

Le corpus ainsi enrichi est maintenant disponible pour être exploité[6].

3.1.2 Les relations causales dans le corpus ANNODIS

Le manuel d'annotation propose une liste de relations. Parmi celles-ci, nous retrouvons les deux relations qui nous intéressent : Explication et Résultat.

Afin de guider les annotateurs, chaque relation est définie puis illustrée par des exemples, une liste de marqueurs potentiels est aussi donnée.

Les relations d'Explication et de Résultat, codées respectivement *explanation* et *result*, ont été caractérisées comme suit :

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. »

[5] Il s'agissait d'étudiants en Licence ne possédant pas de connaissances particulières sur les théories du discours.

[6] La ressource ANNODIS est disponible sur le site REDAC (Ressources Développées à CLLE-ERSS : http://redac.univ-tlse2.fr/), sous licence « Creative Commons ».

- Exemple : [Le chômage baisse en 2008]_1 [parce qu'il y a moins d'actifs.]_2
  Explanation (1,2)
- Marqueurs suggérés : *car*, *parce que*, *à cause de*, *du fait de*, *par la faute de*, *grâce à*, *si* 1 *c'est parce que* 2, *depuis* (si causalité évidente).

Résultat (*result*)

- Définition : « La relation Résultat caractérise des liens entre deux segments portant sur deux éventualités (événements ou états) dont la 2e 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 suggérés : *du coup*, *donc*, *par conséquent*, *en conséquence*, *par suite*, *à la suite de quoi*.

3.2 Constitution d'un corpus annoté pour l'étude des relations causales

Le corpus issu d'ANNODIS constitue le point de départ de notre étude. Nous avons choisi de nous intéresser aux textes annotés après finalisation du manuel d'annotation. Nous nous sommes donc concentrée sur 42 textes qui ont fait l'objet de plusieurs annotations : au moins une annotation *naïve* et une annotation *experte*. Ces textes sont des extraits d'articles issus de l'encyclopédie en ligne *Wikipédia* (27 textes) et du quotidien *Est Républicain* (15 textes). Les relations causales représentent 9% des relations annotées dans ces textes (dont 4,3% d'Explication et 4,7% de Résultat).

Afin de faciliter l'exploitation du corpus pour l'étude des relations causales, nous avons procédé à une ré-annotation de l'ensemble des textes. Pour cela, nous nous sommes appuyée sur les annotations disponibles, confrontant les propositions des différents annotateurs et décidant pour chaque relation d'Explication et de Résultat repérée si nous maintenions l'annotation. Notre corpus ré-annoté compte 61 relations d'Explication et 57 relations de Résultat.

Constatant que les constituants liés pouvaient être de nature différente, nous avons construit une typologie des relations causales. Chaque exemple issu de notre corpus ré-annoté a ensuite fait l'objet d'une classification selon des critères que nous décrirons dans la section 4. Les 42 textes segmentés enrichis de nos propres annotations constituent notre corpus d'étude, corpus adapté spécifiquement pour l'étude des relations causales.

4 Enrichissement de la théorie à partir de l'observation des données

Alors que la SDRT s'est principalement concentrée sur la description des relations causales portant sur les éventualités contenues dans les segments liés, le corpus fait apparaître d'autres types de relations. Nous décrirons ici les différentes relations causales que nous avons distinguées tout en illustrant nos propos par des exemples tirés de notre corpus. Ces relations se distinguent par leurs effets sémantiques. Nous proposons d'enrichir la théorie en caractérisant chaque type de relation à travers la nature du lien causal établi.

4.1 Relations entre éventualités

Selon la SDRT, les relations Explication et Résultat ont pour effet sémantique d'établir un lien causal entre deux éventualités. Nous retrouvons ce type de relation dans le corpus :

(5) [La tour 7 du WTC s'est effondrée dans l'après-midi]_11 [en raison d'incendies et des dégâts occasionnés par la chute des Twin Towers.]_12
Relation annotée : Explication (11,12)

(6) [le côté gauche de la voiture [qui doublait]_9 a mordu l'accotement.]_8 [L'automobile

a perdu sa roue gauche]_10
Relation annotée : Résultat (8,10)

(7) [son chef, [Gérard Pizzetti,]_8 [en désaccord avec le fonctionnement de l'association,]_10 démissionnait.]_9
Relation annotée : Explication (9,10)

(8) [Cette loi du silence règne]_8 [car elle joue sur la peur]_9 [que les mafieux ont de la mafia,]_10 [car ils connaissent les représailles]_11 [qui attendent celui]_12 [qui parlerait.]_13
Relation annotée : Explication (10,11)

Nous pouvons distinguer quatre sous-types de relations selon la nature de l'éventualité (événement ou état) rapportée par chacun des deux segments :

1. éventualité 1 : événement ; éventualité 2 : événement (exemples (5) et (6)) ;
2. éventualité 1 : événement ; éventualité 2 : état (exemple (7)) ;
3. éventualité 1 : état ; éventualité 2 : événement[7] ;
4. éventualité 1 : état ; éventualité 2 : état (exemple (8)).

Lorsque la relation se présente sous la configuration 1., elle est soumise à des contraintes temporelles. En effet, l'événement expliquant précède toujours l'événement expliqué. (Asher et Lascarides, 2003) énoncent les règles suivantes :

Explication_Contraintes_Temporelles $\Phi_{\text{Explication}(\alpha,\beta)} \Rightarrow (\text{event}\ (e_\beta) \Rightarrow (e_\beta < e_\alpha))$

Résultat_Contraintes_Temporelles $\Phi_{\text{Résultat}(\alpha,\beta)} \Rightarrow (\text{event}(e_\alpha) \Rightarrow (e_\alpha < e_\beta))$

Les exemples (5) et (6) vérifient ces contraintes.

Les relations d'Explication et de Résultat, telles que décrites par la SDRT, ne représentent en réalité qu'une moitié des relations causales observables dans le corpus. Nous proposons donc d'enrichir la théorie en introduisant deux types de relations causales supplémentaires : les relations épistémiques et les relations inférentielles.

4.2 Relations épistémiques

Observons l'exemple suivant tiré de notre corpus :

(9) [En ce qui concerne les programmes spatiaux hors MD,]_2 [il est difficile de faire le point des financements proposés à l'heure actuelle,]_3 [car les lignes budgétaires restent éparpillées et le plus souvent non-identifiables dans le projet de budget de la Maison Blanche.]_4
Relation annotée : Explication (3,4)

Nous pouvons voir ici une simple relation entre deux états : la difficulté de faire le point s'explique par l'éparpillement des lignes budgétaires. La relation répondrait alors aux effets sémantiques décrits précédemment :

Explication_Conséquence $\Phi_{\text{Explication}(\alpha,\beta)} \Rightarrow \text{cause}\ (e_\beta, e_\alpha)$

[7] Les relations causales présentant cette configuration sont absentes de notre corpus. Nous observons que lorsque le segment « expliquant » contient un état, celui-ci se trouve au sein du second argument d'une relation d'Explication. De même, lorsque l'état se trouve dans le segment « expliqué », ce dernier prend place en tant que second argument d'une relation de Résultat. Une relation établissant un lien causal entre un événement et un état se présenterait donc plus fréquemment sous la configuration décrite en 2. Nous considérons bien entendu ces observations avec la plus grande précaution, notre corpus nécessitant d'être élargi pour pouvoir vérifier la validité de celles-ci (voir section 5).

Pour que cette règle soit respectée, il faut que la proposition décrite par le segment 3 soit vraie. Autrement dit, la difficulté évoquée doit être avérée et la valeur de vérité de la proposition ne peut être contestée.

Or, nous pouvons interpréter l'énoncé autrement et comprendre que la proposition décrite par le segment 3 n'est vraie que pour le locuteur. Une seconde interprétation peut en effet être envisagée : à travers l'emploi de l'évaluatif *il est difficile de*, il est possible de percevoir la présence de l'énonciateur. Celui-ci exprimerait son point de vue. Les effets sémantiques de la relation ne sont alors plus les mêmes, le lien causal ne s'établit pas entre deux états. Il ne s'agit plus d'une simple Explication, mais de la justification d'une croyance propre à l'énonciateur : celui-ci n'explique pas pourquoi il existe une difficulté, mais pourquoi il pense que cette difficulté existe. Nous dirons qu'il s'agit d'une *relation causale épistémique* puisqu'elle renvoie à une attitude mentale.

Dans notre corpus, nous avons relevé plusieurs relations de ce type. Bien que se présentant plus fréquemment sous l'ordre d'une relation d'Explication, nous avons relevé plusieurs relations de Résultat pouvant recevoir une interprétation épistémique. En voici un exemple :

(10) [Ces attentats ont été vécus presque en temps réel par des centaines de millions de téléspectateurs à travers le monde,]_25 [les images de l'avion heurtant la deuxième tour du World Trade Center ayant été diffusées en direct,]_26 [ainsi que l'effondrement complet en quelques secondes des trois tours du WTC à Manhattan.]_27 [Le choc psychologique a été considérable au plan international.]_28
Relation annotée : Résultat (25,28)

Tout comme pour l'exemple (9), si nous considérons l'énoncé comme un simple récit objectif, la relation sera alors une relation de Résultat liant deux éventualités entre elles. Or, l'adjectif *considérable*, en tant qu'évaluatif, peut être perçu comme exprimant un point de vue subjectif. Le contenu du segment 28 serait vrai pour le locuteur. Selon cette interprétation, la relation serait de type épistémique.

Les exemples (9) et (10) peuvent donc recevoir deux interprétations différentes selon qu'on attribue la vérité du contenu propositionnel d'un segment (premier segment pour une relation d'Explication, second segment pour une relation de Résultat) au locuteur seul ou qu'on le considère vrai pour tous.

Dans certains énoncés, l'ambiguïté est levée par la présence de marques explicites de subjectivité. C'est le cas notamment avec l'emploi de modaux tel que *probablement* :

(11) [« La route moderne [(entre Mariana et Aleria),]_64 [au bas des collines,]_65 est probablement un tracé traditionnel,]_63 [car elle suit tout naturellement la limite du terrain ferme et du terrain alluvial]_66 [et l'Itinéraire a pu choisir ce parcours]_67 ... »
Explication (63,[66,67])

L'emploi de *probablement*, appuyé par la présence des guillemets, indique que le segment 63 rapporte le point de vue du locuteur, celui-ci fait part de sa propre interprétation sur les origines de la route dont il est question. La relation est donc, sans ambigüité, une relation épistémique.

Nous empruntons, pour les relations que nous venons de décrire, l'appellation *épistémique* à (Sweetser, 1990). Celle-ci distingue trois domaines dans lesquels un lien causal peut s'établir : le domaine du contenu propositionnel, le domaine épistémique et le domaine illocutoire. Les relations s'établissant au niveau du contenu propositionnel correspondent aux relations Explication et Résultat décrites par la SDRT, alors que celles s'établissant au niveau illocutoire correspondent aux relations Explication* et Résultat*.

Sanders et *al.* (1992) ont fait le choix de ne pas établir de distinction entre les relations s'établissant au niveau épistémique et celles s'établissant au niveau illocutoire. Ils regroupent ces relations sous un seul type : *les relations pragmatiques*. Or, si nous nous intéressons aux

effets sémantiques des relations, nous constatons qu'ils sont différents.

Une relation épistémique établit un lien causal entre un acte de penser et une éventualité. Cette éventualité décrite dans un segment explique pourquoi le locuteur pense le contenu exprimé dans le segment complémentaire. Nous proposons de rendre compte ci-dessous des effets sémantiques de ce type de relation :

Explication_Epistémique_Conséquence $\Phi_{\text{Explication\_épistémique}(\alpha,\beta)} \Rightarrow \text{cause}\ (e_{\beta},e_{\alpha'})$ avec $[e_{\alpha'}$: penser $(\text{loc}\_\alpha,K_{\alpha})]$[8]

Résultat_Epistémique_Conséquence $\Phi_{\text{Résultat\_épistémique}(\alpha,\beta)} \Rightarrow \text{cause}\ (e_{\alpha},e_{\beta'})$ avec $[e_{\beta'}$: penser $(\text{loc}\_\beta,K_{\beta})]$

La relation pragmatique (Explication* et Résultat* en SDRT), que nous distinguons de la relation épistémique, établit, elle, un lien entre un acte de langage et une éventualité. Le locuteur explique pourquoi il accomplit cet acte :

Explication*_Pragmatique_Conséquence $\Phi_{\text{Explication*\_pragmatique}\ (\alpha,\beta)} \Rightarrow \text{cause}\ (e_{\beta},\alpha)$

Résultat*_Pragmatique_Conséquence $\Phi_{\text{Résultat*\_pragmatique}\ (\alpha,\beta)} \Rightarrow \text{cause}\ (e_{\alpha},\beta)$

Tout comme les relations décrites en 4.1, les relations épistémiques et les relations pragmatiques ont pour effet sémantique d'établir un lien entre deux éventualités. Cependant, une seule des éventualités liées correspond à une éventualité décrite dans le contenu propositionnel d'un segment. La seconde éventualité liée correspond à un acte : l'acte de penser, pour la relation épistémique, et l'acte de langage pour la relation pragmatique.

4.3 Relations inférentielles

Nous avons mis en évidence l'existence de relations pouvant recevoir une interprétation épistémique. Nous avons vu que dans certains cas, l'énoncé pouvait être envisagé selon deux interprétations différentes : une interprétation épistémique ou une interprétation selon laquelle la relation s'établit au niveau du contenu propositionnel, entre deux éventualités.

Nous avons trouvé dans le corpus d'autres relations pouvant recevoir une interprétation épistémique, mais celles-ci présentent des caractéristiques différentes sur le plan du contenu. En voici deux exemples :

(12) [BITNET était différent d'Internet]_7 [parce que c'était un réseau point-à-point de type « stocké puis transmis ».]_8
Relation annotée : Explication (7,8)

(13) [chaque ordre était égal à une voix.]_28 [Il y avait donc deux voix pour les privilégiés,]_29 [et une pour les non-privilégiés]_30
Relation annotée : Résultat (28,[29,30])

Sur le plan du contenu, les segments liés ne rapportent pas des éventualités, mais des faits. De plus, il existe un lien logique entre ces faits.

En (12), le fait rapporté dans le segment 8 implique le fait rapporté dans le segment 7 : être un réseau point-à-point de type « stocké puis transmis » implique d'être différent d'Internet. L'inférence est ici permise par la connaissance de la définition d'Internet : Internet n'est pas un réseau point-à-point de type « stocké puis transmis ».

[8] L'éventualité $e_{\alpha'}$ (ou $e_{\beta'}$) ne correspond pas nécessairement à *penser*. Elle peut renvoyer à d'autres attitudes mentales, comme : *croire*, *réaliser*, *savoir*, etc.

L'énoncé (13) fait référence aux trois ordres suivants : le clergé, la noblesse et le tiers-état. Le clergé et la noblesse correspondent à la part privilégiée de la population, contrairement au tiers-état. Attribuer une voix à chaque ordre implique donc d'attribuer deux voix pour les privilégiés et une voix pour les non-privilégiés. L'inférence relève des mathématiques.

Les exemples (12) et (13) font donc appel à des relations d'implication. La causalité ne pouvant se réduire à une simple implication logique, nous ne pouvons pas parler de lien causal entre les faits exposés.

Cependant, nous avons bien affaire à des relations d'Explication et de Résultat. Il s'agit de relations épistémiques. En effet, tout comme pour les relations décrites en 4.2, le locuteur justifie ses croyances personnelles, à la seule différence que ces croyances sont ici fondées sur une implication dont la valeur de vérité est démontrée et qu'elles se situent donc à un niveau de certitude différent.

Autrement dit, nous pouvons reformuler (11) en (14) et (12) en (15) :

(14) *Je pense que* la route moderne (entre Mariana et Aleria), au bas des collines, est un tracé traditionnel, car elle suit tout naturellement la limite du terrain ferme et du terrain alluvial et l'Itinéraire a pu choisir ce parcours.

(15) *Je sais que* BITNET est différent d'Internet parce que c'est un réseau point-à-point de type « stocké puis transmis ».

Afin de distinguer les relations décrites précédemment de celles décrites ici, nous appellerons les relations qui s'appuient sur l'existence d'un lien logique entre deux faits des relations *inférentielles*.

Ces relations ont en réalité déjà été introduites brièvement en SDRT. Bras, Le Draoulec et Asher (2009), dans une étude portant sur le connecteur *alors*, proposent une analyse d'un exemple issu de Jayez (1998). Nous le reprenons ci-dessous :

(16) Ce nombre est égal à 4. Alors il est pair.

Bras, Le Draoulec et Asher constatent que la relation qui s'établit entre K_α et K_β implique que si K_α est vrai alors normalement K_β est vrai, soit ($K_\alpha > K_\beta$).

Dans les exemples (12) et (13), nous sommes en présence du même type de relation. Ainsi, (13) respecte les effets sémantiques de la relation de Résultat Inférentiel, tels que décrits par Bras, Le Draoulec et Asher :

Résultat_Inférentiel_Conséquence $\Phi_{\text{Résultat\_inférentiel}(\alpha,\beta)} \Rightarrow (K_\alpha \wedge K_\beta \wedge (K_\alpha > K_\beta))$

Nous pouvons en déduire les effets sémantiques propres à la relation d'Explication Inférentielle auxquels satisfait (12) :

Explication_Inférentielle_Conséquence $\Phi_{\text{Explication\_inférentielle}(\alpha,\beta)} \Rightarrow (K_\alpha \wedge K_\beta \wedge (K_\beta > K_\alpha))$

5 Discussion sur la représentativité du corpus

Une première exploitation de notre corpus nous a permis de distinguer et de caractériser différents types de relations causales. Nous avons ainsi pu enrichir notre corpus d'annotations supplémentaires, déterminant pour chaque relation à quel type elle appartenait.

La typologie établie et le corpus ré-annoté, nous envisageons de poursuivre notre étude en procédant à des analyses quantitatives et comparatives.

A travers une réflexion sur la représentativité des corpus, nous présenterons les limites de notre corpus. Cette démarche nous amènera à envisager un élargissement de celui-ci pour la suite de nos analyses, élargissement qui devra être appréhendé selon des critères bien définis.

5.1 Nouvelles perspectives d'exploitation du corpus

L'étude de la répartition des différents types de relations causales, décrits dans la section 4, au sein du corpus semble indiquer des tendances différentes pour les relations d'Explication et les relations de Résultat :

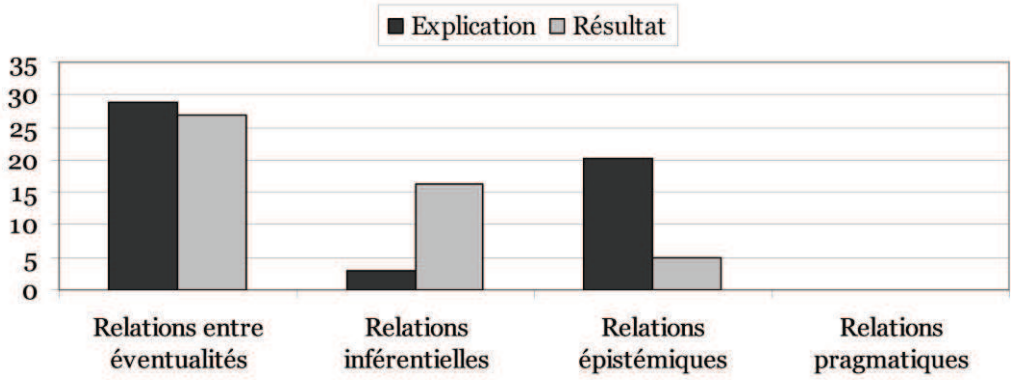

Figure 1 – Répartition des relations causales dans le corpus (en %)

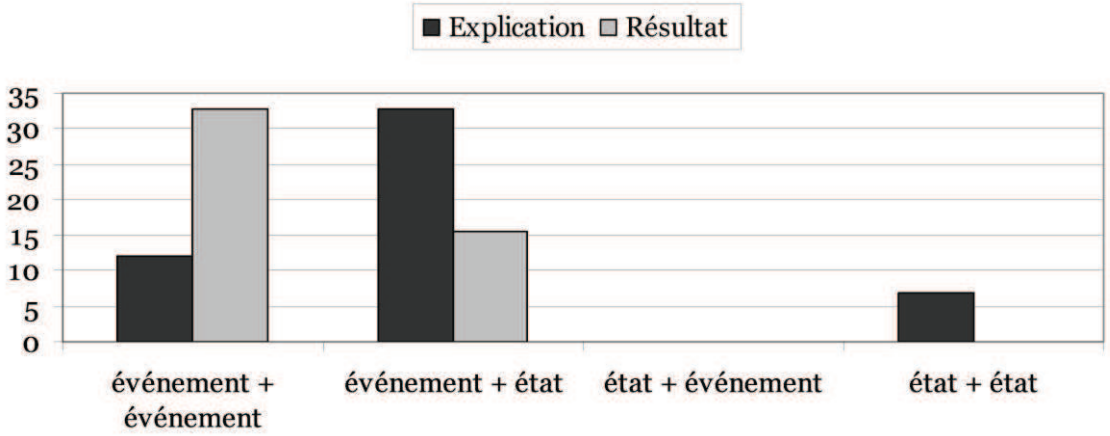

Figure 2 – Répartition dans le corpus des relations causales entre éventualités (en %)

A partir de l'observation des figures 1 et 2, nous pouvons formuler les hypothèses suivantes :

- Les relations inférentielles se présentent plus fréquemment sous la forme de relations de Résultat.
- Les relations épistémiques se présentent plus fréquemment sous la forme de relations d'Explication.
- Les liens causaux entre deux événements s'établissent le plus fréquemment sous la forme de relations de Résultat.
- Les liens causaux entre un événement et un état s'établissent le plus fréquemment sous la forme de relations d'Explication.

Autrement dit, il existerait un lien entre le type de relation causale et l'ordre des arguments respecté par cette relation.

Etant donnée la petite taille de notre corpus (qui compte au total 118 relations causales), nous

ne pouvons nous prononcer sur la validité de ces hypothèses. En effet, la question de la représentativité de notre corpus se pose.

La taille d'un corpus doit satisfaire les exigences de l'exploitation envisagée. Le corpus en l'état actuel nous a permis, lors d'une première exploration, de distinguer et caractériser différents types de relations. Sa petite taille nous a autorisée à procéder à une analyse linguistique de chaque occurrence des relations causales repérées pour répondre à la nécessité d'enrichir la théorie.

Or, pour des analyses quantitatives, la taille du corpus pose problème. Celle-ci doit être augmentée si l'on veut pouvoir obtenir des résultats statistiques pertinents. Habert (2000) parle d'*incertitude* pour désigner ce biais : « L'incertitude survient quand un échantillon est trop petit pour représenter avec précision la population réelle. »

Pour les mêmes raisons, notre corpus ne permet pas une étude satisfaisante des marqueurs potentiels de la causalité. Chaque marqueur relevé présente un nombre d'occurrences bien trop faible dans le corpus.

Pour la suite de nos analyses, la nécessité d'élargir notre corpus apparaît donc comme une évidence. L'intégration de nouveaux textes doit cependant se faire de façon réfléchie.

5.2 Un nouveau corpus pour une meilleure représentativité

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.

Nous avons vu que notre corpus ne satisfaisait pas le premier critère. Qu'en est-il du second ? Notre corpus, permet-il de rendre réellement compte de la causalité et de la diversité de ses réalisations ?

L'absence de relations causales pragmatiques dans notre corpus apporte une première réponse à nos interrogations : notre corpus ne rend pas compte de tous les types de relations causales.

Les effets sémantiques des relations pragmatiques impliquent qu'un des constituants liés soit un acte de langage. Cet acte de langage peut se présenter sous la forme d'un ordre (forme impérative) ou d'une question (forme interrogative). Par conséquent, la relation sera observée de préférence dans un contexte de dialogue. Notre corpus étant exclusivement constitué d'extraits de textes issus de brèves de presse et d'articles encyclopédiques, les situations de dialogue en sont absentes. Nous pourrions donc envisager d'intégrer à notre corpus des textes rapportant ce type de situations.

Cette réflexion sur le contexte d'apparition de relations causales spécifiques nous amène à envisager qu'il existerait un lien entre le genre textuel (ou type de texte) et le type de relation pouvant y être observé.

Dans cette perspective, intéressons-nous aux autres types de relations causales. Les relations épistémiques faisant appel aux attitudes mentales du locuteur, nous émettons l'hypothèse selon laquelle nous devrions pouvoir observer un nombre plus important de relations de ce type dans des textes argumentatifs, textes dans lesquels les marques de subjectivité sont généralement fréquentes. De même, les relations causales s'établissant au niveau du contenu propositionnel (relations d'Explication et de Résultat) devraient être plus fréquentes dans des textes narratifs, textes rapportant des éventualités qui entretiennent entre elles un lien temporel.

Il serait donc intéressant que notre corpus élargi puisse rendre compte de la diversité des genres textuels. Bien entendu, celui-ci ne pourra pas prétendre à l'exhaustivité (la notion de genre textuel étant de plus difficile à appréhender). Cependant, un corpus présentant une hétérogénéité interne permettrait d'envisager une confrontation inter-genres et donc de tester la validité des hypothèses que nous venons d'énoncer.

La construction d'un nouveau corpus, plus grand, plus diversifié et donc plus représentatif, devrait nous permettre de poursuivre nos analyses. Nous pourrons ainsi rendre compte des liens existant entre les quatre paramètres suivants :

- type de relation ;
- ordre des arguments de la relation ;
- genre textuel ;
- marquage de la relation.

6 Conclusion

Nous avons présenté dans cet article la démarche suivie pour une étude des relations de discours causales dans le cadre de la SDRT. Cette démarche se veut originale puisqu'elle envisage le corpus comme point de départ.

En nous appuyant sur le corpus issu du projet ANNODIS, nous avons constitué notre propre corpus de textes enrichis d'annotations dans le but de procéder à une étude approfondie des relations causales dans le discours.

L'observation des relations repérées dans ce corpus nous a permis de constater que la SDRT ne rendait pas compte de la diversité des relations causales observables dans les textes. Nous avons pu par conséquent proposer un enrichissement de la théorie, en proposant notamment une description des relations dites *épistémiques* et des relations dites *inférentielles*. Les différents types de relations causales se distinguent par leurs effets sémantiques. Nous avons, dans la section 4, proposé des règles rendant compte de ces différences.

Pour la suite de nos analyses, nous envisageons d'autres utilisations du corpus. Nous souhaitons en effet mettre en lumière les liens pouvant exister entre différents paramètres : le type de relation causale, l'ordre des arguments de la relation, les marqueurs associés à celles-ci et le genre textuel. Dans la section 5, nous avons évoqué les limites de notre corpus actuel pour de telles exploitations. Afin de rendre compte au mieux de la réalité du discours, notre corpus devra être élargi.

Nous envisageons de constituer un nouveau corpus plus représentatif. Pour cela, il devra présenter une taille suffisante et être construit de façon à rendre compte au mieux de la diversité des usages langagiers.

De plus, la construction d'un corpus selon ces critères devrait permettre à d'autres utilisateurs de l'exploiter pour leurs propres besoins. Par sa taille, la diversité des textes qui y seront représentés et les annotations proposées, notre corpus pourra faire office de corpus de référence pour l'étude des relations de discours causales.

Références

ARISTOTE (2007). *Rhétorique.* Paris, Garnier Flammarion.

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

BRAS, M., LE DRAOULEC, A. et ASHER, N. (2009). A formal analysis of the French Temporal Connective *alors*. In BEHRENS, B. et HANSEN, C. F. (éds.), *Information structure and Explicit versus Implicit Information in Text across languages*. Oslo, Osla.

DUCROT, O. et ANSCOMBRE, J.-C. (1983). *L'Argumentation dans la langue.* Bruxelles, Mardaga.

GROSS, G. (2009). *Sémantique de la cause.* Louvain-Paris, Peeters.

HABERT, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? In BILGER, M.

(éd.), *Linguistique sur corpus. Études et réflexions*. Perpignan, Presses Universitaires de Perpignan, pages 11-58.

HOBBS, R. (1985). *On the Coherence and Structure of Discourse. Report No. CSLI-85-37*. Stanford, Center for the Study of Language and Information, Stanford University.

HOVY, E. et MAIER, E. (1993). *Parsimonious or Profligate: How Many and Which Discourse Structure Relations? Technical report*. Los Angeles, USC Information Sciences Institute, University of Southern California.

HUME, D. (1748). *An Enquiry Concerning Human Understanding.* Oxford, Clarendon Press.

JAYEZ, J. et ROSSARI, C. (2001). The Discourse Level Sensitivity of Consequence Discourse Markers in French. *Cognitive Linguistics*, 12, pages 275-290.

JAYEZ, J. (1988). *Alors*, descriptions et paramètres. *Cahiers de Linguistique Française*, 9, pages 135-175.

KAMP, H. et REYLE, U. (1993). *From Discourse to Logic.* Dordrecht, Kluwer Academic Publishers.

KISTLER, M. (2004). La causalité dans la philosophie contemporaine. *Intellectica*, 38, pages 139-185.

LEWIS, D. (1973). Causation. *Journal of Philosophy*, 70, pages 556-567.

LASCARIDES, A. et ASHER, N. (1993). Temporal Interpretation, Discourse Relations and Commonsense Entailment. *Linguistics and Philosophy*, 16-5, pages 437-493.

MANN, W. C. et THOMPSON, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8-3, pages 243-281.

NAZARENKO, A. (2000). *La cause et son expression en Français*. Paris, Ophrys.

PERY-WOODLEY, M.-P., AFANTENOS, S.-D., HO-DAC, L.-M. et ASHER, N. (2012). Le corpus ANNODIS, un corpus enrichi d'annotations discursives. *TAL,* 53-2.

PERY-WOODLEY, M.-P., ASHER, N., ENJALBERT, P., BANAMARA, F., BRAS, M., FABRE, C., FERRARI, S., HO-DAC, L.-M., LE DRAOULEC, A., MATHET, Y., MULLER, P., PREVOT, L., REBEYROLLE, J., TANGUY, L., VERGEZ-COURET, M., VIEU, L. et WILDÖCHER, A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. *TALN 2009*, Senlis.

PLANTIN, C. (1990). *Essais sur l'argumentation : Introduction linguistique à l'étude de la parole argumentative*. Paris, Kimé.

RUSSEL, B. (1912). *On the Notion of Cause*. London, Routledge.

SANDERS, T., SPOOREN, W. et NOORDMAN, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15-1, pages 1–35.

SANDERS, T. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes*, *24*, pages 119-147.

SWEETSER, E. (1990). *From Etymology to Pragmatics.* Cambridge, Cambridge University Press.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam et Philadelphia, John Benjamins.