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

Vers une algèbre des relations de discours pour la comparaison de structures discursives

Charlotte Roze
Alpage, INRIA Paris–Rocquencourt & Université Paris 7
charlotte.roze@linguist.jussieu.fr

Résumé. Nous proposons une méthodologie pour la construction de règles de déduction de relations de discours, destinées à être intégrées dans une algèbre de ces relations. La construction de ces règles a comme principal objectif de pouvoir calculer la fermeture discursive d'une structure de discours, c'est-à-dire de déduire toutes les relations que la structure contient implicitement. Calculer la fermeture des structures discursives peut permettre d'améliorer leur comparaison, notamment dans le cadre de l'évaluation de systèmes d'analyse automatique du discours. Nous présentons la méthodologie adoptée, que nous illustrons par l'étude d'une règle de déduction.

Abstract. We propose a methodology for the construction of discourse relations inference rules, to be integrated into an algebra of these relations. The construction of these rules has as main objective to allow for the calculation of the discourse closure of a structure, i.e. deduce all the relations implicitly contained in the structure. Calculating the closure of discourse structures improves their comparison, in particular within the evaluation of discourse parsing systems. We present the adopted methodology, which we illustrate by the study of a rule.

Mots-clés : Relation de discours, fermeture discursive, évaluation, déduction.

Keywords: Discourse relation, discourse closure, evaluation, inference.

1 Introduction

L'analyse rhétorique (ou discursive) d'un texte a pour but de représenter sa structure globale, c'est-à-dire les liens qui s'établissent entre les différentes parties du texte, permettant à son lecteur de l'interpréter comme formant un tout cohérent, et pas comme une simple succession de phrases indépendantes les unes des autres. Ces liens sont appelés *relations rhétoriques* ou relations de discours. Ils s'établissent entre des *segments de discours*, qui couvrent des propositions, des phrases et/ou de plus larges portions du texte. Différentes théories et formalismes, comme la RST (*Rhetorical Structure Theory*, Mann & Thompson, 1988), la SDRT (*Segmented Discourse Representation Theory*, Asher & Lascarides, 2003), D–LTAG (*Discourse Lexicalized Tree Adjoining Grammar*, Webber, 2004), et D–STAG (*Discourse Synchronous Tree Adjoining Grammar*, Danlos, 2009), proposent de représenter ce type de structures. Dans le travail présenté ici, le cadre théorique adopté est la SDRT.

Le traitement automatique du discours vise principalement à développer des systèmes permettant de générer des analyses de la structure discursive d'un texte. Dans cette perspective, la constitution de corpus de référence et l'évaluation des annotations produites par les systèmes d'analyse automatique sont des tâches primordiales. Les corpus de référence fournissent des données aux systèmes basés sur des méthodes d'apprentissage et permettent d'évaluer les annotations en sortie d'un analyseur. La constitution de ces corpus nécessite bien souvent la « fusion » de différentes annotations d'un même texte, donc la comparaison de structures discursives. L'évaluation des annotations générées par un système implique elle aussi la comparaison de structures discursives : les structures contenues dans les annotations du système et les structures contenues dans les annotations de référence.

Les questions qui se posent dans un objectif de construction d'une référence ou d'évaluation sont donc les suivantes : comment comparer deux annotations en discours ? quelles structures de discours sont équivalentes ou compatibles ? En effet, deux annotations discursives d'un même texte peuvent différer sans que l'une ou l'autre soit pour autant « fausse » ou « incomplète ». Considérons par exemple le discours en (1), qui contient trois segments de discours, que nous nommons (π_1), (π_2) et (π_3). On peut avoir deux annotations différentes et néanmoins équivalentes pour ce discours : une première annotation A_1, contenant les relations *Result*(π_1, π_2) et

Elaboration(π_2, π_3) ; une seconde annotation A_2 contenant ces deux mêmes relations et la relation *Result*(π_1, π_3). Les deux annotations diffèrent, mais sont équivalentes : dans l'annotation A_1, il est renseigné que (π_3) élabore (π_2), qui lui-même décrit un résultat de (π_1) ; il est donc implicitement renseigné que (π_3) décrit un résultat de (π_1), ce qui signifie que l'on peut déduire l'annotation A_2 à partir de l'annotation A_1.

1. (a) Il a beaucoup plu aujourd'hui. (π_1)
   (b) *Du coup*, Jean n'a pas pu faire ce qu'il avait prévu. (π_2)
   (c) Il n'a pas pu faire son footing, *notamment*. (π_3)

Dans le cadre de la construction d'une annotation de référence pour un texte donné, on veut pouvoir intégrer les informations présentes dans les différentes annotations de ce texte, si bien sûr elles sont compatibles. Dans le cadre de l'évaluation, on veut savoir si certaines informations absentes dans une annotation A_1 (que ce soit la référence ou non) et présentes dans une annotation A_2 sont en réalité implicitement renseignées dans l'annotation A_1. Autrement dit, on voudrait pouvoir déduire toutes les informations (les relations de discours) implicitement présentes dans les deux annotations à comparer.

Étant donnée une structure discursive associée à un texte, composée de relations de discours entre les segments qui le constituent, notre objectif est donc de pouvoir calculer, à l'aide de règles de déduction, la *fermeture discursive* de la structure, c'est-à-dire toutes les relations de discours qui peuvent être déduites à partir des relations déjà annotées. Pour calculer la fermeture discursive d'une structure, des règles de déduction de relations de discours sont nécessaires. Par exemple, pour calculer la fermeture discursive de l'annotation A_1 du discours en (1), nous avons besoin de la règle suivante : *Result*$(\pi_1, \pi_2) \wedge$ *Elaboration*$(\pi_2, \pi_3) \rightarrow$ *Result*(π_1, π_3). Or, les théories du discours ne définissent pas (ou peu) de règles de ce type. Nous proposons d'étudier et de construire des règles de déduction de relations de discours, destinées à être intégrées dans une algèbre des relations de discours, similaire à celle construite par Allen (1983) pour les relations temporelles.

L'ensemble de relations rhétoriques utilisé varie selon les théories, ainsi que la façon dont ces relations sont définies. Néanmoins, il existe un consensus sur un certain nombre de relations, comme par exemple les relations *Elaboration* et *Narration*. De plus, les relations peuvent généralement être mises en correspondance d'une théorie à une autre : par exemple, la relation *Result* de la SDRT recouvre les relations *Volitional Result* et *Non-Volitional Result* de la RST. Nous avons choisi de placer ce travail dans le cadre théorique de la SDRT (Asher & Lascarides, 2003), parce que : d'une part, cette théorie se trouve, en ce qui concerne la définition des relations de discours, à un niveau de granularité intermédiaire entre les approches multiplicatrices comme celle de la RST, qui construisent des listes étendues de relations, et les approches réductionnistes, comme celle de Grosz & Sidner (1986), qui proposent de ne distinguer que deux relations structurelles : *dominates* (dominance) et *satisfaction-precedence* (satisfaction de la précédence) ; d'autre part, elle rend explicites les contraintes sémantiques établies par une relation de discours sur ses arguments, et ces contraintes sont le point de départ de l'étude des règles de déduction. Dans la SDRT, l'établissement d'une relation donnée impose des contraintes (temporelles, causales, structurelles) sur les deux segments qu'elle relie : par exemple, la relation *Narration* implique une précédence temporelle entre les éventualités qu'elle relie, la relation *Result* implique une relation causale entre deux éventualités, etc. (voir section 3.2). La structure hiérarchique du discours est représentée à l'aide d'une distinction entre relations coordonnantes et relations subordonnantes, ce qui permet également de restreindre, dans un discours donné, l'ensemble des segments disponibles pour l'attachement de nouveaux segments dans un discours.

Dans la section 2, nous justifions l'utilisation de règles de déduction dans l'évaluation des structures discursives, en montrant pourquoi les métriques d'évaluation utilisées en analyse syntaxique ne permettent pas d'évaluer correctement des analyses discursives. Dans un second temps, nous présentons des travaux qui s'intéressent aux relations temporelles, et qui proposent d'améliorer l'évaluation de graphes temporels à l'aide de règles de déduction. Dans la section 3, nous présentons une méthodologie pour la construction des règles de déduction. Dans la section 4, nous illustrons cette méthodologie par l'étude et la construction complète d'une règle. Pour terminer, dans la section 5, nous concluons en apportant des perspectives dans la construction des règles de déduction.

2 Évaluation de graphes discursifs

Les travaux en traitement automatique du discours, et plus précisément le développement de systèmes d'analyse automatique de la structure discursive, posent, comme toutes les tâches de TAL, la question cruciale de l'évaluation : il faut évaluer les analyses produites par les systèmes, en prenant en compte les particularités des structures

discursives. L'analyse de la structure discursive d'un texte comprend deux tâches distinctes : la *segmentation* en unités minimales de discours (appelés *segments élémentaires*), et l'*annotation des relations* existant entre les différents segments (certains de ces segments recouvrent plusieurs segments élémentaires, et sont appelés *segments complexes*). Nous ne traiterons ici que de l'évaluation de la seconde étape.

Les structures discursives sont, dans la SDRT, représentées par des graphes, dont les noeuds sont des segments de discours (qui recouvrent des portions plus ou moins larges d'un texte : propositions, phrases, paragraphes, etc.), et dont les arcs sont des relations de discours (par exemple : *Narration*, *Explanation*, etc.). Dans la RST, les structures de discours sont représentées par des arbres binaires étiquetés (Marcu, 1996), dans lesquels les feuilles sont des segments élémentaires, les autres noeuds sont des relations rhétoriques, les étiquettes des arcs décrivant le type des arguments de relations (*Nucleus* ou *Satellite*).

Dans cette section, nous décrivons succinctement certaines métriques utilisées dans l'évaluation de structures proches des structures discursives : les arbres de constituants syntaxiques (proches des arbres de la RST, si l'on fait un parallèle entre les segments de discours et les mots, et un autre entre les relations rhétoriques et les constituants), et les graphes de dépendances syntaxiques (proches des graphes de la SDRT). Nous verrons pourquoi ces métriques ne permettent pas d'évaluer correctement des analyses discursives. Dans un second temps, nous présentons des travaux qui s'intéressent aux relations temporelles, et qui proposent d'améliorer l'évaluation de graphes temporels à l'aide de règles de déduction. Pour terminer, nous proposons d'utiliser des règles de déduction pour l'évaluation de graphes discursifs, et discutons de la forme de ces règles.

Les systèmes de TAL sont généralement évalués en termes de *rappel*, *précision* et *F-score* (adaptés à l'évaluation de la classification d'objets indépendants), mais diverses métriques sont utilisées, selon la tâche et le type de structure à évaluer : métrique Parseval en analyse syntaxique, métrique BLEU (qui s'applique à des séquences de mots) en traduction automatique, etc. En analyse syntaxique, les arbres de constituants sont le plus souvent évalués suivant la métrique Parseval (Black *et al.*, 1991), qui calcule, pour un arbre syntaxique, la précision et le rappel des constituants, en partant du principe suivant : un constituant dans l'arbre calculé par l'analyseur est correct s'il existe un constituant dans l'arbre correspondant du corpus de référence qui domine la même séquence de symboles terminaux et possède le même label. Elle calcule également la moyenne des constituants dans un arbre qui « croisent » des frontières de constituants dans l'autre arbre (*crossing brackets*). Dans l'évaluation de dépendances syntaxiques, les métriques utilisées (Nivre & Scholz, 2004) sont généralement : UAS (*unlabelled attachment score*), qui calcule la proportion de mots pour lesquels le gouverneur assigné est correct ; LAS (*labelled attachment score*), qui calcule la proportion de mots pour lesquels le gouverneur assigné est correct et le type de dépendance est correct ; et LabAcc (*labelled accuracy score*), qui calcule la proportion de mots pour lesquels le type de dépendance est correct.

Si l'on tente d'adapter les métriques utilisées dans l'évaluation de dépendances syntaxiques pour évaluer des graphes discursifs (en remplaçant les mots par les segments de discours, et les dépendances syntaxiques par les relations de discours), on constate qu'elles ne permettent pas toujours une évaluation satisfaisante. Considérons par exemple le discours en (2). Dans ce discours, le segment (π_1) est élaboré par les segments (π_2) et (π_3) : ces deux segments décrivent une partie du repas de Jean mentionné en (π_1). La description du repas est faite dans l'ordre chronologique, les segments (π_2) et (π_3) forment donc une narration. Si l'on adapte les métriques UAS, LAS et LabAcc pour évaluer une annotation A de ce discours contenant $\textit{Elaboration}(\pi_1, \pi_2) \wedge \textit{Narration}(\pi_2, \pi_3)$, en prenant comme référence l'annotation R de ce même discours contenant $\textit{Elaboration}(\pi_1, [\pi_2, \pi_3]) \wedge \textit{Narration}(\pi_2, \pi_3)$, on n'obtient pas un score de 1. Pourtant, étant donné la sémantique des relations en jeu, on peut déduire que les deux structures annotées sont équivalentes. En effet, l'annotation A permet de déduire l'annotation R : les informations a priori « manquantes » en A sont en réalité implicitement présentes.

2. (a) Jean a fait un excellent repas. (π_1)
   (b) Il a mangé un délicieux saumon, (π_2)
   (c) puis s'est régalé d'un copieux plateau de fromages. (π_3)

Il est donc nécessaire d'utiliser des règles de déduction pour procéder à une évaluation précise de structures discursives. Concernant le discours en (2), la SDRT possède une règle qui permet d'établir l'équivalence entre l'annotation A et la référence R : $\textit{Elaboration}(\alpha, \beta) \wedge \textit{Narration}(\beta, \gamma) \rightarrow Elaboration(\alpha, [\beta, \gamma])$. Cependant, la question des équivalences entre structures discursives reste très peu étudiée dans les théories du discours, et pour une majorité de couples de relations de discours (R_1, R_2), nous ne savons pas (par exemple) si la structure $R_1(\alpha, \beta) \wedge R_2(\beta, \gamma)$ contient une information implicite, et si oui, laquelle. Nous proposons donc d'étudier et de définir des règles de déduction permettant une meilleure évaluation des graphes discursifs. L'idée d'utiliser des

règles de déduction pour compléter et évaluer des annotations a déjà été exploitée en ce qui concerne les relations temporelles (Setzer *et al.*, 2003) et les relations de coréférence (Vilain *et al.*, 1995). Par exemple, si une annotation temporelle contient les informations : l'événement e_1 a lieu avant e_2 ($e_1 < e_2$) et l'événement e_2 a lieu avant e_3 ($e_2 < e_3$), alors il est implicite que l'événement e_1 a lieu avant e_3 (et l'on peut déduire $e_1 < e_3$). De la même façon, si dans une annotation en chaînes de coréférence les informations suivantes sont présentes : l'expression e_1 coréfère avec e_2 et l'expression e_2 coréfère avec e_3, alors il est implicite que l'expression e_1 coréfère avec e_3.

En ce qui concerne les relations temporelles, Allen (1983) définit une algèbre temporelle complète, avec des règles de la forme : $r_1(A, B) \wedge r_2(B, C) \rightarrow r_3(A, C)$[1]. L'algèbre utilise 13 relations (*before*, *during*, *overlaps*, etc.). Dans beaucoup de cas, il existe plus d'une relation r_3 déductible entre les intervalles A et C. Par exemple : *overlaps*$(A, B) \wedge$ *overlaps*$(B, C) \rightarrow$ *before*$(A, C) \vee$ *overlaps*$(A, C) \vee$ *meets*(A, C). Setzer *et al.* (2003) introduisent la notion de *fermeture temporelle* d'un graphe temporel, qui est la représentation complète des conséquences temporelles du graphe. Pour aider à la création de corpus de référence, ils proposent de comparer deux annotations temporelles d'un même texte en termes d'équivalence ou de recouvrement de leurs fermetures temporelles, qui sont calculées à partir de règles d'inférences similaires à celles définies par Allen (1983). Ces règles peuvent également servir d'aide à l'annotation. Tannier & Muller (2008) proposent une autre méthode de comparaison de graphes temporels. Ils distinguent deux types de relations dans un graphe temporel : les relations essentielles, et les relations qui peuvent être déduites à partir d'autres relations. Ils proposent d'effectuer la comparaison et l'évaluation d'annotations temporelles uniquement à partir des relations essentielles, ce qui nécessite également l'exploitation de règles de déduction.

Nous proposons de construire une algèbre des relations de discours, inspirée de l'algèbre des relations temporelles construite par Allen (1983), afin d'améliorer la qualité de l'évaluation de graphes discursifs : étant donnée une annotation discursive d'un texte, composée de relations de discours entre les segments qui le constituent, notre objectif est de calculer, à l'aide de règles de déduction, la *fermeture discursive* de l'annotation, c'est-à-dire toutes les relations de discours qui peuvent être déduites à partir des relations déjà annotées. Pour construire une algèbre des relations de discours, au moins deux types de règles semblent nécessaires. En effet, si l'on considère un discours quelconque à trois segments (α), (β) et (γ), il y a potentiellement deux structures (présentées à la Figure 1) pour lesquelles une relation reste indéterminée (en pointillés dans la figure) : dans la première structure, la relation entre (α) et (γ) n'est pas explicite ; dans la seconde, c'est la relation entre (β) et (γ) qui ne l'est pas[2]. Considérant ces deux structures, nous proposons d'utiliser deux types de règles de déduction[3] (représentés à la Figure 1). Notons que la déduction peut être une disjonction de relations, comme dans l'algèbre de Allen.

FIG. 1 – Schémas de déduction pour les deux types de règles : déduction de R_z ou déduction de R_y (où (α), (β) et (γ) représentent trois segments de discours successifs)

3 Méthodologie pour la construction d'une algèbre des relations de discours

Nous détaillons dans cette section la méthodologie adoptée dans l'étude et la construction des règles de déduction. Pour plus de lisibilité, nous ne présentons la méthodologie que pour le premier type de règle proposé (à gauche dans la Figure 1). La construction des règles se fait prémisse par prémisse. Dans la section 3.1, nous décrivons

[1] A, B et C représentent des intervalles temporels, et r_1, r_2 et r_3 des relations temporelles.

[2] Dans la SDRT, la seconde structure n'est valide que si R_x est une relation subordonnante : selon la *contrainte de la frontière droite*, si R_x est coordonnante, (γ) ne peut théoriquement pas être « attaché » à (α). Cependant, ces structures en théorie invalides ne sont pas nécessairement exclues des annotations en discours, nous les prenons donc en compte dans les règles de déduction à étudier.

[3] Si les relations traitées étaient des relations temporelles, le cas $R_x(\alpha, \beta) \wedge R_z(\alpha, \gamma)$ pourrait être ramené au cas : $R_x^{-1}(\beta, \alpha) \wedge R_z(\alpha, \gamma)$, où R_x^{-1} est la relation « inverse » de R_x. En effet, dans l'algèbre de Allen, il n'existe que des règles dont la prémisse est $r_1(A, B) \wedge r_2(B, C)$. En ce qui concerne les relations de discours, il nous faut distinguer les deux cas, puisqu'il n'existe pas de relation « inverse » à chaque relation de discours. De plus, l'ordre des segments dans un discours a un impact sur sa structure et son interprétation.

le travail d'extraction de règles candidates à partir du corpus ANNODIS (Péry-Woodley *et al.*, 2009)[4]. Dans la section 3.2, nous présentons les éléments qui nous permettent de mettre en évidence les différentes déductions possibles pour une prémisse de règle donnée. Dans la section 3.3, nous montrons comment les règles sont validées par l'annotation de données empiriques, extraites à partir de corpus non annotés en discours.

3.1 Extraction de déductions candidates pour la construction des règles

Pour construire une algèbre des relations de discours utilisant environ 20 relations, il y a potentiellement 8000 règles du type $R_x(\alpha,\beta) \wedge R_y(\beta,\gamma) \rightarrow R_z(\alpha,\gamma)$ à étudier (il y a 20×20 prémisses de règles possibles, et pour chaque prémisse 20 déductions possibles), plus 8000 règles du type $R_x(\alpha,\beta) \wedge R_z(\alpha,\gamma) \rightarrow R_y(\beta,\gamma)$, donc environ 16000 règles en tout. Chaque règle nécessitant une étude linguistique, il est impératif de faire une sélection dans les règles à étudier en priorité, et de dégager les règles candidates les plus pertinentes. Pour extraire des déductions candidates, nous exploitons le corpus ANNODIS, constitué de 100 textes annotés en discours, provenant notamment de Wikipédia et du corpus de l'Est Républicain. Après avoir été segmenté en unités minimales de discours, chaque texte du corpus a reçu deux annotations distinctes, avec la SDRT comme point de départ au guide d'annotation. 19 relations de discours sont utilisées dans le corpus. Nous les présentons dans le tableau 1, avec leur nombre d'occurrences dans les annotations.

Relation	Nombre d'occurrences	Relation	Nombre d'occurrences
Elaboration	1662	*Parallel*	154
Entity Elaboration	1169	*Attribution*	151
Continuation	658	*Background*	134
Narration	567	*Flashback*	106
Frame	416	*Description Continuation*	54
Contrast	334	*Conditional*	53
Result	303	*Alternation*	35
Explanation	259	*Source*	18
Goal	238	*Explanation**	12
Commentary	222	*Result**	0

TAB. 1 – Relations utilisées dans le corpus ANNODIS

Nous dégageons à partir de ce corpus des règles de déduction candidates : pour chaque prémisse de règle de la forme $R_x(\alpha,\beta) \wedge R_y(\beta,\gamma)$ nous extrayons la ou les relation(s) $R_z(\alpha,\gamma)$ les plus probables. Nous nous intéressons donc aux annotations contenant des triplets de segments de discours (α, β, γ) pour lesquels les relations sont saturées, c'est-à-dire telles que : une relation (au moins) a été annotée entre (α) et (β), de même entre (β) et (γ), ainsi qu'entre (α) et (γ). Les triplets « saturés » du corpus permettent de calculer, pour toute relation R_z, la probabilité que la relation $R_z(\alpha,\gamma)$ soit établie sachant que les relations $R_x(\alpha,\beta)$ et $R_y(\beta,\gamma)$ sont présentes :

$$P(R_z(\alpha,\gamma) \mid R_x(\alpha,\beta) \wedge R_y(\beta,\gamma)) = \frac{count(R_x(\alpha,\beta),R_y(\beta,\gamma),R_z(\alpha,\gamma))}{count(R_x(\alpha,\beta),R_y(\beta,\gamma),R(\alpha,\gamma))}$$ [5].

Ces probabilités nous donnent, pour une prémisse de règle donnée, une idée de la plausibilité des déductions : plus la probabilité $P(R_z(\alpha,\gamma) \mid R_x(\alpha,\beta) \wedge R_y(\beta,\gamma))$ est grande, plus la règle $R_x(\alpha,\beta) \wedge R_y(\beta,\gamma) \rightarrow R_z(\alpha,\gamma)$ est plausible. Dans l'exploitation du corpus, nous réécrivons les relations impliquant des segments complexes[6] avec les règles de réécriture suivantes : $R_1(\alpha,[\beta,\gamma]) \wedge R_2(\beta,\gamma) \rightarrow_{rew} R_1(\alpha,\beta) \wedge R_1(\alpha,\gamma) \wedge R_2(\beta,\gamma)$; et $R_1(\alpha,\beta) \wedge R_2([\alpha,\beta],\gamma) \rightarrow_{rew} R_1(\alpha,\beta) \wedge R_2(\alpha,\gamma) \wedge R_2(\beta,\gamma)$. La réécriture permet d'exploiter un plus grand nombre de données du corpus. Par exemple, nous obtenons 0 occurrence du triplet (*Result*(α,β), *Contrast*(β,γ), *Contrast*(α,γ)) sans la réécriture, et 127 occurrences avec la réécriture. La réécriture des segments complexes permet donc de dégager une règle candidate comme *Result*$(\alpha,\beta) \wedge$ *Contrast*$(\beta,\gamma) \rightarrow$ *Contrast*(α,β)[7] à partir

[4] Les données du corpus ANNODIS ne sont pas encore disponibles, mais nous ont été gentiment fournies par les membres du projet.

[5] $R(\alpha,\gamma)$ signifie qu'il existe une relation entre (α) et (γ).

[6] Dans un discours, certains segments élémentaires sont à la fois arguments d'une relation de discours, et à la fois partie d'un segment de discours plus large (appelé segment complexe) qui est lui-même argument d'une relation de discours. Par exemple, la structure du discours en (2) de la section 2 est *Elaboration*$(\pi_1,[\pi_2,\pi_3]) \wedge$ *Narration*(π_2,π_3), où (π_2) et (π_3) sont les deux arguments de la relation *Narration*, mais sont aussi inclus dans un segment *complexe* plus large ($[\pi_2,\pi_3]$) qui élabore (π_1).

[7] Voir la section 4 pour une étude détaillée de la prémisse *Result*$(\alpha,\beta) \wedge$ *Contrast*(β,γ).

des structures annotées *Result*$(\alpha, \beta) \wedge$ *Contrast*$([\alpha, \beta], \gamma)$ dans le corpus. Notons que la sémantique de $R_1(\alpha, \beta)$ devient alors : « la relation R_1 est établie entre (α) (ou un segment contenant (α)) et (β) (ou un segment contenant (β)) ».

3.2 Mise en évidence des déductions possibles

Pour construire les règles de déduction, nous effectuons une première étude de la prémisse de règle considérée, qui nous permet de dégager la ou les déduction(s) possible(s), et dans certains cas de dégager des paramètres influant sur la déduction. Cette étude s'appuie sur différents éléments : les déductions candidates extraites à partir du corpus ANNODIS, la sémantique des relations présentes dans la prémisse et dans les déductions candidates, et l'analyse d'exemples construits et attestés[8].

Les probabilités calculées sur le corpus ANNODIS nous permettent de dégager la/les déduction(s) plausible(s) : plus la probabilité $P(R_z(\alpha, \gamma) \mid R_x(\alpha, \beta) \wedge R_y(\beta, \gamma))$ est grande, plus la règle $R_x(\alpha, \beta) \wedge R_y(\beta, \gamma) \rightarrow R_z(\alpha, \gamma)$ est plausible. Cependant, nous ne pouvons pas nous baser sur ces seules informations pour dégager des règles de déduction, en partie parce que ces calculs exploitent des discours où toute l'information est explicitée par l'annotateur, et que nous voulons mettre au jour l'information *implicite* établie par la présence des relations des prémisses de règles. Pour cela, nous devons examiner, pour la prémisse de règle considérée, les conséquences de l'établissement des relations $R_x(\alpha, \beta)$ et $R_y(\beta, \gamma)$. En effet, lorsqu'une relation de discours est établie, elle impose certaines contraintes sur ses arguments et sur les liens qui existent entre eux. La nature de ces contraintes varie selon les relations de discours. Certaines relations, comme *Narration*, établissent des contraintes temporelles sur les éventualités qu'elles relient : si l'on a *Narration*(α, β), alors l'éventualité e_α décrite dans le premier segment (α) a lieu avant l'éventualité e_β décrite dans le second segment. D'autres relations, comme *Explanation*, établissent des relations causales : si l'on a *Explanation*(α, β), alors l'éventualité e_β décrite dans le second segment est la cause de l'éventualité e_α décrite dans le premier segment. Par exemple, dans le discours : *Jean est tombé. Max l'a poussé.*, l'éventualité décrite dans la seconde phrase est la cause de l'éventualité dans la seconde phrase.

Observer les contraintes établies par les relations R_x et R_y permet de mieux caractériser le lien existant entre les segments (α) et (γ), en exploitant la définition théorique des relations de discours en jeu. Les contraintes établies par les relations contenues dans la prémisse de la règle nous permettent ainsi dans certains cas de prédire la déduction (voir section 4.2.3), d'un point de vue théorique au moins. De plus, les contraintes établies par les relations contenues dans la prémisse nous permettent d'exclure certaines relations des déductions possibles, car les conséquences de $R_x(\alpha, \beta)$ et $R_y(\beta, \gamma)$ sont incompatibles avec l'établissement de $R_z(\alpha, \gamma)$. Par exemple, lorsque les relations *Narration*(α, β) et *Narration*(β, γ) sont établies, la relation *Flashback* ne peut lier (α) et (β). Cette incompatibilité est illustrée dans le discours en (3), où l'on a *Narration*(π_1, π_2) et *Narration*(π_2, π_3), avec pour conséquences temporelles : l'événement décrit en (π_1) a lieu avant l'événement décrit en (π_2) $(e_{\pi_1} < e_{\pi_2})$; l'événement décrit en (π_2) a lieu avant l'événement décrit en (π_3) $(e_{\pi_2} < e_{\pi_3})$; donc l'événement décrit en (π_1) a lieu avant l'événement décrit en (π_3) $(e_{\pi_1} < e_{\pi_3})$, ce qui est incompatible avec l'établissement de la relation *Flashback*(π_1, π_3), car sa conséquence temporelle est : $e_{\pi_3} < e_{\pi_1}$.

3. (a) Aujourd'hui, Julie est allée voir une expo. (π_1)
   (b) *Ensuite*, elle a déjeuné avec des amis. (π_2)
   (c) *Puis* elle a fait des courses au marché. (π_3)

L'étude des règles de déduction nécessite également un travail d'introspection, consistant à construire des discours comportant la prémisse de règle étudiée. Ce travail permet de vérifier les hypothèses formulées à partir des calculs effectués sur le corpus et des contraintes établies par les relations de discours. Dans la construction des discours, on essaie de couvrir le plus grand nombre de cas de figure. Pour cela, selon les relations présentes dans la prémisse, nous faisons varier certains paramètres. Par exemple, lors de la construction de discours impliquant la relation *Contrast*, on peut établir des contrastes où : une même entité présente deux propriétés distinctes, deux entités distinctes présentent chacune une propriété, une négation est présente dans une des propositions, etc. ; pour les discours contenant la relation *Explanation*(α, β), l'éventualité décrite dans le segment (α) peut être un événement, un état, une cause future, etc. Pour analyser les données construites ou attestées, et vérifier la déduction d'une relation entre (α) et (γ), on utilise les tests d'insertion d'un connecteur et de réorganisation du discours, décrits dans les deux paragraphes suivants.

[8] Ces exemples attestés proviennent du corpus ANNODIS et des corpus de discours extraits grâce à l'outil présenté à la section 3.3.

Insertion d'un connecteur L'insertion d'un connecteur permet de tester la présence d'une relation de discours entre (α) et (γ) dans l'étude de règles de la forme : $R_x(\alpha,\beta) \wedge R_y(\beta,\gamma) \rightarrow R_z(\alpha,\gamma)$. Si l'on veut vérifier la présence d'une relation R_z donnée entre les segments (α) et (γ) d'un discours, on peut utiliser le test suivant : si, après avoir inséré dans le segment (γ) un connecteur adverbial lexicalisant R_z, le discours reste cohérent et que son interprétation est inchangée, alors la présence de la relation $R_z(\alpha,\gamma)$ est vérifiée. Observons les discours en (4), dans lesquels les relations *Result*(π_1,π_2) et *Explanation*(π_2,π_3) sont établies. On constate à l'aide du test d'insertion que : pour (4c-i), le connecteur de résultat *du coup* peut être inséré sans rendre le discours incohérent, et sans en modifier l'interprétation, donc la relation *Result*(π_1,π_3) peut être inférée ; pour le discours en (4c-ii), en revanche, l'insertion d'un connecteur de résultat rend le discours incohérent, donc la relation *Result*(π_1,π_3) ne peut pas être établie.

4. (a) L'électricité est revenue ce matin. (π_1)
   (b) Les habitants sont très contents, (π_2)
   (c) i. *car* ***du coup*** ils ont pu regagner leurs appartements. (π_3)
   ii. *car* (**# *ainsi* / # *du coup***) ils ont besoin de chauffage. (π_3)

Réorganisation du discours Une autre méthode pour mettre au jour la relation établie entre les segments (α) et (γ) est de réorganiser le discours en échangeant la position des segments (β) et (γ). Le connecteur qui lexicalise la relation $R_y(\beta,\gamma)$ est remplacé par un connecteur lexicalisant la relation « inverse » ${R_y}^{-1}(\gamma,\beta)$ si elle existe : par exemple, si l'on a la relation *Result*(β,γ), on utilise un marqueur de la relation *Explanation*(γ,β). Dans le discours ainsi formé, si un connecteur lexicalisant R_z peut être inséré entre les segments (α) et (γ) sans rendre le discours incohérent et en conservant l'interprétation d'origine, alors on peut déduire que la relation $R_z(\alpha,\gamma)$ s'établit dans le discours d'origine. Par exemple, en (5b-i), la réorganisation du discours en (4c-i) nous permet de mettre en évidence la présence de la relation *Result*(π_1,π_3), car le connecteur de résultat *du coup* peut être inséré. En revanche, la réorganisation du discours (4c-ii) nous permet d'exclure la présence de la relation *Result* entre les segments (π_1) et (π_3), car le discours en (5b-ii) est incohérent.

5. (a) L'électricité est revenue ce matin. (π_1)
   (b) i. *Du coup*, les habitants ont pu regagner leurs appartements. (π_3)
   ii. # *Du coup*, ils ont besoin de chauffage. (π_3)
   (c) *Donc* ils sont très contents. (π_2)

3.3 Annotation pour la validation des règles construites

Après avoir mis en évidence les déductions possibles pour une prémisse donnée, on réalise une étude systématique sur un corpus de plusieurs centaines de discours contenant la prémisse étudiée : pour chaque discours du corpus collecté, on annote la relation déduite[9]. Cette annotation permet d'une part de vérifier la validité des hypothèses formulées quant aux déductions possibles, et aux paramètres ayant un impact sur la déduction ; d'autre part, elle permet, dans le cas où plusieurs déductions sont possibles pour une prémisse donnée, de connaître la fréquence de chacune des déductions.

Pour mener une telle étude, le corpus ANNODIS ne contient généralement pas suffisamment de triplets de segments (α,β,γ) au sein desquels les prémisses étudiées (et seulement les prémisses) sont établies. Pour constituer un corpus d'exemples suffisamment grand, nous avons donc développé un outil permettant d'extraire des discours contenant les prémisses de règles à partir de données non-annotées en discours. L'identification de la présence des relations de discours en jeu dans la prémisse de règle considérée est effectuée grâce à la présence de marques de surface : les connecteurs de discours. L'extraction est effectuée sur le corpus de l'Est Républicain, annoté en dépendances syntaxiques par l'analyseur BONSAI (Candito *et al.*, 2009), et exploite un lexique des connecteurs discursifs du français, LEXCONN (Roze *et al.*, 2010), qui contient 330 connecteurs auxquels sont associées une catégorie syntaxique et la relation de discours qu'il établissent. Afin de mieux exploiter le lexique, nous avons complété dans celui-ci un certain nombre de contraintes concernant les positions pouvant être occupées par les différents connecteurs lorsqu'ils lexicalisent une relation donnée. En effet, certains connecteurs n'établissent une relation de discours que lorsqu'ils occupent certaines positions, ou bien peuvent établir des relations différentes selon la position occupée. Pour les adverbes, c'est la position dans la proposition qui peut avoir une importance ; pour les conjonctions de subordination, c'est la position de la proposition subordonnée par rapport à la principale.

[9] Dans la phase d'annotation, nous utilisons les tests d'insertion d'un connecteur et de réorganisation du discours.

Pour extraire des occurrences d'une prémisse de règle $R_x(\alpha, \beta) \wedge R_y(\beta, \gamma)$, on recherche dans le corpus des contextes dans lesquels on a : p_1 *[conn$_x$]*p_2 *[conn$_y$]*p_3, où *conn$_x$* est un connecteur accueilli par la proposition p_2 et lexicalise la relation R_x, et *conn$_y$* est un connecteur accueilli par la proposition p_3 et lexicalise la relation R_y. Soit $conn_y$ se trouve dans la même phrase que $conn_x$, soit dans la phrase suivante. Par exemple, pour la prémisse de règle *Explanation*$(\alpha, \beta) \wedge$ *Result*(β, γ), on extrait des discours comme en (6), où la conjonction *car* marque la présence de la relation *Explanation*, et l'adverbe *alors* marque la présence de la relation *Result*. De la même façon, pour la prémisse de règle *Result*$(\alpha, \beta) \wedge$ *Contrast*(β, γ), on extrait des discours comme en (7).

6. Malgré l'annonce de la fin possible des combats, ils n'ont plus du tout confiance *car*, lors des années passées, ils ont vu la guerre et la paix se succéder. *Alors*, ils se disent que, cette fois encore, la guerre pourrait revenir...
7. Mme Mulot, assistante sociale DVIS, est en absence de longue durée. Ses permanences sont *donc* annulées. La prise en charge des urgences reste *néanmoins* assurée : joindre la circonscription DVIS centre-Vosges...

4 Construction d'une règle de déduction

Dans cette section, nous déroulons pour la prémisse de règle *Result*$(\alpha, \beta) \wedge$ *Contrast*(β, γ) la méthodologie présentée à la section 3. Il nous a paru intéressant d'étudier cette prémisse de règle, et d'observer les conséquences de l'établissement des deux relations relations en jeu, car elles appartiennent à deux groupes de relations distincts : les relations causales (pour *Result*) et les relations adversatives (pour *Contrast*), selon la classification de (Halliday & Hasan, 1976). La première relie deux éventualités dont l'une est la cause de l'autre, et c'est une relation de cohérence, c'est-à-dire une relation liée au contenu sémantique des propositions reliées. La seconde, en revanche, relie des éventualités présentées par le locuteur comme étant en opposition ou en contradiction. Elle est généralement considérée comme une relation intentionnelle, liée aux buts communicatifs.

4.1 Extraction de déductions candidates

L'exploitation du corpus ANNODIS permet d'obtenir les probabilités présentées dans le tableau 2. Les calculs sont effectués à partir des 150 contextes du corpus où l'on a : *Result*(α, β), *Contrast*(β, γ) et une relation $R_z(\alpha, \gamma)$, après réécriture des relations entre segments complexes. Le calcul des probabilités nous permet de dégager la règle candidate : *Result*$(\alpha, \beta) \wedge$ *Contrast*$(\beta, \gamma) \rightarrow$ *Contrast*$(\alpha, \gamma) \vee$ *Result*(α, γ). En effet, on ne retient pas la déduction de *Goal* dans la règle candidate, car la probabilité observée pour cette règle provient d'une seule annotation du corpus, au sein de laquelle les relations annotées ne sont pas cohérentes. On constate que la déduction candidate la plus probable est *Contrast*.

Relation $R_z(\alpha, \gamma)$	$p(R_z(\alpha, \gamma) \mid$ *Result*(α, β), *Contrast*$(\beta, \gamma))$
$R_z =$ *Contrast*	0.847
$R_z =$ *Result*	0.127
$R_z =$ *Goal*	0.027
$R_z \notin \{$*Result*, *Contrast*, *Goal*$\}$	0.0

TAB. 2 – Probabilités $p(R_z(\alpha, \gamma) \mid$ *Result*(α, β), *Contrast*$(\beta, \gamma))$ pour toute relation R_z

4.2 Mise en évidence des déductions possibles

Nous cherchons ici à mettre en évidence les déductions possibles pour la prémisse de règle *Result*$(\alpha, \beta) \wedge$ *Contrast*(β, γ). Dans la section 4.2.1, nous fournissons une définition de ces deux relations, et distinguons deux sous-cas pour la relation *Contrast* : un premier où elle est de type *opposition sémantique*, et un second où elle de type *concession* ou *violation d'attente*. Cette distinction nous amène à présenter l'étude des déductions possibles dans deux sections séparées : 4.2.2 pour le premier sous-cas, et 4.2.3 pour le second.

4.2.1 Définition des relations *Result* et *Contrast*

Au sein de la SDRT, la relation *Result* peut être établie par : des marques linguistiques, comme certains connecteurs de discours (*donc*, *du coup*, *alors*, etc.) et des verbes causatifs (*provoquer*, *entraîner*, etc.) ; des connaissances (extra-)linguistiques, comme *pousser*$(x, y) >$ *tomber*(y)[10] (qui permet par exemple l'interprétation de : *Léa a poussé Max. Il est tombé.*). Elle a pour conséquence sémantique l'établissement d'une relation causale entre les deux éventualités reliées, à savoir que l'éventualité décrite dans le premier argument de la relation est la cause de l'éventualité décrite dans le second argument : *cause*(e_α, e_β) (pour *Result*(α, β)). Cette relation causale recouvre l'implication non monotone $K_\alpha > K_\beta$ (K_α désigne le contenu sémantique du segment α). *Result* possède également des effets temporels : l'éventualité du premier argument a lieu avant l'éventualité du second argument.

La relation *Contrast* relie des segments qui présentent une dissimilarité sémantique. Cette relation recouvre dans la SDRT trois relations qui sont parfois distinguées dans la littérature (Busquets, 2007) : *opposition sémantique* ou *contraste formel*, *concession*, et *violation d'attente*. L'*opposition sémantique* (OS) est définie par Spooren (1989) comme une relation entre deux propositions qui ont deux sujets distincts, auxquels sont attribués des propriétés qui s'excluent mutuellement dans le contexte. Selon Oversteegen (1997), l'opposition sémantique ne nécessite pas la présence de deux entités distinctes : il peut aussi n'y avoir qu'une seule entité, à laquelle différentes propriétés sont assignées, à des localisations temporelles ou spatiales distinctes, où dans différents mondes possibles. Dans la SDRT, cette relation implique une similarité structurelle entre les segments reliés. Busquets (2007) résume la relation d'opposition sémantique comme un contraste ou une dissimilarité entre les éléments comparés (des états, des événements ou des individus) sans contradiction entre eux. En revanche, les éléments impliqués dans une relation de *concession* ou de *violation d'attente* sont en contradiction. Ils s'inscrivent dans le schéma suivant : $(A > C) \wedge (B > \neg C)$. En ce qui concerne la *concession* (CS), les propositions A et B sont explicites (Gröte *et al.*, 1995). Par exemple, dans le discours en (8), on infère de (8a) la proposition *nous avons mangé* $(= C)$, et de (8b) la proposition *nous n'avons pas mangé* $(= \neg C)$. En ce qui concerne la *violation d'attente* (VA), la proposition $\neg C$ est présente. Par exemple, en (9) on infère *Pierre n'aime pas le foot* $(= C)$ de la proposition en (9a) et l'on a $\neg C$ en (9b). On note que les deux éléments en relation de VA peuvent apparaître dans l'ordre inverse.

8. (a) Nous avions faim, $>$ *Nous avons mangé.*
   (b) *mais* les restaurants étaient fermés. $>$ *Nous n'avons pas mangé.*
9. (a) Pierre n'aime pas le sport, $>$ *Pierre n'aime pas le foot.*
   (b) *mais* il aime le foot.

Il nous faut noter que deux segments de discours peuvent à la fois entretenir une relation de *Result* et une relation de *Contrast*. Par exemple, dans le discours en (10), on infère que l'accident de Marie est la cause de ses fractures, ce qui correspond à l'établissement de la relation *Result*. On infère également que l'accident aurait normalement dû causer des blessures plus graves que des fractures, mais qu'il n'en a pas causé. Dans ce discours, l'établissement de la relation *Result* a pour conséquence sémantique $K_{\pi_1} > K_{\pi_2}$, et l'établissement de la relation *Contrast* (de type CS) a pour conséquence sémantique $(K_{\pi_1} > P) \wedge (K_{\pi_2} > \neg P)$ (où $P =$ *Marie a des blessures graves*). Cet exemple nous montre qu'il peut exister des cas dans lesquels on déduira à la fois la relation *Result* et la relation *Contrast*.

10. (a) Marie a eu un accident de voiture, (π_1)
    (b) *mais* elle n'a que quelques fractures. (π_2)

4.2.2 Déduction dans le cas où *Contrast* est de type *opposition sémantique*

Dans le cas où la relation *Contrast*(β, γ) est de type OS il semble que trois déductions soient possibles : soit on déduit *Result*(α, γ) comme dans le discours en (11), où l'on infère que Julie est aux anges *parce que* la France a perdu le match, et où l'éventualité en (α) a des effets opposés sur deux entités distinctes (*Marie* et *Julie*) ; soit on déduit *Contrast*(α, γ) comme dans le discours en (12a), où l'on infère que *malgré* l'accident, Julie n'est pas blessée ; soit on déduit *Result*$(\alpha, \gamma) \wedge$ *Contrast*(α, γ) comme dans le discours en (12b), où, par le contenu de (α) et le contenu de (β), on infère que Julie aurait *normalement* dû être plus gravement blessée, ce qui est contredit par le segment (γ), qui décrit parallèlement un état causé par l'accident décrit en (α).

[10] L'opérateur conditionnel non monotone permet d'exprimer des règles défaisables (ou révisables) (Asher & Lascarides, 2003). Par exemple, $A > B$ signifie : « si A est vrai, alors normalement, B est vrai ».

11. (α) La France a perdu le match.
    (β) *Du coup*, Marie est très déçue.
    (γ) *Par contre*, Julie est aux anges.
12. (α) Marie et Julie ont eu un accident de voiture.
    (β) Marie a une jambe cassée.
    (γ) a. *En revanche*, Julie est indemne.
    b. *En revanche*, Julie n'a que quelques égratignures.

Notons que lorsque la déduction de *Contrast* a lieu, le type de contraste établi entre (α) et (γ) n'est pas le même que celui établi entre (β) et (γ). Par exemple, dans le discours en (13), le contraste entre (β) et (γ) est de type OS car deux entités distinctes (*Julie* et *Léa*) présentent des propriétés opposées. En revanche, entre les segments (α) et (γ), c'est un contraste de type VA qui s'établit. En effet, on a : $K_\alpha > \neg K_\gamma$ (si Marie se fait agresser, alors normalement Léa devrait intervenir). D'ailleurs, si l'on réorganise le discours en (13) en intervertissant les positions de (β) et (γ), on lexicalise la relation entre (α) et (γ) par le connecteur *mais*, qui établit la relation *Contrast* (*Marie s'est fait agresser.* ***Mais*** *Léa n'a pas bougé.* ***Par contre****, Julie a tenté de la défendre.*).

13. (α) Marie s'est fait agresser.
    (β) *Du coup*, Julie a tenté de la défendre.
    (γ) *Par contre*, Léa n'a pas bougé.

4.2.3 Déduction dans le cas où *Contrast* est de type *concession* ou *violation d'attente*

Dans le cas où la relation *Contrast*(β, γ) est de type CS ou VA, il semble que l'on puisse déduire la relation *Contrast*(α, γ) directement à partir des conséquences sémantiques de l'établissement des relations de la prémisse. Notons que la présence de la relation *Result*(α, β) a toujours pour conséquence $K_\alpha > K_\beta$. Si le contraste est de type CS, comme dans le discours en (14), alors on a de plus : $(K_\beta > P)$ et $(K_\gamma > \neg P)$. On a donc $K_\alpha > K_\beta > P$, et l'on peut déduire : $K_\alpha > P$. Comme $K_\gamma > \neg P$, on retrouve les conséquences sémantiques de l'établissement d'une relation de CS entre (α) et (γ).

14. (α) Nous n'avions pas mangé de la journée.
    (β) *Donc* nous avions très faim. $> P$ (*Nous avons mangé.*)
    (γ) *Mais* les restaurants étaient fermés. $> \neg P$ (*Nous n'avons pas mangé.*)

Si le contraste est de type VA, on a deux cas possibles, représentés par les discours en (15) et (16). Dans le premier cas, on a $(K_\beta > P)$ et $(K_\gamma = \neg P)$. On a donc $K_\alpha > K_\beta > P$, et l'on peut déduire : $K_\alpha > P$. Comme $K_\gamma = \neg P$, on retrouve les conséquences de l'établissement d'une relation de VA entre (α) et (γ). Dans le second cas, illustré en (16), on a $(K_\beta = P)$ et $(K_\gamma > \neg P)$. On a $K_\alpha > K_\beta = P$, et l'on peut déduire $K_\alpha > P$. On a donc : $(K_\alpha > P)$ et $(K_\gamma > \neg P)$, ce qui correspond à l'établissement d'une relation de CS entre (α) et (γ).

15. (α) Pierre n'aime pas courir.
    (β) *Du coup* il n'aime pas le sport. $> P$ (*Pierre n'aime pas le foot.*)
    (γ) *Mais* il aime le foot. $(\neg P)$
16. (α) Tous les copains de Pierre jouent au foot.
    (β) *Du coup*, il s'est inscrit avec eux. (P)
    (γ) *Pourtant* il n'aime pas le sport. $> \neg P$ (*Pierre ne s'est pas inscrit au foot.*)

4.3 Résultats de l'annotation

Nous présentons dans cette section les résultats de l'annotation effectuée sur les discours extraits automatiquement en utilisant les connecteurs de discours comme marques des relations recherchées. Parmi les 360 discours analysés, 189 (soit 52.5%) ne contiennent pas la prémisse[11] ou sont difficilement analysables, par manque de contexte (extra-)linguistique. En revanche, 171 discours contiennent bien la prémisse à étudier (soit 47.5%). Nous présentons dans

Relation(s) déduite(s)	Pourcentage	Nombre
Contrast	73.7	126
Result	12.3	21
Unknown	5.8	10
None	5.3	9
Contrast et *Result*	2.9	5
Total	100	171

TAB. 3 – Pourcentages des relations déduites entre (α) et (γ) pour les discours extraits contenant la prémisse *Result*$(\alpha, \beta) \wedge$ *Contrast*(β, γ)

le tableau 3 les résultats obtenus pour ces discours. On regroupe les cas où la relation déduite est ambiguë sous *Unknown*, et les cas où aucune relation n'est déduite à partir de la prémisse sous *None*.

Ces résultats nous montrent que dans une majorité de cas, la relation *Contrast* est déduite (76.6%). Nous proposons donc d'établir une règle défaisable : *Result*$(\alpha, \beta) \wedge$ *Contrast*$(\beta, \gamma) >$ *Contrast*(α, γ). De façon plus générale, l'étude de la prémisse permet de formuler une règle dure : *Result*$(\alpha, \beta) \wedge$ *Contrast*$(\beta, \gamma) \rightarrow$ *Contrast*$(\alpha, \gamma) \vee$ *Result*$(\alpha, \gamma) \vee$*None*(α, γ). Pour que les règles puissent s'inscrire dans une algèbre des relations de discours, il nous faut définir une relation artificielle *None*, exprimant le fait qu'il n'existe aucune relation de discours entre deux segments, qui est exclusive de toutes les autres relations de la déduction. Concernant les cas où l'on déduit *None*(α, γ) pour la prémisse étudiée, nous avons observé que d'autres relations que celles contenues dans la prémisse sont généralement présentes. Par exemple, pour le discours en (17), deux relations temporelles (de recouvrement temporel plus précisément) viennent s'ajouter aux relations de la prémisse étudiée : on a $Background_{forward}(\pi_1, \pi_2)$ et $Background_{forward}(\pi_2, \pi_3)$. Nous formulons l'hypothèse que la présence d'autres relations que celles contenues dans la prémisse peuvent « bloquer » la déduction. Les interactions entre les différentes règles de déduction restent donc à étudier, ainsi que les relations incompatibles.

17. (a) Le généraliste était bien connu dans sa petite localité pour ses problèmes d'alcool. (π_1)
    (b) Entre avril et août 2001 , on lui interdisait *donc* de continuer à exercer. (π_2)
    (c) *Or* la CPAM a continué à recevoir des feuilles de remboursement de patients. (π_3)

Pour raffiner la règle générale formulée, on peut distinguer des sous-cas nous permettant de mieux prédire la déduction, en nous basant sur le type de contraste établi, comme nous l'avons montré dans la section 4.2 : *Result*$(\alpha, \beta) \wedge$ (*Contrast*$_{\text{CS}}(\beta, \gamma) \vee$ *Contrast*$_{\text{VA}}(\beta, \gamma)) \rightarrow$ *Contrast*(α, γ).

5 Conclusion et perspectives

Nous avons proposé une méthodologie pour la construction de règles de déduction de relations de discours, destinées à être intégrées dans une algèbre (complète) de ces relations. La construction d'une telle algèbre a comme principal objectif de permettre une meilleure comparaison des structures dans le cadre de l'évaluation de systèmes d'analyse automatique du discours et dans le cadre de la construction de corpus de référence. Elle peut également aider à la détection d'incohérences dans des structures discursives, ce qui peut servir à améliorer l'annotation discursive manuelle ou automatique. Nous avons présenté l'étude complète d'une prémisse de règle, qui a servi à l'élaboration de la méthodologie proposée. Cette étude nous amène à formuler : des règles dures, établies grâce à des éléments théoriques sur les relations de discours, des données construites et des données attestées ; des règles molles, établies grâce à des probabilités de déduction calculées à partir de l'annotation manuelle de données extraites automatiquement. Pour extraire ces données, nous avons développé un outil qui exploite les connecteurs de discours pour détecter la présence des relations de discours recherchées.

La construction des règles de déduction est en cours, et nous avons pour objectif de dégager des traits linguistiques permettant de prédire la relation déduite lorsqu'une règle donne lieu à la déduction d'une disjonction de relations, en nous basant sur l'étude linguistique des règles et sur l'exploitation des données annotées par des méthodes statistiques. Au fur et à mesure de l'étude des règles, nous allons chercher à établir des généralisations sur les règles,

[11] Certains de ces exemples impliquent les relations *Result* et *Contrast* sans contenir précisément la séquence recherchée : les segments (α), (β) et (γ) ne sont pas consécutifs, et l'on a par exemple la structure *Result*$(\alpha, \beta_1) \wedge R(\beta_1, \beta_2) \wedge$ *Contrast*(β_2, γ).

et tenter de déterminer si le type de relation (coordonnante ou subordonnante) a un impact sur la déduction, et si des relations partageant certaines conséquences sémantiques ont un comportement similaire dans la déduction.

Remerciements

Je remercie Laurence Danlos, Pascal Denis et Philippe Muller pour leurs conseils et relectures.

Références

ALLEN J. (1983). Maintaining knowledge about temporal intervals. In *Communications of the* ACM : ACM Press.

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

BLACK E., ABNEY S., FLICKENGER D., GDANIEC C., GRISHMAN R., HARRISON P., HINDLE D., INGRIA R., JELINEK F., KLAVANS J., LIBERMAN M., MARCUS M., ROUKOS S., SANTORINI B. & STRZALKOWSKI T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Speech and Natural Language : Proceedings of a Workshop Held at Pacific Grove, California*.

BUSQUETS J. (2007). Discourse contrast : Types an tokens. *Language, Representation and Reasoning. Memorial Volume to Isabel Gómez Txurruka*, p. 103–123.

CANDITO M., CRABBÉ B., DENIS P. & GUÉRIN F. (2009). Analyse syntaxique du français : des constituants aux dépendances. In *Proceedings of TALN'09*, Senlis, France.

DANLOS L. (2009). D-STAG : un formalisme d'analyse automatique de discours basé sur les TAG synchrones. *Revue TAL*, **50**, 1–30.

GROSZ B. & SIDNER C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, **12**, 175–204.

GRÖTE B., LENKE N. & STEDE M. (1995). Ma(r)king concessions in English and German. In LEIDEN, Ed., *Proceedings of the Fifth European Workshop on Natural Language Generation*, Netherlands.

HALLIDAY M. A. K. & HASAN R. (1976). *Cohesion in English*. London : Longman.

MANN W. & THOMPSON S. (1988). Rhetorical structure theory : Towards a functional theory of text organization. *Text*, **8**, 243–281.

MARCU D. (1996). Building up rhetorical structure trees. In *Proceedings of 13th National Conference on Artificial Intelligence*, volume 2, p. 1069–1074, Portland, Oregon.

NIVRE J. & SCHOLZ M. (2004). Deterministic dependency parsing of english text. In *COLING 2004*, p. 64–70, Geneva, Switzerland.

OVERSTEEGEN L. E. (1997). On the pragmatic nature of causal and contrastive connectives. *Discourse Processes*, **24**, 51–85.

PÉRY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., DRAOULEC A. L., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., COURET M. V., VIEU L. & WIDLÖCHER A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives (poster). In *Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlis, France.

ROZE C., DANLOS L. & MULLER P. (2010). LEXCONN : a French Lexicon of Discourse Connectives. In *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010)*, Moissac, France.

SETZER A., GAIZAUSKAS R. & HEPPLE M. (2003). Using semantic inferences for temporal annotation comparison. In *Proceedings of the Fourth International Workshop on Inference in Computational Semantics (ICoS-4)*.

SPOOREN W. (1989). *Some aspects of the form and interpretation of global contrastive coherence relations*. PhD thesis, K.U. Nijmegen.

TANNIER X. & MULLER P. (2008). Evaluation metrics for automatic temporal annotation of texts. In *Language Resources and Evaluation Conference* (LREC 2008), Marrakech.

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D. & HIRSCHMAN L. (1995). A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference* (MUC–6), Columbia, Maryland.

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