

Connecteurs et analyses de corpus : de l'analyse manuelle à l'analyse automatisée

Liesbeth Degand et Yves Bestgen

Fonds National de la Recherche Scientifique/ Université catholique de Louvain

1. Introduction

L'objectif premier de cet article est de présenter la complémentarité entre analyses de corpus qualitatives (traditionnellement manuelles) et analyses de corpus quantitatives (traditionnellement semi-automatisées), et de montrer qu'en opérationnalisant de manière adéquate les facteurs linguistiques étudiés, l'on peut en quelque sorte aboutir à une analyse automatisée alliant quantité et qualité. Pour illustrer notre propos, nous nous concentrerons sur la sémantique des connecteurs causaux en néerlandais et en français, et, plus particulièrement sur les facteurs linguistiques influençant le sens et l'usage de ceux-ci. Une manière classique de procéder est d'effectuer une analyse de corpus dont l'objectif premier est de mettre au jour le contexte discursif dans lequel les connecteurs sont utilisés afin d'en dériver les contraintes (syntaxiques et sémantiques) d'emploi. Traditionnellement, on procède en analysant de manière exhaustive des corpus de petites tailles, ou en extrayant aléatoirement de corpus plus vastes un petit échantillon (en général cinquante, parfois cent occurrences). Cette limitation quantitative trouve son origine dans le caractère manuel des analyses effectuées. Du point de vue qualitatif, ces analyses sont précieuses, car elles permettent une description fine d'un grand nombre de traits syntaxiques, sémantiques et/ou pragmatiques. Toutefois, la petite taille des échantillons et le fait que ces analyses soient dépendantes du linguiste qui les a effectuées les rendent difficilement généralisables et répliquables. Pour dépasser ces limitations, nous proposons une méthodologie qui permet de traiter de manière exhaustive et automatisée de vastes corpus contenant des milliers d'occurrences d'un même phénomène linguistique en implémentant les procédures d'analyse afin de les rendre indépendantes de l'analyste. Pour atteindre cet objectif, nous avons besoin (i) d'hypothèses linguistiques à tester, et (ii) d'outils de Traitement Automatique du Langage (TAL).

Une des thèses défendues dans cet article est que la construction des hypothèses linguistiques doit se faire à partir d'études qualitatives empiriques manuelles. Pour illustrer notre propos, nous nous concentrerons sur trois connecteurs causaux pour le français, à savoir, *car*, *parce que* et *puisque*, et quatre connecteurs causaux pour le néerlandais, *doordat* ('à cause du fait que'), *omdat* ('parce que'), *want* ('car'), et *aangezien* ('puisque').

Nous détaillerons la méthodologie suivie pour catégoriser les relations causales et leurs marqueurs linguistiques en termes d'Implication du Locuteur (section 2). Les avantages et les inconvénients méthodologiques seront illustrés (section 2.2). Nous passerons ensuite en revue comment les résultats recueillis sur base empirique peuvent être généralisés et répliqués dans une étude automatisée (section 3). Après une description des outils d'analyse syntaxique (section 3.1) et sémantiques (section 3.2), nous présenterons les résultats obtenus (sections 4 et 5). Nous concluons ce travail avec un nombre de perspectives de travail futur.

2. Analyses manuelles: construction théorique

D'un point de vue méthodologique, nous considérons que les analyses de corpus manuelles permettent de tester une série d'hypothèses élaborées dans le cadre d'une modélisation théorique. Il s'agit de vérifier la validité du modèle théorique face à la réalité des données linguistiques authentiques et de le modifier si nécessaire. Cette démarche est illustrée dans la section suivante.

2.1. Connecteurs et Implication du Locuteur

On admet généralement que les relations de cohérence et leurs marqueurs linguistiques peuvent exprimer des significations à différents "niveaux" du discours. Une manière classique d'en rendre compte est de faire usage de classifications très générales du type sémantique vs. pragmatique (VAN DIJK 1979) ou d'établir une distinction trichotomique entre des relations établies dans le domaine du contenu (exemple 1), des relations établies dans le domaine épistémique (exemple 2) et des relations établies dans le domaine interactionnel (exemple 3) (SWEETSER 1990).

- (1) *John came back because he loves her.*
- (2) *John loves her because he came back.*
- (3) *What are you doing tonight, because there's a good movie on.*

Ce type de catégorisation est néanmoins trop rigide pour rendre compte des propriétés distributives complexes de certains connecteurs qui peuvent prendre des positions intermédiaires par rapport aux catégories générales proposées (DEGAND 1998; PANDER MAAT, SANDERS 1995). Dans ce contexte, DEGAND et PANDER MAAT ont proposé une classification alternative consistant en une conceptualisation scalaire des propriétés des connecteurs en termes d'*Implication du Locuteur* ('Speaker Involvement') (DEGAND, PANDER MAAT 1999, 2003; PANDER MAAT, DEGAND 2001). Brièvement, l'Implication du Locuteur (IdL) fait référence au degré avec lequel le locuteur joue implicitement un rôle actif dans la construction de la relation (causale). Selon cette hypothèse, les différents connecteurs sont ordonnés sur une échelle allant d'une implication minimale du locuteur (relation objective) à une implication maximale (relation subjective). Ainsi, dans le domaine de la causalité, l'échelle comporte en ordre croissant d'IdL les relations suivantes : non-volitionnelle (4), volitionnelle (5), épistémique (6) (déductive (6a), non-causale (6b) et abductive (6c)) et interactionnelle (7) (illocutoire (7a) et discursive (7b)).

- (4) *Blanche-Neige est morte. Elle a croqué une pomme empoisonnée.*
- (5) *Cendrillon s'est enfuie à minuit. Le charme allait se rompre.*
- (6) a) *Blanche-Neige doit mourir. Elle est trop belle.*
- (6) b) *Le prince était très beau. Les princes sont toujours beaux.*
- (6) c) *Pinnocchio mentait. Son nez s'allongeait.*
- (7) a) *Veux-tu que je te lise l'histoire de Bambi ? Il est l'heure d'aller dormir.*
- (7) b) *Les succès de Walt Disney – [car] il faut bien parler de succès – ont conduit à un véritable empire du dessin animé.*

Quatre caractéristiques discursives déterminent le niveau d'IdL d'une relation (PANDER MAAT, DEGAND 2001):

- a) **le degré d'iconicité de la relation causale** : moins la relation est iconique, plus le niveau d'IdL est élevé. Ainsi, l'exemple (8) illustre une relation à IdL assez élevée. Il s'agit d'une relation interactionnelle discursive qui est non-iconique avec la causalité dans le monde réel. Cette relation est adéquatement marquée par le connecteur *puisque*. L'exemple (9), par contre, illustre une relation à niveau d'IdL plus bas, puisque la relation (volitionnelle) est iconique avec la causalité dans le monde réel.

- (8) *Apparemment, le président Mobutu est rentré, vendredi, à Kinshasa. Apparemment, puisque personne (...) ne l'a vu descendre de l'avion dans lequel on l'avait vu embarquer, en matinée, à Nice. (Le soir, 1997)*

(9) *Le président avait embarqué pour Nice, mais il est descendu en évitant toute personne parce qu'il n'avait pas envie de rencontrer la presse.*

b) **le degré d'implication subjective d'un protagoniste conscient** : plus l'implication est subjective, plus le niveau d'IdL est élevé. Ainsi, l'exemple (4) *Blanche-Neige est morte. Elle a croqué une pomme empoisonnée.* exprime une relation ne faisant intervenir aucun protagoniste conscient. En effet, Blanche Neige n'agit pas de manière consciente et délibérée quand elle croque la pomme, elle n'a pas le contrôle de la situation causale. Il s'agit dès lors d'une relation causale à niveau d'IdL plus bas que les relations exprimées en (9) ou (10), par exemple, où les protagonistes, respectivement *le président* et *le locuteur*, sont bel et bien responsable de la situation causale exprimée en agissant de manière délibérée.

c) **la distance par rapport au locuteur et au temps présent** : plus la distance est petite, plus l'IdL est élevée. Ainsi la relation en (10) exprime un niveau d'IdL maximal sur ce trait, car la relation se produit dans le *ici et maintenant* de l'énonciation, ce qui n'est pas le cas pour (11), qui induit un niveau d'IdL plus bas.

(10) *(...) je crois que ça s'appelle en français mais excusez-moi parce que je vais peut-être (...) estropier le mot hein / un goupillon là (Valibel)*

(11) *Il s'excusa parce qu'il pensait que sa prononciation était inadéquate.*

d) **le degré plus ou moins implicite de la réalisation du protagoniste** : plus la réalisation est implicite, plus l'IdL est élevée. Pour cette raison, (12) à un niveau d'IdL plus élevé que (13).

(12) *Ce transfert de souveraineté est génial , parce que je vais fièrement pouvoir dire à l'avenir que je suis une vraie Chinoise. (Le soir, 1997)*

(13) *Je trouve que ce transfert de souveraineté est génial , parce que je vais fièrement pouvoir dire à l'avenir que je suis une vraie Chinoise.*

La place qu'occupe un connecteur sur l'échelle se reflète dans son comportement discursif, ce qui permet de faire des prédictions sur l'emploi des marqueurs (DEGAND 2000, PANDER MAAT, DEGAND 2001). En particulier, la représentation scalaire rend compte du fait que les connecteurs ne sont pas strictement liés à un "domaine" ou "niveau" spécifique de signification (*contenu, épistémique, interactionnel, ..*), mais qu'ils imposent néanmoins des contraintes sur les contextes dans lesquels ils peuvent apparaître, certains contextes étant plus "naturels" que d'autres. Ainsi, l'acceptabilité d'un connecteur peut être mesurée en termes d'IdL : tout connecteur encode un certain niveau d'IdL qui contribue à l'interprétation de son environnement discursif. Lorsque ce niveau est trop bas ou trop élevé pour être combinable avec cet environnement, l'usage du connecteur est inapproprié ou forcera une lecture différente, comme illustré dans l'exemple (14).

(14) - *Tu es vraiment casse-pieds !*

- *Puisque / ?# Parce que je suis si casse-pieds, tu iras à ta soirée sans moi !*

Pour découvrir la nature sémantique d'un connecteur donné ainsi que son interaction avec le discours environnant, nous faisons appel à des analyses systématiques de corpus combinant données distributionnelles et intuitions sémantiques. Il s'agit donc de découvrir comment les connecteurs causaux se comportent par rapport aux quatre caractéristiques discursives déterminant le niveau d'IdL d'une relation (cf. *supra*). Nous ne développerons pas ici en détail les résultats obtenus dans le cadre

de la modélisation théorique (PANDER MAAT,DEGAND 2001; DEGAND,PANDER MAAT 2003) ; nous nous concentrerons plutôt sur les aspects méthodologiques de la démarche poursuivie.

2.2 Analyse manuelle de l'usage des connecteurs causaux en contexte

Afin de découvrir le profil sémantique des connecteurs causaux en termes d'Implication du Locuteur, un échantillon de données a été constitué. Au moyen d'un programme d'extraction automatique (dans notre cas, WinGrep3.01), 50 occurrences de chacun des connecteurs sont extraites aléatoirement d'un vaste corpus¹. On obtient ainsi des fragments constitués du connecteur avec au moins deux phrases avant et deux phrases après celui-ci. Deux juges au minimum analysent chacune des occurrences en contexte. L'analyse manuelle consiste à coder chaque fragment selon une série de traits internes à l'analyse (nom du codeur, numéro du fragment, date de codage), ainsi qu'une série de traits linguistiques indispensables pour tester nos hypothèses quant à la position de chacun des connecteurs sur l'échelle d'IdL. Plus précisément, nous avons codé par exemple,

- le type de relation exprimé par le connecteur (non-volitionnelle, volitionnelle, épistémique, interactionnelle, ...);
- la modalité exprimée dans le segment précédant le connecteur (S1) et le segment suivant le connecteur (S2);
- la présence d'un protagoniste conscient;
- l'expression linguistique du protagoniste;
- la continuité du protagoniste entre S1 et S2;
- le temps verbal;
- la position du connecteur;
- ...

Chaque fragment est donc passé au crible d'une vingtaine de traits sémantiques et syntaxiques, comme partiellement illustré en (15), où « 7 » identifie le corpus (Le Soir 1997), « 14 » le connecteur *parce que*, « 08 » le numéro du fragment, « 5 » et « 2 » respectivement la modalité de S1 (action) et de S2 (opinion), « 9 » le type de relation (volitionnel), etc.

(15) (...) *On a pardonné à certains d'avoir collaboré **parce qu'il ne fallait pas affaiblir le camp anticomuniste.*** (...) (Le Soir 1997)

7- 14 -08 - 5 - 2 - 9 - 1 - 3 - 6 - 3 - 1 - 11 - 11 - 1 - 6 - 3 - 3 - 1 - 2

Les difficultés principales de ce type d'analyse (manuelle) sont de deux ordres. Elles concernent (i) le choix des catégories et des traits sémantiques, et (ii) la fiabilité des analyses. La réponse à la première difficulté dépend étroitement du modèle théorique et des hypothèses étudiés. En outre, il faut veiller à ne pas multiplier inutilement le nombre de traits sémantiques à l'intérieur d'une catégorie, vu le nombre relativement restreint d'occurrences analysées. La deuxième difficulté concerne les désaccords entre juges sur le codage des extraits, désaccords qui sont en général le résultat d'une interprétation différente de ceux-ci. Ainsi, pour le codage du type de relation causale exprimé dans l'exemple (16), l'un des juges pourrait estimer qu'il s'agit d'une relation volitionnelle (le *je* explique pourquoi il accomplit l'action de sortir ces dossiers cachés) et l'autre pourrait estimer qu'il s'agit en premier lieu d'une relation épistémique, car subordonnée à un *si* hypothétique. En (17), la détermination de la modalité du premier segment peut également être une source de confusion. S'agit-il d'un fait, d'une expérience ou d'une action?

¹ Il peut s'agir de corpus de presse écrite (Pander Maat & Degand, 2001; Degand & Pander Maat, 1999, 2003), de corpus oraux (Degand et al., 2002; Simon & Degand, soumis.), de corpus de traduction (Degand 2004, sous presse), ou de corpus d'apprenants (Perrez & Degand, en prép.).

- (16) *Si j' accepte aujourd'hui de sortir ces dossiers cachés au fond des tiroirs , c' est parce que le formidable travail qui a été fait ne peut rester oublié.* (Le Soir 1997)
- (17) *Scène 2 : suite à un accident, la voiture de Madame P. est déclarée en perte totale car le coût de la réparation dépasse la valeur intrinsèque du véhicule.* (Le Soir 1997)

Pour répondre à ce genre de difficultés, nous avons mis un soin particulier à opérationnaliser le codage au maximum en explicitant le processus interprétatif. Ainsi, l'ensemble des variables codées (type de relation, modalité du segment 1, modalité du segment 2, identité du protagoniste en S1, identité du protagoniste en S2, etc.) est repris dans un *cahier de codage* décrivant le plus précisément possible les critères de sélection de chacune des valeurs catégoriales. Malgré cette opérationnalisation, les problèmes de désaccords inter-juges subsistent. Face à cela, deux positions sont envisageables : (i) les données non fiables doivent être rejetées et on ne peut dès lors rien en conclure ; (ii) les phénomènes codés étant subjectifs par nature, il est possible que nous n'atteignons jamais les taux d'accord nécessaires². Néanmoins, le fait même que nous considérons que ces phénomènes subjectifs valent la peine d'être étudiés montre que nous sommes en quelque sorte prêts à nous baser sur des données imparfaites. Nous pensons que cela est méthodologiquement acceptable pour autant que nous restions conscients des limites d'un schéma de codage qui nous procure des données imparfaites en termes de généralisation (voir aussi, DEGAND, SPOOREN, BESTGEN 2004). Dans le cadre des analyses manuelles, l'option consiste en général à discuter des cas de désaccords jusqu'à ce qu'un consensus soit trouvé entre les codeurs. Néanmoins, afin de réaliser une avancée substantielle dans ce domaine, nous pensons qu'il faut continuer à développer des techniques de codage qui soient moins dépendantes de l'analyste. L'automatisation et l'implémentation des procédures de codage nous semblent être un pas dans cette direction (voir aussi, DEGAND, SPOOREN, BESTGEN 2004). Nous développerons cet aspect dans la section 3.

Sur la base de l'analyse des analyses manuelles décrites, nous avons pu vérifier les hypothèses suivantes (PANDER MAAT, DEGAND 2001; DEGAND, PANDER MAAT 2003):

- Le potentiel expressif de chacun des connecteurs causaux peut être représenté comme une zone continue sur l'échelle d'Implication du Locuteur.
- Les connecteurs les plus fréquents divergent significativement en termes de zones d'occupation sur l'échelle.
- L'échelle est constante pour des langues différentes, les connecteurs peuvent diverger par les zones qu'ils occupent.

Ainsi, nous avons pu établir que les connecteurs étudiés s'échelonnent comme suit en termes d'Implication du Locuteur:

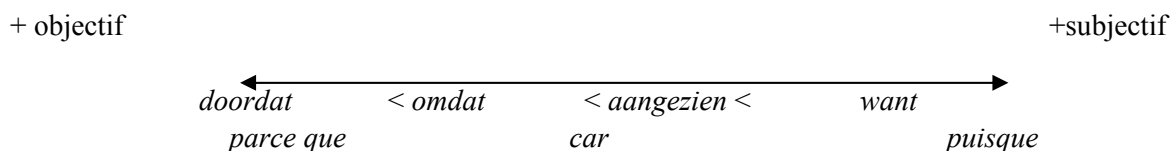


Figure 1: Echellonnement des connecteurs causaux en termes d'IdL

En néerlandais, *doordat* serait le connecteur le plus objectif, *want* le plus subjectif, *omdat* et *aangezien* occupant des zones intermédiaires. En français, *parce que* serait le plus objectif, suivi de *car* et *puisque* en termes de subjectivité. Nous avons par ailleurs pu montrer que le modèle de l'IdL peut rendre compte de la variété d'usage des connecteurs causaux et des effets de substitution d'un connecteur par un connecteur à niveau d'IdL différent, qu'il peut mettre au jour des divergences très fines entre connecteurs, et qu'il permet de contraster des équivalents dans des langues différentes (

² Dans les analyses de corpus on estime en général que l'accord inter-juges est satisfaisant à 80%.

In: Sylvie Porhiel, Dominique Klingler (éds.) (2004) : *L'Unité texte*. Pleyben, pp. 49-73.

DEGAND 2004, sous presse; DEGAND,PANDER MAAT 1999, 2003; PANDER MAAT, DEGAND 2001).

3. Vers une analyse automatisée : validation méthodologique et généralisation des résultats

Nous avons soulevé plus haut les problèmes inhérents à toute analyse manuelle, à savoir, d'une part, la restriction quantitative, et d'autre part, le manque de généralisation et de réplication dues à la nature sémantique des données analysées. Nous présentons ici une méthode automatisée que nous avons développée (BESTGEN, DEGAND,SPOOREN 2003, soumis; DEGAND, SPOOREN, BESTGEN 2004) afin d'analyser de manière automatique plusieurs milliers d'occurrences des connecteurs étudiés. Afin de valider la procédure automatique, nous avons choisi de vérifier plusieurs hypothèses linguistiques concernant la sémantique des connecteurs causaux en néerlandais et en français. Nous en présenterons deux ici qui font intervenir deux types d'analyse différents : l'échelonnement des connecteurs en termes d'IdL (cf. *supra*) que l'on étudiera en faisant appel à des techniques d'analyse de contenu (section 3.2.1.), et l'emploi des connecteurs comme marqueurs de rupture de la perspective discursive (section 5), pour lesquels on utilisera l'analyse sémantique latente (3.2.2.).

La procédure automatique repose sur une série de techniques issues du traitement automatique du langage (TAL). Elles doivent nous permettre d'une part d'identifier et d'extraire le matériel linguistique pertinent, et d'autre part, d'analyser ce matériel linguistique en fonction des hypothèses.

3.1. Identification et extraction du matériel linguistique pertinent

Pour extraire les connecteurs et leur contexte d'utilisation, nous avons employé des techniques classiques en traitement automatique des langues naturelles telles que des étiqueteurs morpho-syntaxiques qui ont également permis de segmenter les textes en phrases. Dans un deuxième temps, nous avons localisé les connecteurs et extrait les phrases qui les entourent. Ensuite, ces phrases ont été analysées au moyen d'une série de règles syntaxiques afin d'identifier le segment exprimant la cause et celui exprimant la conséquence.

Pour le néerlandais, le corpus dont ont été extraits les connecteurs et leur contexte était constitué de 6 mois du journal hollandais *De Volkskrant* 1997 un quotidien de grande diffusion et à visée généraliste. De ce corpus de départ, nous avons supprimé les articles souvent très courts dont le contenu sémantique et syntaxique est peu pertinent pour notre étude comme les mots croisés, les programmes de télévision, les cotations boursières, les résultats sportifs... Ce corpus a été lemmatisé et étiqueté par les logiciels MBLEM et MBT (VAN DEN BOSCH & DAELEMANS, 1999; DAELEMANS, ZAVREL, BERCK & GILLIS, 1996). Les erreurs d'étiquetages provoquées par des incompatibilités entre le fichier source et les programmes d'étiquetage ont été supprimées. Dans la plupart des cas, il s'agissait de non-mots introduits par le lemmatiseur. Après toutes ces étapes, le corpus était composé d'approximativement 16.5 millions d'éléments incluant bien sûr les mots, mais aussi les signes de ponctuation, les chiffres et les caractères spéciaux.

La même méthodologie a été employée lors de la construction du matériel pour le français. Les articles ont été extraits de six mois du journal *Le Soir* 1997, un quotidien belge francophone à grande diffusion et à visée généraliste. Les articles dont le contenu sémantique et syntaxique est peu pertinent pour notre étude ont également été éliminés. Ce corpus a été lemmatisé et étiqueté par le logiciel TreeTagger (SCHMIDT, 1994). Après ces étapes, le corpus était composé d'approximativement 13.5 millions d'éléments.

Après avoir extrait les connecteurs et les phrases qui les entourent, il a fallu élaborer une série de règles heuristiques à base syntaxique afin d'identifier le segment exprimant la cause (P) et celui exprimant la conséquence (Q). Pour le néerlandais nous avons abouti à l'élaboration de 21 règles

heuristiques basées principalement sur (i) la position du connecteur dans la phrase (prise en compte du nombre et types de mots précédant le connecteur), (ii) le nombre, la position et l'ordre des verbes conjugués dans le segment, (iii) la présence ou l'absence de marqueurs de ponctuation, en particulier les virgules. Un exemple illustratif est donné en (19).

(19)

- a) Si *CONN* = *omdat*, *doordat* ou *aangezien*; et
- b) si *CONN* est en position initiale, recherche du premier verbe conjugué (*Vc*), si *Vc* apparaît dans segment <... *vc*, *vc* ...> ou <... *vc vc* ...> alors couper avant second *Vc*, et le segment contenant *CONN* est *P*, l'autre segment est *Q*.
- c) Si *CONN* est en position initiale et il y a un seul *Vc*, alors segment contenant *CONN* est *P*, et phrase précédente est *Q*.

L'adéquation des règles a été vérifiée manuellement sur de larges échantillons de données et a donné lieu finalement à la perte de 1,4% des données parce qu'un des segments (*P* ou *Q*) était manquant, ou parce qu'aucune des procédures n'aboutissait à l'identification de *P* et *Q*. En fin de compte, les règles heuristiques ont permis d'identifier les segments de cause et de conséquence pour 14181 phrases : 246 pour *aangezien*, 815 pour *doordat*, 7531 pour *omdat*, et 5589 pour *want* (voir BESTGEN, DEGAND, & SPOOREN, 2003).

Pour l'identification des segments de cause et de conséquence dans le corpus francophone le travail est toujours en cours. Nous exposons ci-dessous les difficultés majeures auxquelles nous sommes confrontés dans l'établissement des règles heuristiques.

Les connecteurs *car*, *parce que* et *puisque* ont des distributions syntaxiques divergentes dont il faut tenir compte dans la mise au point des règles permettant l'identification des segments de cause et de conséquence dans les extraits.

Car occupe toujours une position médiane. Il est toujours précédé du segment de conséquence (*Q*) et suivi du segment de cause (*P*). Travaillant avec des extraits découpés en phrases, nous avons estimé que le segment *P* commence toujours avec le connecteur et se termine à la ponctuation finale de la phrase. Reste alors à définir la taille du segment *Q*. Si le connecteur se trouve en position initiale, le segment *Q* est composé de la phrase précédente (20) (construction médiane externe), sinon le segment *Q* commence avec le début de la phrase et se termine avant le connecteur (21) (construction médiane interne).

- (20) [*Q*] *Le paradoxe veut que ce scénario, quasiment immuable depuis le début de saison, n'empêche pas Valentino Rossi de posséder une belle marge de sécurité au championnat. [P] Car s'il n'a pas souvent les moyens de fausser compagnie à ses adversaires, le jeune Italien n'a pas son pareil pour emballer le sprint final. (Le Soir 1997)*
- (21) [*Q*] *Il faut, selon lui, travailler dans les institutions, [P] car c'est le contexte qui permet à une population de vivre. (Le Soir 1997)*

Pour *puisque* et pour *parce que* les règles seront différentes, car la distribution syntaxique de ces connecteurs est divergente. Ainsi il faut pouvoir identifier correctement les segments de cause et de conséquence dans des constructions antéposées (22-23), médianes internes (24-25) et externes (26-27), et enfin des constructions focalisées (pour *parce que* uniquement, voir l'exemple 28).

- (22) [*P*] *Et puisque la Villa a toujours été et reste une maison de luxe, [Q] cela ne vaut pas la peine de regretter des prix élevés. (Le Soir 1997)*

- (23) [P] **Parce que** les travaux ont été réalisés sans autorisation , [Q] ils demandent la remise en état des lieux d'ailleurs situés en zone agricole et forestière et l' interdiction des expositions. (*Le Soir* 1997)
- (24) [Q] La circulation ne sera pas trop gênée normalement , [P] **puisque** les travaux se concentrent sur la zone des parkings. (*Le Soir* 1997)
- (25) [Q] Pour un coach , c'est vraiment l' homme idéal [P] **parce qu'** il arrive à mettre en pratique les consignes que je donne. (*Le Soir* 1997)
- (26) [Q] Aujourd'hui, donc, on ne se cache plus pour acheter des produits aux emballages « blancs » , sans autre marque que celle de la grande surface où on les distribue. [P] **Puisque** , la crise aidant , on a pu apprécier depuis longtemps la qualité de ces produits. (*Le Soir* 1997)
- (27) Il distille un grand bonheur. [P] **Parce qu'** il restitue justement une tranche de vie. (*Le Soir* 1997)
- (28) [P] Ce n' est pas **parce qu'** on se connaît, [Q] qu' on obéit systématiquement. (*Le Soir* 1997)

Si la démarche pour établir les règles d'identification est fondamentalement la même pour le français que pour le néerlandais, la liberté syntaxique plus grande en français (principalement dans l'ordre des constituants, l'emploi d'incises, et le caractère non obligatoire de la ponctuation) complique la définition des règles heuristiques. D'autres problèmes concernent la hiérarchisation des règles par défaut et des règles « prioritaires », la détermination des frontières des segments et la taille minimale des segments pour une analyse sémantique. Vu la complexité de la tâche, seule une partie des extraits ont pu être identifiés en termes de segments de cause et de conséquence. Ces résultats provisoires sont présentés dans le tableau 1 :

Tableau 1 : Constructions causales pour le français (résultats provisoires)

Connecteur	Médian externe	Médian interne	Antéposé	Total
Car	1898	2801	0	4699
Parce que	123	2397	173	2693
Puisque	3	2704	117	2824

Au départ, 4803 occurrences de *car*, 3912 occurrences de *parce que* et 3420 occurrences de *puisque* avaient été extraites du corpus. Si l'on compare ces données au Tableau 1, force est de constater que si la quasi totalité des occurrences de *car* sont couvertes par les analyses (97,8%), ceci n'est pas le cas pour *puisque* (82,6%) et encore moins pour *parce que* (68,8%). Afin d'améliorer ces scores, nous travaillons pour le moment sur d'autres techniques permettant d'identifier la construction causale exprimée par les connecteurs, en particulier des techniques faisant intervenir une analyse syntaxique automatisée (BOURIGAULT & FABRE, 2000).

3.2. Techniques d'extraction d'information sémantique

Comme indiqué plus haut, nous voulons illustrer la procédure automatisée sur base de deux ensembles d'hypothèses. Les premières portent sur le fait que certains segments contiennent plus que d'autres segments des mots appartenant à une catégorie sémantique ou grammaticale spécifique. Pour tester celles-ci, nous avons eu recours à une méthodologie classique tant en analyse de corpus par ordinateur (BIBER, 1988) qu'en analyse de contenu (STONE, 1997) : l'identification automatique de catégories lexicales. Les secondes portent sur la proximité sémantique entre deux segments de textes. Pour mesurer cette proximité sémantique, nous avons employé l'analyse sémantique latente (ASL), une technique à la base d'un nombre de plus en plus important de recherches en psycholinguistique (LANDAUER, FOLTZ, & LAHAM, 1998).

3.2.1. Analyse de contenu thématique

Afin de tester notre premier groupe d'hypothèses, nous avons employé une technique dénommée "identification automatique de traits linguistiques" (BIBER, 1988) ou analyse de contenu thématique (POPPING, 2000; STONE, 1997), dont le but est de déterminer si certaines catégories de mots (mots exprimant une opinion, un fait) ou certaines catégories grammaticales (pronoms personnels) sont plus fréquentes dans certains types de segments par exemple ceux qui sont introduits par un connecteur donné.

La première étape de ce genre d'analyse consiste en la construction d'un dictionnaire contenant des listes de mots qui relèvent des différentes catégories à étudier. Ces catégories peuvent correspondre à des classes grammaticales, mais aussi à des regroupements thématiques de mots. Ensuite, ces listes de mots sont comparées aux mots présents dans chaque segment de textes à analyser afin de déterminer la fréquence de chaque catégorie dans chacun de ceux-ci. Ces données sont rassemblées dans une matrice dont les lignes correspondent aux segments et les colonnes aux différentes catégories. Finalement, cette matrice est analysée pour déterminer si certaines catégories sont plus fréquentes dans certains types de segments de textes.

En analyse du contenu assistée par ordinateur, cette technique est fréquemment employée pour inférer des caractéristiques de l'auteur d'un texte sur la base des thèmes qu'il aborde dans son discours, mais aussi de la façon dont il s'exprime (POPPING, 2000; PENNEBAKER, MEHL, & NIEDERHOFFER 2003). In analyse linguistique de corpus, un domaine plus proche de notre recherche, cette technique a été principalement popularisée par BIBER (1988; CONRAD & BIBER, 2001) qui l'a employée pour réaliser son étude à grande échelle des traits linguistiques qui distinguent les genres et les registres en anglais. Les traits linguistiques pris en compte étaient très nombreux allant de caractéristiques linguistiques comme le passif ou la nominalisation à des catégories de verbes comme les verbes publics (prétendre, dire, ...) en passant par des classes fermées comme les pronoms ou les conjonctions. Il a montré que genres et registres se distinguaient quant à la fréquence de ce genre de traits.

Les deux difficultés majeures auxquelles cette technique fait face dans notre cas sont la très petite taille des segments de textes analysés (une phrase ou même moins) et la difficulté, voir l'impossibilité, de constituer une liste exhaustive des mots qui appartiennent à une catégorie aussi générale que les mots d'opinion. Pour ce qui est de la première difficulté, nous pensons que la petitesse des segments sera compensée par le grand nombre de segments de chaque type qui sera analysé. Pour la deuxième difficulté, nous pensons qu'il devrait être possible d'étendre automatiquement les listes de mots, par exemple en calculant les plus proches voisins (BESTGEN, 2002) des entrées lexicales et de les ajouter au dictionnaires s'ils sont adéquats afin d'élargir le dictionnaire (DEGAND, SPOOREN & BESTGEN, 2004).

3.2.2. L'analyse sémantique latente (Latent Semantic Analysis)

L'analyse sémantique latente (ASL) est une technique mathématique qui vise à extraire un espace sémantique de très grande dimension à partir de l'analyse statistique de l'ensemble des cooccurrences dans un corpus de textes. Comme le souligne LANDAUER *et al.* (1998), cette technique peut être vue de deux manières. A un niveau théorique, elle peut servir de base pour développer des simulations des processus psycholinguistiques à l'œuvre lors de la compréhension du langage (LANDAUER *et al.*, 1997), incluant, par exemple, un "modèle computationnel" du traitement des métaphores (KINTSCH, 2000 ; LEMAIRE, BIANCO, SYLVESTRE, & NOVECK, 2001; BESTGEN & CABIAUX, 2002), mais aussi l'analyse de la cohérence dans des textes (FOLTZ, KINTSCH, & LANDAUER, 1998 ; BESTGEN, 2004). A un niveau plus appliqué, c'est une technique permettant d'inférer et de représenter le sens de mots sur la base de leur usage dans des textes afin de pouvoir estimer les similarités sémantiques entre des mots, des phrases ou des paragraphes

(BESTGEN, 2002; CHOI, WIEMER-HASTINGS, & MOORE, 2001). C'est ce second emploi qui nous intéresse ici.

Le point de départ de l'analyse est un tableau lexical (LEBART & SALEM, 1992) qui contient le nombre d'occurrences de chaque mot dans chacun des documents, un document pouvant être un texte, un paragraphe ou même une phrase. Pour dériver d'un tableau lexical les relations sémantiques entre les mots, la simple analyse des cooccurrences brutes se heurte à un problème majeur. Même dans un grand corpus de textes, la plus grande partie des mots sont relativement rares. Il s'ensuit que les cooccurrences le sont encore plus. Leur rareté les rend particulièrement sensibles à des variations aléatoires (BURGESS, LIVESAY, & LUND, 1998 ; KINTSCH, 2001). L'ASL résout ce problème en remplaçant le tableau de fréquences original par une approximation qui produit une sorte de lissage des associations. Dans ce but, le tableau de fréquence fait l'objet d'une décomposition en valeurs singulières avant d'être recomposé à partir d'une fraction seulement de l'information qu'il contient. Les milliers de mots caractérisant les documents sont ainsi remplacés par des combinaisons linéaires ou 'dimensions sémantiques' sur lesquelles peuvent être situés les mots originaux. Contrairement à une analyse factorielle classique, les dimensions extraites sont très nombreuses (plusieurs centaines) et non interprétables. Elles peuvent toutefois être vues comme analogues aux traits sémantiques fréquemment postulés pour décrire le sens des mots (LANDAUER et al., 1998).

Tant les mots que les segments originaux sont positionnés dans cet espace sémantique, ce qui permet de mesurer leur proximité. Plus précisément, le sens de chaque mot y est représenté par un vecteur. Pour mesurer la similarité sémantique entre deux mots, on calcule le cosinus entre les vecteurs qui les représentent. Plus deux mots sont sémantiquement proches, plus les deux vecteurs qui les représentent pointent dans la même direction et donc plus leur cosinus se rapproche de 1. Un cosinus de 0 indique une absence de similarité puisque les vecteurs correspondants sont orthogonaux. De la même manière, en calculant le cosinus entre les vecteurs qui représentent deux segments de textes, on peut déterminer la proximité sémantique entre ceux-ci.

Le plus important toutefois est que cette technique est encore plus générale puisqu'elle permet de calculer le vecteur qui correspond à une phrase ou un paragraphe même si cette phrase ou ce paragraphe ne constitue pas un document analysé en tant que tel. Ce vecteur correspond au centroïde des mots qui composent l'unité en question, obtenu en calculant la somme pondérée des vecteurs représentant les mots qui composent cette unité. Cela permet de déterminer la proximité sémantique entre deux phrases que celles-ci fassent partie du corpus de départ ou non, que le corpus de départ ait été segmenté en documents correspondant à des phrases ou non. Nous emploierons cette technique pour évaluer la proximité sémantique entre P et Q et entre ces deux segments et la phrase qui les précède ou qui les suit.

Pour construire l'espace sémantique pour l'analyse des extraits néerlandais, nous avons employé le corpus de textes lemmatisé duquel les connecteurs ont été extraits (cf. section 3.1). Avant de construire le tableau lexical, une série de prétraitements ont été effectués. Tout d'abord, les chiffres, caractères spéciaux et les signes de ponctuation ont été supprimés ainsi que tous les mots appartenant à une liste de formes fonctionnelles (pronoms, articles, auxiliaires, ...), réduisant le nombre total de mots à plus ou moins 6.5 millions de mots. Ce corpus a été segmenté en fonction des articles. La taille de ceux-ci variant fortement, nous avons éliminé les articles très courts (moins de 24 mots) et les articles très longs (plus de 523 mots). Ces limites ont été définies sur la base d'une analyse de la distribution de fréquence des longueurs d'articles afin d'éliminer les 10% d'articles les plus courts et les 10% les plus longs. Tous les mots dont la fréquence dans le corpus était inférieure à 10 ont été également supprimés. Après cette dernière étape, il restait 36630 mots différents et 28640 documents, ces nombres constituant les dimensions du tableau lexical décomposé en valeurs singulières par le programme SVDPACKC (BERRY, 1992 ; BERRY, DO, O'BRIEN, KRISHNA, & VARADHAN, 1993). Les 300 premiers vecteurs propres ont été conservés.

Pour le français, nous avons également utilisé le corpus de presse écrite pour construire l'espace sémantique. Comme pour l'analyse du matériel néerlandophone, les éléments non pertinents (chiffres, signes de ponctuation, mots fonctionnels, ...) ont été éliminés, laissant un total de 5 millions de mots. Ce corpus a été segmenté en articles dont les 10% les plus courts (moins de 114 mots) ainsi que les 10% les plus longs (plus de 856 mots) ont été supprimés. Après suppression des mots de fréquence inférieure à 10, le tableau lexical, 16304 mots différents et 18735 documents, a été décomposé en valeurs singulières et les 300 premiers vecteurs propres ont été conservés.

4. Analyse automatique du niveau d'Implication du Locuteur

Sur base des analyses manuelles présentées dans la section 2, nous faisons l'hypothèse que les connecteurs diffèrent les uns des autres par le niveau d'IdL qu'ils encodent. En néerlandais, *doordat* encoderait un niveau d'IdL bas (non-volitif, objectif, factuel), *want* un niveau d'IdL élevé (épistémique-interactionnel, subjectif, opinion-argument), alors que *omdat* et *aangezien* occuperaient une position intermédiaire (volitionnelle, épistémique). En français, *parce que* exprimerait des relations plutôt objectives, à tendance subjective (volitionnelles, parfois épistémiques), *car* des relations plutôt subjectives (épistémiques, parfois volitionnelles), et *puisque* principalement des relations subjectives (épistémiques).

Pour tester ces hypothèses, nous proposons de faire appel à l'analyse de contenu thématique (Section 3.2.1). A cette fin, nous avons construit un « dictionnaire de subjectivité » contenant trois catégories: une liste de mots exprimant la *factualité*, une liste exprimant l'*action* et une liste relevant de l'*opinion*. La catégorie *opinion* comprend, par exemple, des verbes, adverbes et adjectifs exprimant une opinion telles que *estimer*, *croire*, *probablement*, *particulièrement*, *magnifique*, ... Pour élaborer ces listes de mots (entre cinquante et cent cinquante par catégorie) nous nous sommes basés sur des thésaurus existants (BROUWERS, 1997 pour le néerlandais; et un atlas sémantique en ligne (http://dico.isc.cnrs.fr/dico_html/index.html) pour le français). Nous en avons extraits tous les lemmes (non-ambigus) correspondant à une des catégories sus-mentionnées. Nous avons alors confrontés ces listes aux segments de conséquence (Q) pour vérifier si ceux-ci exprimaient une information plutôt objective (*fait*, *action*) ou subjective (*opinion*). En ce qui concerne la relation entre la nature sémantique du segment de conséquence et les connecteurs étudiés, les hypothèses étaient les suivantes:

- En néerlandais, les segments de <conséquence> liés par *doordat* contiennent des mots factuels, ceux liés par *omdat* contiennent des mots d'action et d'opinion, et ceux liés par *aangezien* et *want* contiennent des mots d'opinion.
- En français, les segments de <conséquence> liés par *parce que* contiennent des mots d'action et d'opinion, ceux liés par *car* et *puisque* contiennent des mots d'opinion.

Notons qu'en français, nous n'avons pas d'hypothèse forte pour les mots factuels, car, contrairement au néerlandais, le français n'a pas de connecteur "spécialisé" dans le domaine non-volitionnel. Nous pensons plutôt que la catégorie factuelle va se répartir sur les trois connecteurs induisant une lecture différente en fonction du connecteur (plus objectif en co-occurrence avec *parce que*, plus subjectif en co-occurrence avec *car* et *puisque*).

Pour le néerlandais, l'analyse automatique confirme les analyses de corpus manuelles. Une analyse de variance (ANOVA) avec les connecteurs comme variable indépendantes et la présence ou l'absence dans le segment Q d'une des entrées lexicales du dictionnaire de subjectivité comme variable dépendante indique que les quatre connecteurs divergent significativement les uns des autres que ce soit pour la catégorie opinion ($F(3, 14177) = 28.10, p < .0001$), action ($F(3, 14177) = 15.96, p < .0001$), ou fait ($F(3, 14177) = 29.24; p < .0001$). *Doordat* co-occure significativement plus avec des

segments factuels que les autres connecteurs, *omdat* plus avec des segments d'action et *want* et *aangezien* plus avec des segments d'opinion (cf. BESTGEN, DEGAND, SPOOREN 2003, soumis).

Pour le français, les résultats doivent être pris avec quelques réserves puisque les données ne sont pas complètes. L'analyse automatique confirme en partie les analyses de corpus manuelles: les trois connecteurs divergent significativement les uns des autres lorsque le segment de conséquence exprime une opinion ($F(2, 10215) = 9.91, p < .0001$), une action ($F(2, 10215) = 12.68, p < .0001$) ou un fait ($F(2, 10215) = 4.57, p < .01$). *Car* co-occure significativement plus avec des segments d'opinion et factuels que les autres connecteurs, alors que *parce que* et *puisque* ne divergent pas significativement l'un de l'autre, même si *parce que* a tendance à survenir plus avec des segments d'action et *puisque* plus avec des segments factuels. Contrairement à nos attentes, *puisque* ne semble pas survenir significativement plus dans des contextes subjectifs. Une analyse qualitative devrait nous apprendre si la présence de ce connecteur dans des contextes factuels induit de lui-même une lecture subjective. Notons également que les données analysées ne sont pas complètes pour *puisque* et *parce que*. Ces résultats devraient donc être confirmées sur des données plus complètes.

Un second trait d'IdL que nous avons vérifié au moyen d'une analyse de contenu thématique est la présence ou l'absence de pronoms personnels. En effet, les pronoms personnels font référence à un protagoniste conscient dans l'événement causal, et de ce fait on peut les considérer comme des marqueurs linguistiques de subjectivité (DEGAND & PANDER MAAT, 2003 ; PIT, 2003). Les hypothèses (pour le néerlandais et le français) sont les suivantes:

- Les connecteurs subjectifs (à IdL élevé) devraient survenir plus avec des pronoms personnels que les connecteurs objectifs (à IdL basse).
- Les connecteurs subjectifs devraient être plus fréquents avec des pronoms personnels à la première personne et les connecteurs objectifs plus fréquents avec des pronoms à la troisième personne.

Nous avons réalisé une nouvelle analyse de contenu thématique sur base d'un dictionnaire des pronoms personnels nominatifs. Les résultats d'une analyse ANOVA avec les connecteurs comme variable indépendante et la présence ou l'absence des pronoms personnels (nominatifs) dans le segment de conséquence comme variable dépendante montre que l'hypothèse se vérifie pour le néerlandais ($F(3, 14177) = 80.58; p < 0.0001$): les segments avec *doordat* contiennent le moins de pronoms personnels, suivi de *aangezien*, *omdat* et *want*. En ce qui concerne la seconde hypothèse, elle n'est que partiellement confirmée. En effet, **tous** les connecteurs sont plus fréquents avec des pronoms à la troisième personne qu'avec des pronoms personnels à la première personne. Néanmoins, la proportion des segments avec *want* contenant un pronom à la première personne est plus élevée (37,9%), ce qui tend à confirmer que *want* est le connecteur le plus subjectif et que les autres connecteurs se retrouvent dans la partie plus objective de l'échelle (cf. BESTGEN, DEGAND & SPOOREN, 2003).

Pour le français, aucune des deux hypothèses ne se confirment, puisque c'est le connecteur *parce que*, censé le plus objectif, qui survient significativement plus avec des pronoms personnels, suivi de *car* et de *puisque* ($F(2, 10215) = 83.05 ; p < 0.0001$). Et si, comme en néerlandais, tous les connecteurs co-occurrent plus avec des pronoms à la troisième personne, c'est à nouveau *parce que* qui forme la proportion la plus importante de segments à la première personne (21,4%), suivi de *car* (17,1%) et de *puisque* (6,8%). Une possible explication est que le connecteur *parce que* survient fréquemment dans du discours rapporté ou semi-rapporté dans notre corpus (presse écrite), c'est-à-dire dans un contexte pseudo-oral. Or, nous savons qu'à l'oral *parce que* s'illustre par un emploi beaucoup plus subjectif, se substituant fréquemment à *car* (quasi inexistant en discours oral spontané) (SIMON & DEGAND, soumis). Cette hypothèse devrait être confirmée par une analyse qualitative.

5. Connecteurs et perspectivation

Différents auteurs avancent l'hypothèse que certains connecteurs marquent une rupture de perspective entre les segments causaux, et que d'autres ne le font pas. Cette perspectivisation rend compte du fait qu'un texte peut être polyphonique. Elle semble jouer un rôle dans les divergences de sens entre *want* (rupture de perspective, connecteur polyphonique) et *omdat* (pas de rupture, connecteur monophonique), mais les analyses de corpus manuelles ne semblent pas confirmer l'hypothèse de manière univoque (DEGAND, 2001; OVERSTEEGEN, 1997). De la même manière, le connecteur *puisque* et *car* ont été décrit comme polyphoniques et *parce que* comme monophonique (GROUPE λ-1, 1975; IORDANSKAIA, 1993). Pour étudier cette question de manière quantitative et automatisée nous proposons d'opérationnaliser la perspectivation en termes de proximité sémantique entre les segments causaux dans le cadre d'une analyse de sémantique latente (Section 3.2.2). Cela revient à une conception de la rupture de perspective comme une rupture dans la cohésion sémantique des segments liés par les connecteurs. Une rupture de perspective devrait impliquer une diminution de la cohésion sémantique entre les segments connectés. Les hypothèses sont les suivantes :

- Le cosinus entre les segments Q et P liés par un connecteur monophonique sera plus élevé que le cosinus entre les segments Q et P liés par un connecteur polyphonique.
- Le cosinus entre la phrase précédant et la phrase suivant les segments causaux sera plus élevé pour les connecteurs monophoniques que pour les connecteurs polyphoniques.

Pour le néerlandais, les deux hypothèses se confirment (BESTGEN, DEGAND, SPOOREN, 2003). Les segments reliés par *omdat* (monophonique) sont sémantiquement plus proches que les segments reliés par *want* (polyphonique), et *omdat* va de pair avec une continuité topicale entre la phrase précédente et la phrase suivante, ceci est moins le cas pour *want*. En outre, une analyse de contenu thématique complémentaire sur base d'un dictionnaire d'«Indicateurs de perspective» (adverbes d'attitude, « intensifieurs », « évaluateurs », ...) confirme que cette rupture dans la cohésion sémantique entre les segments causaux peut être interprétée comme une rupture de perspective. En effet, quand les segments causaux sont reliés par *want*, le segment Q contient des indicateurs de perspective, et P n'en contient pas, alors que lorsque les segments sont reliés par *omdat* une telle rupture ne se produit pas (perspective uniforme ou absence de perspective) (cf. DEGAND, SPOOREN & BESTGEN, 2004 pour les résultats détaillés).

Pour le français, les résultats vont dans le même sens. Deux analyses ANOVA furent effectuées. La première avec les connecteurs comme variable indépendante et le cosinus entre P et Q comme variable dépendante montre que la première hypothèse est quasi confirmée : Les segments reliés par *parce que* (monophonique) sont sémantiquement plus proches que les segments reliés par *car* (polyphonique) ($F(2, 8617) = 7,96, p < 0.0005$). L'hypothèse ne se confirme pas pour *puisque* qui ne diverge pas significativement de *parce que* dans cette analyse. La deuxième ANOVA avec les connecteurs comme variable indépendante et le cosinus entre la phrase précédente et la phrase suivante comme variable dépendante confirme la seconde hypothèse : Le connecteur *parce que* (monophonique) va de pair avec une continuité topicale entre la phrase précédente et la phrase suivante, ceci est (significativement) moins le cas pour *car* (polyphonique) et encore moins pour *puisque* (polyphonique) ($F(2, 8617) = 18,01, p < 0.0001$).

6. Conclusions

Dans la présente étude, nous avons combiné des approches manuelles et automatisées afin de décrire les facteurs linguistiques qui déterminent le sens et l'usage de connecteurs causaux en néerlandais et en français. Les approches manuelles sont nécessaires pour proposer des hypothèses linguistiques solides. Elles ont donc une valeur heuristique importante. Dans le cas présent, elles nous

ont permis de proposer une conceptualisation de ces connecteurs sur une échelle d'implication du locuteur allant d'un pôle objectif (implication minimale) à un pôle subjectif (implication maximale). Toutefois, la généralisabilité des résultats est loin d'être garantie. Non seulement, ces analyses ne peuvent être effectuées que sur de très petits échantillons, mais elles sont dépendantes des intuitions du chercheur qui les effectue. Ceci se marque, par exemple, dans les divergences, présentes dans la littérature scientifique, à propos de la validité de l'hypothèse de rupture de perspective selon laquelle un connecteur comme *want* serait polyphonique et introduirait une rupture de perspective à l'inverse d'un connecteur comme *omdat*.

Pour dépasser ces limitations, nous avons proposé d'employer un ensemble de techniques issues de différentes disciplines dont l'objectif est de promouvoir le traitement automatique des langues naturelles. Il s'agit tant d'étiqueteurs lexico-syntaxiques que d'analyses sémantiques issus de la linguistique computationnelle, de la psycholinguistique et de l'analyse linguistique de corpus. L'automatisme de ces procédures d'analyses permet de traiter d'une manière objective et répliquable de très grands corpus contenant plusieurs milliers d'occurrences de chaque connecteur. Nous avons pu ainsi confirmer les résultats des analyses manuelles et même apporter des arguments empiriques en faveur de l'hypothèse de perspectivation.

Les premiers résultats, rapportés ici, pour le français sont toutefois moins clairs que ceux obtenus pour le néerlandais (Bestgen, Degand et Spooren, 2003). Comprendre l'origine de cette divergence devrait permettre de tracer des pistes à explorer dans de futures recherches. Evidemment, la première question à laquelle il faudra répondre porte sur le caractère incomplet des données disponibles. Nous n'avons pu prendre en compte qu'à peine deux tiers des occurrences de *parce que* et quatre cinquièmes de celles de *puisque*. C'est tout particulièrement l'identification de ces connecteurs en position médiane externe et antéposée qui pose problème en raison de la forte variabilité des structures syntaxiques qu'autorise le français, un problème qui ne se pose pas en néerlandais. Pour améliorer l'efficacité de nos procédures, le recours à des analyses syntaxiques plus complètes comme celles développées par BOURIGAULT et FABRE (2000) semble nécessaire. Il s'agit là de la voie de développement prioritaire. L'amélioration des procédures employées pour construire les dictionnaires sémantiques nous semble aussi indispensable. Employer l'analyse sémantique latente pour étendre d'une manière semi-automatique les dictionnaires est la voie que nous explorons actuellement. Ces différents développements donneront plus crédit à nos résultats et confirmeront ou non les divergences entre les deux langues. En cas de confirmation, il deviendra indispensable de nuancer les propositions théoriques développées dans l'introduction à propos de la position des connecteurs causaux *parce que*, *car* et *puisque* sur l'échelle d'implication du locuteur. Les différences quant au fonctionnement de ces connecteurs à l'oral et à l'écrit attestées dans des études comparatives (SIMON & DEGAND, soumis) suggèrent des pistes intéressantes. Cette discussion ne doit néanmoins pas occulter le fait que, malgré le caractère exploratoire de notre analyse du corpus francophone, les résultats obtenus sont globalement conformes aux hypothèses dérivées des analyses manuelles. En conclusion, combiner ces deux types d'approche permet de contrôler les faiblesses de chacune et de tirer parti de leurs avantages réciproques.

Remerciements

L. Degand et Y. Bestgen sont chercheurs qualifiés du Fonds National de la Recherche (FNRS). Cette recherche est financée par le projet FRFC n° 2.4535.02 et par une "Action de Recherche concertée" du Gouvernement de la Communauté française de Belgique.

Bibliographie

BERRY M.W. (1992), "Large scale singular value computation", *International journal of Supercomputer Application*, 6, 13-49.

In: Sylvie Porhiel, Dominique Klingler (éds.) (2004) : *L'Unité texte*. Pleyben, pp. 49-73.

BERRY M.W., DO T., O'BRIEN G., KRISHNA V., & VARADHAN S. (1993), "*SVDPACKC: Version 1.0 User's Guide*", Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN, October 1993.

BESTGEN Y. (2002), "Détermination de la valence affective de termes dans de grands corpus de textes", *Actes du Colloque International sur la Fouille de Texte CIFT'02* (pp. 81-94). Nancy : INRIA.

BESTGEN Y. (2004), "Analyse sémantique latente et segmentation automatique des textes", in PURNELLE G., FAIRON C., & DISTER A. (Eds.), *Actes des 7^{es} Journées internationales d'Analyse statistique des Données Textuelles* (pp. 171-181), Louvain-la-Neuve : Presses universitaires de Louvain.

BESTGEN Y., & CABIAUX, A.F. (2002), "L'analyse sémantique latente et l'identification des métaphores", *Actes de la 9^{ème} Conférence annuelle sur le traitement automatique des langues naturelles*. INRIA, Nancy, pp. 331-337.

BESTGEN Y., DEGAND L., & SPOOREN W. (2003), "On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: an explorative study", in LAGERWERF L., SPOOREN W., & DEGAND L. (Eds.), *Determination of information and tenor in texts: Multidisciplinary approaches to discourse 2003* (pp. 189-202). Münster: Nodus Publikationen.

BESTGEN Y., DEGAND L., & SPOOREN W. (soumis), Towards automatic determination of the semantics of connectives in large newspaper corpora, *Discourse Processes*.

BIBER D. (1998), *Variation across speech and writing*, Cambridge: Cambridge University Press.

BOURIGAULT, D. & FABRE, C. (2000), Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de grammaire, 25, Université Toulouse Le Mirail, p. 131-151

BROUWERS, L. (1997). *Het juiste woord, betekeniswoordenboek*. 6^{ème} édition. (édité par F. Claes). Antwerpen etc.: Standaard.

BURGESS C., LIVESAY K., LUND K., (1998), "Explorations in Context Space : Words, Sentences, Discourse", *Discourse Processes*, 25, 211-257.

CHOI F., WIEMER-HASTINGS P., & MOORE J. (2001), "Latent Semantic Analysis for Text Segmentation", in LEE L., & HARMAN D. (Eds.), *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing* (pp. 109-117).

CONRAD S., & BIBER D. (Eds., 2001), *Variation in English: Multidimensional Studies*. Harlow: Longman.

DAELEMANS W., ZAVREL J., BERCK P., & GILLIS S. (1996), "MBT: A Memory-Based Part of Speech Tagger-Generator", in EJERHED E., & DAGAN I. (Eds.), *Proceedings of the Fourth Workshop on Very Large Corpora* (pp. 14-27). Copenhagen, Denmark.

DEERWESTER S., DUMAIS S.T., FURNAS G.W., LANDAUER T.K., & HARSHMAN R. (1990), "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, 41, 391-407.

DEGAND L. (1998), "Het ideationele gebruik van *want* en *omdat*: een geval van vrije variatie?" [The ideational use of *want* and *omdat*: A case of free variation?], *Nederlandse Taalkunde*, 4, 309-326.

DEGAND L. (2000), "Contextual constraints on causal sequencing in informational texts", *Functions of Language*, 7 (2), 173-201.

DEGAND L. (2001), *Form and Function of Causation. A Theoretical and Empirical Investigation of Causal Constructions in Dutch*. Leuven: Peeters. [Studies op het gebied van de Nederlandse taalkunde 5].

DEGAND L. (in press), "De l'analyse contrastive à la traduction: le cas de la paire *puisque* - *aangezien*", in WILLIAMS G. (Ed.) *Actes des deuxième journées de linguistique de corpus*. P.U.R.

DEGAND L., & PANDER MAAT H., (1999), "Scaling causal relations in terms of Speaker Involvement", *Levels of Representation in Discourse, Working Notes of the International Workshop on Text Representation*, Edinburgh University, 45-54.

In: Sylvie Porhiel, Dominique Klingler (éds.) (2004) : *L'Unité texte*. Pleyben, pp. 49-73.

DEGAND L., & PANDER MAAT H. (2003), "A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale", in VERHAGEN A., & VAN DE WEIJER J. (Eds.), *Usage based approaches to Dutch*, Utrecht: LOT, 175-199.

DEGAND L., & SANDERS T., (1999), "Causal connectives in language use. Theoretical and methodological aspects of the classification of coherence relations and connectives", *Levels of Representation in Discourse, Working Notes of the International Workshop on Text Representation*, Edinburgh University, 3-12.

DEGAND L. SPOOREN W. & BESTGEN Y. (2004), "On the Use of Automatic Tools for Large-scale Semantic Analyses of Causal Connectives", *Discourse Annotation: A Workshop in conjunction with ACL'04*, July 25-26, Barcelona, Spain

FOLTZ P.W., KINTSCH W., & LANDAUER T.K. (1998), "The measurement of textual coherence with Latent Semantic Analysis", *Discourse Processes*, 25, 285-307.

GROUPE λ-1 (1975), Car, parce que, puisque. "Revue Romane", 10, 248-280.

IORDANSKAIA, L. (1993), Pour une description lexicographique des conjonctions du français contemporain. *Le Français moderne* 2, 159-190.

KINTSCH W. (2000), "Metaphor comprehension: A computational theory", *Psychonomic Bulletin and Review*, 7, 257-266.

LANDAUER T.K., FOLTZ P.W., & LAHAM D. (1998), "An introduction to Latent Semantic Analysis", *Discourse Processes*, 25, 259-284.

LEBART L., & SALEM A. (1992), *Statistique textuelle*, Paris, Dunod.

LEMAIRE B., BIANCO M., SYLVESTRE E., & NOVECK I. (2001), "Un modèle de compréhension de textes fondé sur l'analyse de la sémantique latente", in PAUGAM MOISY H., NYCKEES V., CARON-PARGUE J. (Eds.), *La Cognition entre Individu et Société : Actes du Colloque de l'ARCo*, Paris: Hermès, 309-320.

OVERSTEEGEN, L. (1997), On the pragmatic nature of causal and contrastive connectives. *Discourse Processes*, 24, 51-86.

PANDER MAAT H., & DEGAND L. (2001), "Scaling causal relations and connectives in terms of Speaker Involvement", *Cognitive Linguistics*, 12 (3), 211-245.

PENNEBAKER J.W., MEHL M.R., & NIEDERHOFFER K.G. (2003), "Psychological aspects of natural language use : Our words, our selves", *Annual Review of Psychology*, 54, 547-577.

PERREZ, J. & DEGAND, L. (en prép.), « On the metadiscursive use of causal and contrastive connectives in Dutch Learners », manuscrit Université catholique de Louvain.

PIT M. (2003). *How to Express Yourself with a Causal Connective. Subjectivity and Causal Connectives in Dutch, German and French*. Amsterdam: Rodopi.

POPPING, R. (2000), *Computer-assisted Text Analysis*. London: Sage.

SCHMIDT H. (1994), "Probabilistic Part-of-Speech Tagging Using Decision Trees". Version électronique disponible sur [<http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html>].

SIMON, A-C & DEGAND, L. (soumis), « Connecteurs de causalité, implication du locuteur et configuration prosodique. Le cas de *car* et de *parce que*. » *French Language Studies*.

STONE P.J. (1997), "Thematic text analysis: New agendas for analyzing text content", in ROBERTS C.W. (Eds.). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*, Mahwah, NJ: Erlbaum, 35-54.

SWEETSER E. (1990), *From Etymology to Pragmatics: Metaphorical and cultural aspects of semantic structure*, Cambridge, C.U.P.,

VAN DEN BOSCH A., & DAELEMANS W. (1999), "Memory-based morphological analysis", in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99*, New Brunswick, NJ: ACL, 285-292.

VAN DIJK T.A. (1979), Pragmatic connectives, *Journal of Pragmatics*, 3, 447-456.