Une analyse des connecteurs pragmatiques fondée sur la théorie de la pertinence et son application au TALN

Sandrine Zufferey
École de traduction et d'interprétation
Université de Genève
<sandrine.zufferey@eti.unige.ch>

Résumé

Dans cet article, nous proposons d'appliquer une analyse des connecteurs pragmatiques issue de la théorie de la pertinence au traitement automatique des langues naturelles (TALN). Nous commencerons par montrer les conséquences de l'application de la théorie de la pertinence sur l'étude des connecteurs, en prenant pour exemple les connecteurs de l'anglais. Dans un deuxième temps, nous exposerons les contraintes imposées par les limites actuelles du TALN. Enfin, nous proposerons un schéma d'étude des connecteurs pragmatiques pour le TALN ainsi qu'une esquisse d'application de cette méthode.

1. Introduction

Depuis plus d'une vingtaine d'années, les connecteurs pragmatiques ont fait l'objet d'un grand nombre d'études en linguistique, notamment dans le cadre des modèles d'analyse de discours fondés sur la notion de cohérence. Traditionnellement, leur exploitation en traitement automatique des langues naturelles (ci-après TALN) s'est inscrite dans le prolongement de ces théories. Un courant de recherche s'est notamment formé en intelligence artificielle autour de la *Rhetorical Structure Theory (RST)* développée par Mann & Thompson (1988). Il inclut les travaux de Hovy (1993, 1995) et de Marcu (2000), par exemple.

La RST est fondée sur la notion de relation de discours, le but de cette théorie étant de définir le nombre de relations de discours nécessaires pour décrire la structure rhétorique interne des textes et des dialogues. Dans cette optique, les connecteurs pragmatiques jouent un rôle fondamental. En effet, ce sont précisément ces éléments qui doivent servir à détecter le type de relation de discours qui lie deux énoncés. Par exemple, en anglais, *but, however* et *nevetheless* introduisent une relation dite de *contraste*. L'un des principaux problèmes qui résulte de cet emploi des connecteurs pragmatiques vient justement de ce manque de précision. En effet, si plusieurs connecteurs peuvent

servir à exprimer la même relation, il devient difficile de définir le rôle exact de chacun d'entre eux dans le discours. Notamment, il est impossible d'expliquer pourquoi, dans certains cas, l'un de ces connecteurs peut être utilisé mais pas les deux autres[1].

Certains auteurs, conscients des limites d'une définition aussi locale de la cohérence, ont tenté de la replacer à un niveau plus général en introduisant la notion de *thème*. En linguistique informatique, cette vision de la cohérence se retrouve notamment dans les travaux de Grosz & Sidner (1986), ainsi que chez Reichman (1985). Dans la théorie du discours de Grosz et Sidner, les relations de discours servent uniquement à indiquer une hiérarchie de structures de type mère-fille, par exemple. Dans ce contexte, les connecteurs pragmatiques doivent avant tout servir d'indices servant à segmenter le discours, sans spécifier forcément la nature de la relation entre les énoncés. Toutefois, comme l'ont montré certains travaux récents, l'utilisation de la notion de thème en tant que moteur de la cohérence globale du discours est également problématique[2].

Enfin, dans le cadre de travaux sur la modélisation du dialogue, les chercheurs ont également eu recours aux connecteurs pragmatiques pour détecter des *actes de dialogue.* Bien que ce terme recouvre en partie la notion d'actes de langage, il convient de ne pas les confondre. En effet, la notion d'acte de dialogue est plus large de celle d'acte de langage. Jurafsky (2003) définit les actes de dialogue de la manière suivante : « An act with internal structure related specifically to its dialogue function ». Ainsi, chaque énoncé correspond à un acte de dialogue, lequel contient une ou plusieurs étiquettes. Ces étiquettes regroupent des informations aussi diverses que *question, suggestion, changement de sujet*, *excuses*, *reformulation,* etc. Dans ce contexte, l'utilité des connecteurs pragmatiques se situerait dans la détection automatique d'un acte de dialogue. Dans le cadre de leurs travaux sur le corpus *Trains*, Byron & Heeman (1997) ont notamment montré l'existence d'une corrélation entre les tours de parole introduits par un connecteur particulier et certaines articulations du dialogue.

Ainsi, on constate que le traitement des connecteurs pragmatiques utilisé actuellement en TALN est fortement lié aux postulats de certaines théories pragmatiques comme la théorie des actes de langage et l'analyse de discours fondée sur la notion de cohérence. Bien que ces modèles permettent un traitement peu profond de la structure des dialogues, leur exploitation des connecteurs est incomplète. C'est pourquoi, nous avons choisi de fonder notre modèle des connecteurs pragmatiques sur une autre théorie : la théorie

[1] A ce sujet, voir également Taboada (2003).

[2] Nous renvoyons le lecteur intéressé à Wilson (1998).

post-gricéenne de la pertinence (Sperber & Wilson, 1986). Nous pensons que cette théorie fournit les pistes nécessaires pour obtenir un traitement plus satisfaisant des connecteurs pragmatiques, d'un point de vue à la fois linguistique et informatique.

2. Connecteurs pragmatiques et théorie de la pertinence

2.1. La notion d'encodage procédural

En théorie de la pertinence, la caractéristique principale qui permet de définir les connecteurs pragmatiques est qu'ils encodent de *l'information procédurale*. Ce type d'information sert à contraindre ou à guider (selon la terminologie de Luscher 1999) la phase inférentielle de la communication en restreignant le nombre d'hypothèses que le locuteur doit considérer pour arriver à comprendre un énoncé. Cette définition des connecteurs pragmatiques entraîne deux conséquences majeures pour le traitement de cette classe d'éléments.

Premièrement, elle conduit à reconsidérer le rôle des connecteurs pragmatiques dans le discours. En effet, dans les approches théoriques de type *analyse de discours*, la fonction première attribuée aux connecteurs pragmatiques est de lier les éléments du discours. Le fait que les connecteurs pragmatiques aient la propriété de lier des éléments semble intuitivement très plausible et difficilement contestable. Pourtant, en théorie de la pertinence, ce n'est plus leur fonction principale. Les connecteurs pragmatiques sont désormais considérés comme des *marques procédurales* qui ont un rôle à jouer dans le traitement des informations au niveau du système central de la pensée, donc au niveau pragmatique. Ils vont notamment servir à déterminer les effets contextuels de l'énoncé et à faciliter le traitement de l'information en minimisant les efforts cognitifs. En résumé, leur rôle n'est plus de lier des éléments mais de guider l'interprétation des énoncés en donnant des instructions sur la manière de construire le contexte et de tirer des implications contextuelles.

Deuxièmement, le fait d'étudier les connecteurs pragmatiques selon la théorie de la pertinence détermine aussi la liste d'éléments à inclure dans cette classe. En effet, il n'existe pas de liste de connecteurs pragmatiques unanimement acceptée, et les choix des auteurs varient considérablement selon l'approche théorique envisagée, et selon la fonction attribuée aux connecteurs. Par exemple, pour les connecteurs pragmatiques de l'anglais, Fraser (1990) dresse une liste de 32 éléments alors que Schiffrin (1987) n'en dénombre que 23. Qui plus est, leurs listes ne comptent que 5 éléments communs. Cette absence de consensus reflète à la fois la grande diversité des approches employées pour étudier les connecteurs pragmatiques et le grand nombre de fonctions qui leur ont été attribuées.

En théorie de la pertinence, les éléments considérés comme des connecteurs pragmatiques sont ceux qui encodent de l'information procédurale. Selon cette définition, les connecteurs pragmatiques ne forment pas une classe unique et homogène. En effet, certains éléments traditionnellement considérés comme des connecteurs, comme *and* et *because*, n'encodent pas de l'information procédurale mais de l'information conceptuelle. En revanche, d'autres éléments comme *well* ou *like*, plus fréquemment étiquetés comme des marqueurs, sont inclus dans cette catégorie, car ils encodent de l'information procédurale.

Pourtant, il semble intuitivement que deux éléments tels que *like* et *although* n'apportent pas exactement le même type de contribution aux énoncés. C'est d'ailleurs pour cette raison que de nombreux auteurs opèrent une distinction entre d'une part les connecteurs pragmatiques (*although*, *but*, *however*) et d'autre part les marqueurs pragmatiques (*like*, *well*, *now*). Cette intuition trouve naturellement son explication dans la théorie de la pertinence. En effet, bien que tous ces éléments encodent de l'information procédurale, cette dernière est de nature différente selon les cas. Pour ce qui est des éléments traditionnellement répertoriés comme des connecteurs, leur procédure encode un effet cognitif précis (ajout d'une nouvelle hypothèse, modification, etc.). En revanche, dans le cas du deuxième groupe d'éléments, leur information procédurale ne porte pas sur un effet cognitif précis. Par exemple *well* encode simplement une garantie de pertinence de la part du locuteur qui l'emploie.

2.2. Critères pour la détection du contenu procédural

Étant donné que la notion d'encodage procédural est au cœur de notre définition des connecteurs pragmatiques, il est essentiel de pouvoir déterminer si un élément encode ce type d'information. Il existe un certain nombre de critères couramment employés pour établir la présence d'un contenu procédural. Nous proposons ici une synthèse des critères les plus fréquemment cités[3], que nous avons divisés en deux catégories : d'une part ceux s'appuyant sur des éléments d'ordre cognitif et d'autre part, ceux reposant sur un principe de comparaison avec d'autres éléments.

2.1.1. Critères cognitifs

Les éléments qui encodent de l'information procédurale ne sont pas facilement paraphrasables. En effet, contrairement à un mot comme *arbre,* que l'on peut facilement définir comme un végétal possédant un tronc, des racines, des branches et des feuilles, par exemple, il est difficile de donner une

[3] Voir notamment Blakemore (2002) et Rouchota (1998)

définition de *mais* sans se contenter de recourir à des exemples. Par ailleurs, même des locuteurs natifs sont incapables de déterminer si deux connecteurs pragmatiques apparemment synonymes (par exemple *but* et *however*) le sont réellement sans essayer de remplacer l'un par l'autre dans de nombreux contextes.

Les éléments qui encodent de l'information procédurale sont notoirement difficiles à traduire et posent souvent de nombreux problèmes aux locuteurs qui apprennent une langue étrangère. Selon Wilson & Sperber (1993), cette difficulté est étroitement liée au fait que ces éléments encodent une procédure. Selon eux, seules les représentations conceptuelles peuvent être rendues conscientes, alors que nous n'avons accès directement ni aux règles de grammaire ni aux opérations procédurales.

2.1.2. Critères comparatifs

Les éléments qui encodent de l'information procédurale et qui sont polysémiques ne sont pas synonymes quand ils sont employés de manière procédurale ou conceptuelle. En effet, la distinction entre éléments conceptuels et procéduraux ne correspond pas toujours à la distinction entre éléments vériconditionnels et non-vériconditionnels. Or, il existe des différences entre les éléments non-vériconditionnels qui encodent de l'information conceptuelle et ceux qui encodent de l'information procédurale. Par exemple, certains adverbes de phrase sont non-vériconditionnels mais encodent de l'information conceptuelle. On constate que ces adverbes gardent la même signification quand ils sont employés de manière non-vériconditionnelle et vériconditionnelle, par exemple dans un syntagme verbal. Ce n'est pas le cas des connecteurs pragmatiques qui ont un équivalent vériconditionnel, comme le montrent les exemples suivants :

(1) Frankly, you have to go now.

(2) Tell me frankly what you think of it.

(3) Well, I'll see what I can do.

(4) You did very well at your exam.

Les emplois de l'adverbe *frankly* en (1) et (2) sont synonymes, bien qu'en (1), il soit non-vériconditionnel et vériconditionnel en (2). En revanche, en (3) et (4), les deux emplois de *well* (connecteur pragmatique et adjectif) ne sont pas synonymes.

Les éléments qui encodent de l'information procédurale ne peuvent pas se combiner entre eux pour former des représentations complexes. Ils se distinguent en cela également des adverbes de phrase non-vériconditionnels. Prenons un exemple :

(5) In total, absolute confidence, I will tell you my story.

Dans cet exemple, *total* et *absolute* peuvent se combiner pour former une représentation conceptuelle complexe. Ce n'est pas le cas des connecteurs pragmatiques, comme le montre cet exemple :

(6) ? Peter doesn't like classical music. Totally nevertheless, Marie loves operas.

A nouveau, cette différence s'explique très bien en vertu de la distinction entre encodage conceptuel et procédural. En effet, étant donné que ces adverbes de phrase encodent des concepts, il est normal que ces derniers puissent se combiner pour former des représentations conceptuelles complexes, en vertu des règles de compositionnalité sémantique. Ce n'est pas le cas des procédures.

Plusieurs connecteurs procéduraux peuvent être employés dans un même énoncé sans que ce dernier devienne illogique ou redondant. Ce n'est pas le cas des connecteurs conceptuels. Prenons un exemple, emprunté à Rouchota :

(7) There's a bird in the garden.

(8) *So* the cat didn't eat them all *then*.

(9) ? Consequently, the cat didn't eat them all, *in conclusion*.

Dans l'exemple (8), s*o* et *then* encodent le même type d'information procédurale. On pourrait définir très grossièrement l'instruction qu'ils contiennent de la manière suivante : « traiter l'énoncé comme une conclusion ». Pourtant, la phrase n'est pas redondante. Par contre, si on remplace ces connecteurs procéduraux par des connecteurs conceptuels, le résultat devient clairement redondant, comme le montre l'exemple (9).

3. Contraintes imposées par le traitement automatique des langues naturelles (TALN)

Nous avons vu jusqu'à présent comment analyser les connecteurs pragmatiques en appliquant la théorie de la pertinence. Passons maintenant au volet suivant de notre étude, à savoir l'application de cette méthode en TALN.

L'un des principaux défis de la modélisation informatique consiste à concilier des analyses linguistiques souvent très fines avec les limites actuelles imposées par le TALN. En effet, les théories proposées par les pragmaticiens ne sont généralement pas applicables telles quelles. C'est pourquoi, depuis une dizaine d'années, la tendance en TALN a été de s'en détourner et de se contenter de formuler quelques règles simples et facilement implémentables, fondées sur des concepts *ad hoc* comme celui d'acte de dialogue.

Nous avons choisi de chercher un autre compromis, en reconnaissant la nécessité de fonder notre analyse sur une théorie générale qui fournisse un modèle cognitivement plausible des aspects pragmatiques de la communication. C'est ainsi que nous nous sommes tournée vers la théorie de la perti-

nence pour trouver des éléments qui nous permettent de modéliser l'information contenue dans les connecteurs pragmatiques.

Toutefois, certains récents travaux montrent que la notion d'encodage procédural est plus complexe qu'on ne le pensait jusque là. En analysant certains connecteurs pragmatiques du français, Moeschler (2002) a notamment montré que certains connecteurs ont à la fois un contenu procédural faible et un contenu conceptuel faible (*et*), d'autres ont un contenu procédural moyen et un contenu conceptuel faible (*ensuite*) et d'autres encore ont un contenu procédural fort et un contenu conceptuel faible (*mais*). Nous n'avons pas pu opérer de distinction aussi précise dans notre modèle, car des nuances aussi fines sont impossibles à capter dans une application informatique. Nous nous sommes donc limitée à une distinction binaire entre les éléments procéduraux et conceptuels, en concentrant notre analyse sur les premiers.

Nous pensons en effet qu'il n'est ni nécessaire ni réaliste de tenter de reproduire la finesse de ces analyses en TALN. Toute la difficulté consiste donc à trouver le niveau de granularité adéquat, à la fois applicable et qui nous permette de capter les informations contenues dans les connecteurs pragmatiques, et que nous souhaitons exploiter.

Pour trouver la granularité adéquate, il convient selon nous de retourner le problème. En effet, après s'être demandé quelles étaient les informations que les connecteurs pragmatiques pouvaient fournir, il est selon nous nécessaire de s'interroger sur le type d'information que nous souhaitons capter, au minimum, dans notre modèle de TALN. Nous avons ainsi défini un schéma d'étude pour les connecteurs pragmatiques en TALN.

4. Définition d'un schéma d'étude des connecteurs pragmatiques pour le TALN

Notre modèle pour l'étude des connecteurs pragmatiques en TALN se divise en trois étapes, que nous présentons tour à tour.

La première a trait à la *désambiguïsation des connecteurs pragmatiques*. En effet, presque tous les connecteurs sont des éléments lexicaux polysémiques. Prenons quelques exemples :

(10) I like music so much.

(11) So, what do you have to say ?

(12) However hard I try, I cannot do it.

(13) I would like to come, however I cannot afford the ticket.

Dans les exemples (10) et (12), *so* et *however* n'ont pas la fonction de connecteur pragmatique, alors qu'il sont bien des connecteurs en (11) et en (13). Le premier pas pour pouvoir traiter adéquatement ces éléments de ma-

nière automatique est donc de savoir les reconnaître et n'extraire que les occurrences dans lesquelles le connecteur a une fonction pragmatique.

En plus de fournir la première étape de notre analyse des connecteurs, la désambiguïsation des occurrences pragmatiques des connecteurs permet également d'améliorer la qualité de l'analyse morpho-syntaxique en réduisant le nombre d'erreurs. Par exemple, un connecteur aussi ambigu que *like* pourrait être étiqueté à tort comme un verbe ou un adverbe, ce qui conduirait à l'échec de l'analyse grammaticale de l'énoncé. Cet avantage n'est pas négligeable si l'on considère que des corpus de dialogues, près de 30% des occurrences de *like* correspondent à des connecteurs pragmatiques. Nous reviendrons sur ce problème dans la suite de cet article lorsque nous monterons un exemple de désambiguïsation de *like*.

La deuxième tâche consiste à *définir formellement l'instruction de chaque connecteur*. Cette tâche est avant tout linguistique et peut se faire indépendamment de tout traitement informatique. Elle reste néanmoins très complexe car il est possible d'obtenir des résultats contradictoires en appliquant la même théorie[4]. Nous pensons que dans les cas où plusieurs instructions semblent envisageables pour un même connecteur, elles doivent être départagées empiriquement.

Enfin, la troisième tâche consiste à *définir la portée de l'instruction procédurale* contenue dans le connecteur pragmatique. A titre d'illustration, prenons l'exemple de *like*. L'analyse linguistique menée par Andersen (1998) l'a conduit à déterminer que *like* est un connecteur qui indique une approximation (*loose talk marker*). Or, en fonction des cas, l'approximation véhiculée par *like* peut porter sur des éléments très divers. Prenons deux exemples :

(14) It was like twenty minutes before she came.

(15) I was like oh okay !

Dans l'exemple (14), l'instruction de *like* porte sur la durée qui suit le connecteur, c'est-à-dire 20 minutes. Il indique que cette durée est approximative. Dans l'exemple (15), *like* sert à introduire une citation et son instruction porte sur le reste de l'énoncé (*oh okay*). Selon nous, il ne suffit donc pas d'établir que *like* introduit une approximation dans l'énoncé, il est indispensable de déterminer sur quels éléments exactement porte cette instruction. Cette troisième tâche constitue, de même que la désambiguïsation, un autre argument en faveur d'un traitement individualisé pour chaque connecteur pragmatique.

---

[4] Voir par exemple les analyses divergentes que Jucker (1993) et Blakemore (2002) proposent de *well,* en se fondant tous deux sur la théorie de la pertinence.

En résumé, nous avons défini trois tâches pour le traitement informatique des connecteurs pragmatiques, définis comme des éléments linguistiques encodant de l'information procédurale. La première consiste à déterminer automatiquement le statut (pragmatique ou non) de chaque occurrence du connecteur. La deuxième a pour objectif de définir formellement une seule instruction pour chaque connecteur pragmatique, qui permette de décrire la fonction de toutes ses occurrences. Enfin, la troisième consiste à trouver des critères pour déterminer la portée de l'instruction de chaque connecteur, en établissant notamment des catégories d'emplois.

Dans le reste de cet article, nous montrons une esquisse d'application de la première tâche définie dans ce modèle, à savoir la détection des connecteurs pragmatiques.

5. Détection automatique des connecteurs pragmatiques

A titre d'exemple, nous avons choisi d'appliquer cette tâche au connecteur pragmatique *like*, qui est certainement l'un des plus difficiles à traiter du point de vue de l'ambiguïté. Cet élément compte en effet de nombreux autres emplois que celui de connecteur pragmatique. Il peut notamment avoir le rôle de préposition (16), d'adjectif (17), de conjonction (18), d'adverbe (19), de nom (20) et de verbe (21), comme le montrent respectivement les exemples suivants[5] :

(16) He was *like* a son to me.

(17) Cooking, ironing and *like* chores.

(18) Nobody can sing that song *like* he did.

(19) It's nothing *like* as nice a their previous house !

(20) Scenes of unrest the *like(s)* of which had never been seen before in the city.

(21) I *like* chocolate very much.

Avant d'envisager un traitement informatique pour tenter d'extraire automatiquement les occurrences pragmatiques de *like*, nous avons décidé d'effectuer deux premières expériences avec des annotateurs humains. Ces expériences sont utiles pour plusieurs raisons. Avant tout, elles permettent de prendre la mesure de la difficulté de cette tâche. Dans cette optique, ce type d'expériences fournit notamment un point de comparaison pour évaluer la qualité des performances obtenues par traitement informatique, lesquelles seront appréciées en fonction des résultats de référence obtenus par les humains. Nous commencerons par relater brièvement les résultats obtenus dans ces deux expériences, puis nous proposerons une méthode permettant d'automatiser la tâche de désambiguïsation.

---

[5] Adaptés du *Dictionnaire Hachette Oxford*, Oxford, OUP, 1994.

5.1. Expériences avec des juges humains

5.1.1. Description des expériences

Nous avons conçu notre test d'annotation en deux parties. Dans une première expérience, nous avons retenu deux groupes de trois annotateurs humains. Le premier était constitué de locuteurs francophones possédant une bonne maîtrise de l'anglais et le second comprenait des locuteurs natifs de l'anglais.

Nous leur avons soumis deux séries d'énoncés provenant de corpus différents, comptant respectivement 26 et 49 occurrences de *like*. Les extraits du premier corpus étaient tirés des dialogues du film *Pretty Woman*. Les énoncés du second corpus correspondaient à une réunion d'une heure tirée du corpus de dialogues réalisé par *l'International Computer Science Institute* (ICSI), à Berkeley (cf. Janin 2003).

La tâche d'annotation était définie de la manière suivante. Chaque annotateur reçoit par écrit quelques brèves explications concernant le rôle de *like* en tant que connecteur pragmatique, ainsi que des exemples d'emplois pragmatiques et non-pragmatiques de ce connecteur. Chaque annotateur doit décider pour toutes les occurrences de *like* s'il s'agit d'un connecteur pragmatique ou non. Ils doivent en outre spécifier le degré de certitude de leurs réponses sur une échelle allant de 1 à 3 (1 = bonne certitude, 2 = certitude moyenne, 3 = hésitation).

Dans la première expérience, les annotateurs se fondaient uniquement sur la lecture des énoncés pour décider du statut d'une occurrence de *like*. Or, à l'issue de ce premier test, plusieurs de nos annotateurs nous ont fait savoir qu'ils auraient souhaité pouvoir entendre les énoncés. C'est pourquoi, nous avons exploré la possibilité d'améliorer le niveau d'accord entre les annotateurs par une désambiguïsation prosodique.

Ainsi, nous avons élaboré une seconde expérience d'annotation sur le même principe que la première, en donnant aux annotateurs l'accès à l'enregistrement audio de la réunion. Par ailleurs, nous n'avons plus retenu de dialogues de film. Étant donné que cette expérience était plus difficile à réaliser pour des raisons techniques et plus contraignante pour les participants, nous avons limité l'expérience à un groupe mixte de trois annotateurs. Deux des sujets avaient déjà participé à la première expérience. Les consignes données aux annotateurs étaient identiques à celles de la première expérience, mais elles comprenaient également quelques informations concernant les propriétés prosodiques de *like* lorsqu'il est utilisé comme connecteur pragmatique. Aucune contrainte de temps n'a été imposée, les sujets ont donc pu écouter chaque énoncé autant de fois qu'ils le souhaitaient.

5.1.2. Résultats

Nous avons tout d'abord constaté que la tâche d'annotation des connecteurs pragmatiques est complexe, même pour des annotateurs humains. En effet, dans la première expérience, l'accord mesuré[6] était très faible (k=0.40) pour les dialogues naturels et moyen pour les dialogues planifiés (k=0.65). En revanche, lorsqu'ils ont accès à des indices prosodique, l'accord devient nettement plus fiable[7] (k=0.74). Ces deux expériences nous ont également permis de tirer les conclusions suivantes.

En réalisant ces expériences, nous pensions que l'emploi des connecteurs pragmatiques variait entre des dialogues planifiés comme ceux d'un roman ou d'un film et des dialogues réels comme ceux des réunions. Nous sommes partie du principe que leur désambiguïsation devrait être nettement moins problématique dans le cas de dialogues planifiés et que donc, l'accord entre les annotateurs serait meilleur pour les extraits du premier corpus. Comme nous le pensions, les occurrences provenant de dialogues planifiés se sont avérées plus faciles à annoter que celles provenant de dialogues naturels. Ce résultat confirme que même si les dialogues de films sont conçus pour reproduire des dialogues naturels, ils ne comportent jamais autant d'ambiguïtés que des dialogues réels, notamment en raison du fait qu'ils reflètent l'intention informative globale d'une seule personne.

Notre seconde hypothèse concernait la différence entre les anglophones et les non-anglophones. Nous pensions notamment que l'usage des connecteurs pragmatiques se faisait de manière instinctive chez des locuteurs natifs et que l'accord serait meilleur parmi le groupe des anglophones. Contrairement à notre prédiction, le groupe des francophones a obtenu un accord extrêmement similaire à celui des anglophones, dans les deux types de corpus.

Enfin, nous voulions tester la présence d'une corrélation entre le degré de certitude avec lequel une réponse était donnée et l'accord entre les annotateurs pour cette occurrence. Nous n'avons pu relever aucune corrélation entre la certitude exprimée par les annotateurs et le niveau d'accord observé. Ainsi, la capacité à juger de son intuition sur ce point ne semble pas très avancée.

En se penchant plus précisément sur le contenu des énoncés litigieux, on constate que certains types d'occurrences de *like* sont plus problématiques pour les annotateurs dans les deux expériences. Notamment, les cas de désaccord portent presque toujours sur les occurrences où *like* a la fonction de

---

[6] L'accord a été mesuré au moyen du coefficient *kappa* en utilisant l'échelle de Krippendorff. Voir par exemple Di Eugenio (2000).

[7] Selon l'échelle de Krippendorff, un résultat supérieur ou égal à 0.8 permet de tirer des conclusions certaines et un résultat compris entre 0.67 et 0.8 permet d'établir la présence probable d'un accord.

préposition. Par exemple, un des annotateurs considérait systématiquement les cas où la préposition *like* sert à exprimer une ressemblance comme un connecteur pragmatique. Ainsi, toutes les occurrences du type *sounds like*, *seems like*, *feels like*, etc. étaient systématiquement classées sous l'étiquette de connecteur pragmatique. Ce résultat n'est toutefois pas vraiment surprenant si l'on garde à l'esprit que la fonction pragmatique de *like* est issue d'un processus de grammaticalisation. Or, l'emploi de *like* en tant que préposition semble être à la forme dont dérive le connecteur pragmatique. Comme le confirme Andersen (2001, 294) : « The fundamental assumption here is that the pragmatic marker *like* originates in a lexical item, that is, a preposition with the inherent meaning "similar to" ». Ce type de constatation laisse à penser que des explications plus approfondies concernant le rôle de *like* ainsi qu'un bon entraînement permettrait, dans une certaine mesure, d'améliorer les résultats.

En résumé, ces deux expériences nous ont notamment permis de quantifier le niveau d'accord entre des annotateurs humains et de confirmer l'utilité de recourir à des critères prosodiques pour désambiguïser efficacement les occurrences d'un connecteur.

5.2. Vers une désambiguïsation automatique de *like*

Après avoir évalué la difficulté pour des annotateurs humains à étiqueter les occurrences d'un connecteur pragmatique ambigu, nous allons proposer une liste de critères utiles pour déterminer le statut d'un connecteur et discuterons de la possibilité d'automatiser l'utilisation de ces critères.

5.2.1. Critères pour déterminer la présence d'un connecteur

En dépit de la difficulté posée par l'ambiguïté des connecteurs pragmatiques, l'extraction manuelle de cette classe d'éléments reste néanmoins une tâche réalisable de manière fiable dans la plupart des cas. Trois critères sont particulièrement importants pour déterminer le statut d'un connecteur. Nous allons les passer en revue tour à tour.

Le premier critère correspond à la présence de *collocations*. Par exemple, le connecteur pragmatique *like* est très souvent immédiatement suivi ou précédé de *sort of, kind of* ou encore de *you know*. De même, lorsque *well* sert à marquer un changement de thème, il apparaît presque toujours sous la forme *well you know*, *well now*, *well I think* ou *oh well*. Quand il sert au contraire à clore un sujet, *well* se trouve souvent dans des combinaisons du type *Ok well* ou *well anyway/anyhow*. Ce critère peut également s'appliquer *a contrario*, à savoir pour déterminer des situations dans lesquelles l'élément en question n'a certainement pas le rôle de connecteur pragmatique, par exemple, lorsque *like* apparaît dans des combinaisons du type *I/you like*, *seems/feels like*, *just*

like ou lorsque *well* apparaît dans des constructions du type : *very well*, *as well*, *quite well*, etc.

Le deuxième critère est celui de la *position dans l'énoncé*. A nouveau, en fonction de l'élément en question, son emplacement au sein de l'énoncé peut servir à conclure à la présence d'un connecteur ou au contraire à l'exclure. Par exemple, le connecteur pragmatique *well* est presque toujours situé au début de l'énoncé ou du moins au début d'une phrase prosodique. Dans d'autres cas, le critère de la position dans l'énoncé implique que, pour avoir le statut de connecteur pragmatique, l'élément doive se situer ailleurs dans l'énoncé. Selon Aijmer (2002, 30) :

> « Some of the discourse particle […] (*actually, sort of*) can, for instance, be inserted parenthetically or finally, often with little difference in meaning, after a sentence, clause, turn, tone unit as a post-end filed constituent. »

Le troisième critère utile pour détecter les connecteurs est celui de la *prosodie*. L'importance de ce critère a été soulignée par de nombreux auteurs d'études sur les connecteurs. Schiffrin (1987, 328) a notamment relevé que « [a discourse particle] has to have a range of prosodic contours e.g. tonic stress and followed by a pause, phonological reduction ».

Toutefois, bien que ces critères rendent l'extraction manuelle des connecteurs réalisable dans la grande majorité des cas, certaines rares occurrences restent néanmoins ambiguës. Certains usages se situent à la limite entre l'usage pragmatique et non pragmatique et il n'est pas possible de trancher de manière certaine, les deux interprétations étant également possibles.

5.2.2. Automatisation des critères de détection

Les trois critères que nous venons de détailler permettent de réaliser l'extraction *manuelle* des connecteurs pragmatiques. Voyons maintenant s'il est possible de les appliquer automatiquement, en vue de les utiliser en TALN.

Notons tout d'abord que chacun de ces critères est relativement aisé à automatiser individuellement. Par exemple, il est très facile d'automatiser le critère des collocations dès lors qu'une liste fiable a été établie. Il peut s'agir d'une liste impliquant la présence d'un connecteur ou au contraire excluant définitivement sa présence. Dans certains cas, ce critère se révèle en effet plus efficace *a contrario*. Il est également relativement simple d'automatiser le critère de la position dans l'énoncé, lorsque cette dernière est très fortement contrainte (par exemple, au début de l'énoncé). Quant au critère prosodique, il est également facilement utilisable sur des données pourvues d'une annotation adéquate.

Toutefois, aucun de ces critères ne peut servir à lui tout seul à automatiser entièrement la tâche d'extraction, même si, dans certains cas, un seul critère

suffit à fournir de très bons résultats. Ainsi, pour le connecteur *well*, le seul critère de la position dans l'énoncé permet d'extraire correctement une grande partie des occurrences. Il est malgré tout insuffisant. En effet, *well* se situe très fréquemment au début de l'énoncé, mais il peut aussi se situer au début d'une phrase prosodique qui ne se trouve pas au début de l'énoncé. Dans ce cas, le recours à la prosodie devient indispensable. De même, l'exclusion de collocations de type *very well*, *as well*, etc. permettra de résoudre les derniers cas problématiques.

En résumé, l'utilisation de chacun de ces facteurs permet d'automatiser partiellement la tâche d'extraction et peut s'avérer très utile pour faciliter la tâche de l'annotateur humain. Par exemple, nous avons mené une expérience sur la possibilité d'utiliser une liste de collocations à exclure pour extraire automatiquement *like*. Cette méthode a fourni des résultats encourageants (le rappel était de 100% et la précision de 60% environ sur les deux corpus testés).

En revanche, une automatisation complète de l'extraction en vue du TALN ne peut se faire que par l'interaction de ces trois critères. D'un point de vue informatique, cette tâche apparaît comme nettement plus complexe. Pourtant, l'émergence de nouveaux outils, et notamment de programmes d'apprentissage automatique, constitue une piste prometteuse pour cette automatisation[8]. De plus amples travaux dans ce sens font partie du programme de recherche de l'auteur.

6. Conclusion

Dans cet article, nous avons mis en relief la nécessité d'ancrer l'analyse des connecteurs pragmatiques dans un modèle de la communication cognitivement plausible comme la théorie de la pertinence. La modélisation des connecteurs pragmatiques fondée sur les principes énoncés ci-dessus s'inscrit dans une vision du TALN à long terme. Nous pensons que la multiplication de ces efforts dans le temps permettra à l'avenir un traitement nettement plus satisfaisant des connecteurs pragmatiques que celui proposé aujourd'hui. Par ailleurs, bien que l'application de l'ensemble de la théorie de la pertinence au TALN semble être une tâche irréalisable pour le moment, l'addition de ces micro-analyses permettra à terme un traitement global des aspects pragmatiques de la communication.

[8] Ces outils fournissent déjà de bons résultats pour l'extraction des actes de dialogue. Voir notamment Jurafsky (2003).

Remerciements

L'auteur remercie chaleureusement le professeur Jacques Moeschler (Université de Genève) pour l'avoir encouragée dans la rédaction de cet article ainsi que dans la réflexion qui l'a précédée.

Bibliographie

AIJMER K. (2002), *English Discourse Particles : Evidence from a Corpus*, Amsterdam, John Benjamins.

ANDERSEN G. (1998), « The pragmatic marker like from a relevance-theoretic perspective », in JUCKER A. & ZIV Y. (eds), *Discourse Markers*, Amsterdam, John Benjamins, 147-170.

ANDERSEN G. (2001), *Pragmatic Markers of Sociolinguistic Variation : a Relevance-Theoretic Approach to the Language of Adolescents*, Amsterdam, John Benjamins.

BLAKEMORE D. (2002), *Meaning and Relevance : the Semantics and Pragmatics of Discourse Markers*, Cambridge, Cambridge University Press.

BYRON D. & HEEMAN P. (1997), « Discourse markers use in task-oriented spoken-dialog », *Proceedings of Eurospeech*, Rhodes (Grèce).

DI EUGENIO B. (2000), « On the usage of Kappa to evaluate agreement on coding tasks », *Proceedings of LREC*, Athènes (Grèce).

FRASER B. (1990), « An approach to discourse markers », *Journal of Pragmatics* 14, 383-395.

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

HOVY E. (1993), « Automated discourse generation using discourse structure relations », *Artificial Intelligence* 63, 341-385.

HOVY E. (1995), « The multifunctionality of discourse markers », *Proceedings of Workshop on Discourse Markers*, Egmond-aan-Zee (Les Pays-Bas).

JANIN A. et al. (2003), « The ICSI meeting corpus », *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Hong-Kong, publication électronique disponible à l'adresse : http://www.icsi.berkeley.edu/ftp/global/pub/speech/papers/icassp03-janin.pdf.

JUCKER A. (1993), « The discourse marker *well* : a relevance-theoretic account », *Journal of Pragmatics* 19, 435-452.

JURAFSKY D. (2003), « Pragmatics and computational linguistics », in WARD G. & HORN L. (eds), *Handbook of Pragmatics*, Oxford, Blackwell.

LUSCHER J.-M. (1999), *Eléments d'une pragmatique procédurale. Le rôle des marques linguistiques dans l'interprétation*, Thèse de doctorat, Université de Genève.

MANN W. & THOMPSON S. (1988), « Rhetorical structure theory : toward a functional theory of text organisation », *Text* 8, 243-281.

MARCU D. (2000), *The Theory and Practice of Discourse Parsing and Summarization*, Cambridge (Mass.), The MIT Press.

MOESCHLER J. (2002), « Connecteurs, encodage conceptuel et encodage procédural », *Cahiers de Linguistique Française* 24, 265-292.

REICHMAN R. (1985), *Getting Computers to Talk like You and Me : Discourse Context, Focus and Semantics (an ATN Model)*, Cambridge (Mass.), The MIT Press.

ROUCHOTA V. (1998), « Connectives, coherence and relevance », in ROUCHOTA V. & JUCKER A. (eds), *Current Issues in Relevance Theory*, Amsterdam, John Benjamins, 11-57.

SCHIFFRIN D. (1987), *Discourse Markers*, Cambridge, Cambridge University Press.

SPERBER D. & WILSON D. (1986), *Relevance : Communication and Cognition*, Oxford, Blackwell.

TABOADA M. (2003), « Discourse markers as signals (or not) of rhetorical relations in conversation », *Paper presented at the 8th International Pragmatics Conference*, Toronto (CA).

WILSON D. (1998), « Discourse, coherence and relevance : a reply to Rachel Giora », *Journal of Pragmatics*, 29, 57-74.

WILSON D. & SPERBER D. (1993), « Linguistic form and relevance », *Lingua*, 90, 1-25.