



HAL
open science

Détection automatique de structures fines de texte

Nicolas Hernandez, Brigitte Grau

► **To cite this version:**

Nicolas Hernandez, Brigitte Grau. Détection automatique de structures fines de texte. Actes de la 12e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005), May 2005, Dourdan, France. hal-02327366

HAL Id: hal-02327366

<https://hal.science/hal-02327366>

Submitted on 22 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Détection Automatique de Structures Fines de Texte

Nicolas Hernandez et Brigitte Grau
LIMSI/CNRS - LIR – Université de Paris-Sud
BP 133, F-91403 ORSAY CEDEX (France)
Hernandez|Grau@limsi.fr

Mots-clefs : Navigation intra-documentaire, analyse thématique, structures du discours, relations discursives, subordination et coordination, parallélisme lexico-syntaxico-sémantique, modèle d'apprentissage, analyses linguistiques

Keywords: Text browsing, topic analysis, text structures, discursive relations, subordination and coordination, lexical, syntactic and semantic parallelism, learning model, linguistic analysis

Résumé Dans ce papier, nous présentons un système de Détection de Structures fines de Texte (appelé *DST*). *DST* utilise un modèle prédictif obtenu par un algorithme d'apprentissage qui, pour une configuration d'indices discursifs donnés, prédit le type de relation de dépendance existant entre deux énoncés. Trois types d'indices discursifs ont été considérés (des relations lexicales, des connecteurs et un parallélisme syntaxico-sémantique) ; leur repérage repose sur des heuristiques. Nous montrons que notre système se classe parmi les plus performants.

Abstract In this paper, we present a system which aims at detecting fine-grained text structures (we call it *DST*). Based on discursive clues, *DST* uses a learning model to predict dependency relations between two given utterances. As discourse clues, we consider lexical relations, connectors and key phrases, and parallelism. We show that our system implements an improvement over current systems.

1 Introduction

Comme le souligne l'annonce du 14 décembre 2004 de la société Google de numériser et de rendre disponible en ligne 15 millions de livres appartenants à 5 des plus célèbres bibliothèques anglo-saxonnes du monde¹, le besoin d'accéder facilement et rapidement au contenu d'un document électronique est plus que jamais un enjeu d'actualité.

Cette facilitation passe par la structuration du contenu informationnel du document. De nombreux travaux (principalement au sein de la communauté de résumé automatique) en ont montré l'intérêt par une présentation organisée de l'information : hiérarchiser l'information, manipuler des unités de texte de différentes granularités (i.e. différents degrés informationnels), fournir un contexte à une information ciblée et permettre une navigation intra-documentaire, etc. (Moens & Busser, 2001; Choi, 2002; Couto *et al.*, 2004).

En particulier nous nous focalisons sur la micro-structure d'un texte (niveau phrastique voire propositionnel). Nous apportons ainsi une complémentarité à une approche globale qui délimite et rapproche des segments thématiques en offrant la possibilité de raffiner ce modèle. En effet qu'elles supposent une organisation plate et linéaire du flot d'informations communiqué (Hearst, 1997; Choi, 2002), ou bien une organisation plus riche en arbres (Moens & Busser, 2001; Couto *et al.*, 2004), les approches globales sont généralement fondées sur des mesures de cohésion lexicale (notamment à travers le suivi de chaînes lexicales) qui souffrent d'un manque de précision quant à la délimitation des unités de texte (appelées segment). De plus elles prennent rarement en compte dans leur analyse les phénomènes discursifs locaux (e.g. les annonces thématiques, les structures énumératives, les marques transitions, etc.).

Notre approche se situe parmi les travaux qui proposent de rechercher le point d'attache optimal d'un énoncé entrant dans la structure en cours de construction. Parmi les approches existantes, Marcu (1999a) propose un système pour la détection automatique de la structure rhétorique d'un texte, Choi (2002) s'intéresse à une structuration thématique fine, Kruijff-Korbayová & Kruijff (1996) analysent le discours en terme de progression thématique, Polanyi *et al.* (2004) modélisent la structure sémantique du discours. Ces systèmes constituent de sérieuses avancées mais requièrent encore la prise en compte de plus d'indices discursifs et de modèles plus souples pour appréhender les différents mécanismes de structuration du discours.

Dans ce papier, nous présentons un système de Détection de Structures fines de Texte (*DST*). Comme Marcu et Choi, *DST* utilise un modèle prédictif obtenu par un apprentissage supervisé. Celui-ci prédit la relation de dépendance existante entre deux énoncés donnés en fonction des indices discursifs présents. A la différence des systèmes de Marcu et de Polanyi *et al.*, nous proposons, comme Choi, un modèle théorique simplifié de la structure du discours. Nous nous intéressons uniquement au rapport structurel élémentaire liant deux énoncés (relation de subordination, de coordination, et absence de relation), indépendamment d'un éventuel étiquetage sémantico-rhétorique de la relation². Par ailleurs, nous dissocions le problème de la reconnaissance de relation de celui de la recherche du point d'attache de l'énoncé entrant, ce qui nous permet d'envisager différents algorithmes de structuration. Un de nos apports est de proposer une mesure pour reconnaître le parallélisme lexico-syntaxico-sémantique de deux énoncés, indice discursif peu considéré jusqu'à présent. Nous avons travaillé sur des articles scientifiques en anglais mais notre démarche est adaptable à d'autres langues comme le français.

¹New York Public Library, University of Michigan, Stanford, Harvard (USA), Oxford (GB).

²Cette tâche sera abordée ultérieurement.

2 L'accès au contenu

Le processus de compréhension requiert d'une part *d'identifier des unités discursives (informationnelles, intentionnelles, ayant une mise en forme visuelle, ou autres)* et d'autre part *d'établir des relations entre ces unités*. Cette reconnaissance de la cohérence peut nécessiter des connaissances sémantiques et pragmatiques, non disponibles dans le texte. Néanmoins nous partons du postulat qu'il est possible de mettre en place des analyses automatiques à partir des indices du discours (chaînes lexicales, connecteurs, introducteurs de cadres, etc.) pour permettre de reconnaître cette cohérence.

L'une des caractéristiques majeure transversale à la plupart des théories du discours est la considération d'un modèle de dépendance qui définit la nature de la relation structurelle existante entre deux énoncés en terme de *subordination* ou de *coordination* (Mann & Thompson, 1987; Polanyi, 1988; Virbel, 1989; Asher & Lascarides, 1994). Les différences entre théories viennent de la signification qu'elles donnent à la nature de ces relations, mais aussi des contraintes structurelles d'assemblage des unités discursives. Au niveau de la micro-structure, les chercheurs ont tendance à considérer que l'unité élémentaire de référence est proche de celle de la proposition (Mann & Thompson, 1987; Polanyi, 1988). Afin de faciliter le repérage automatique, nous considérons comme Choi (2002) la phrase syntaxique comme unité élémentaire.

Suivant le genre de texte considéré (expositif, narratif, dialogue, etc.), les théories du discours mettent en évidence un ou plusieurs plans d'organisation de l'information: rhétorique, logico-visuelle, informationnelle, etc. Les interactions entre ces différentes structures sont encore très floues, c'est pourquoi nous avons décidé de nous concentrer sur le plan informationnel que nous considérons comme pertinent pour les textes scientifiques. Notre description du plan informationnel repose sur la théorie de la RST³ (Mann & Thompson, 1987), le LDM (Polanyi, 1988), et aussi la progression thématique de la phrase au discours (Kruijff-Korbayová & Kruijff, 1996). Globalement cela signifie que la relation entre deux énoncés est déterminée en fonction de leur contenu informationnel indépendamment de l'intention rhétorique de l'auteur.

Dans notre modèle, un énoncé entrant se rattache au discours selon une relation de subordination ou de coordination (ou bien les deux). Un énoncé est interprété en fonction de son thème (ce dont il parle), de son propos (ce qui est dit au sujet du thème) et de sa fonction sémantico-rhétorique. Ces éléments sont identifiés à partir d'indices présents dans l'énoncé et dans son contexte ce qui permet de déduire avec quelles parties du discours il est lié et comment.

Nous illustrons ces relations à l'aide du texte de la figure 1 extrait d'un passage de notre corpus. Les indices discursifs aisément représentables visuellement sont soulignés dans le texte. Les couples d'énoncés (1, 2) et (1, 6) décrivent des relations de subordination. 1 est un modèle classique d'annonce thématique avec un quantifieur *two*, une phrase syntaxiquement incomplète et un caractère de ponctuation annonce " : ". Les énoncés 2 et 6, quant à eux, contiennent des marques qui caractérisent des items d'une énumération ("1." et "2."). Ces deux énoncés présentent aussi une relation de coordination explicite l'un envers l'autre, soulignée notamment par un parallélisme syntaxique :

NUM. NOM, *whereby* ADJ NOM be+conj='présent' VERB+conj='participe passé' PREP

Le couple d'énoncés (2, 3) constitue un exemple de subordination où le deuxième énoncé 3 correspond au développement d'un des aspects du premier. Cette subordination est marquée par une progression thématique de rhème en thème (i.e. le terme *importance* qui est repris dans 3). Le couple d'énoncés (4, 5) décrit, quant à lui, un exemple de coordination implicite.

³Rhetorical Structure Theory (RST), Linguistic Discourse Model (LDM).

- Two traditional approaches to automatic abstracting are: (1)
1. Extraction, whereby specific sentences are selected from the source text according to some assessment of their *importance*. (2)
- Importance* indicators include the concentration of topic-relevant terms [...]; the occurrence of expressions, such as "important", "to sum up"; and the position of the sentence within the text. (3)
- This approach is exemplified by Pollock and Zamora's ADAM system [1]. (4)
- The problems with this approach are that importance clues are often not reliable, and that the extracted sentences do not always constitute a coherent text, since they often contain cross-references. (5)
2. Summarisation, whereby detailed semantic analysis is applied to the text, and a representation such as a semantic net is produced, from which a summary is then generated. (6)

Figure 1: Exemples de relations de subordination, et de coordination explicite et implicite

En effet, tous les deux sont subordonnés à une même entité, *l'approche en terme d'extraction*, et chacun d'eux en traite de manière indépendante, le premier en présentant un exemple et le second en décrivant les problèmes.

3 Algorithme de structuration “*shift and reduce*”

La structure de texte en arbre unique est une simplification de la réalité, néanmoins nous adoptons une modélisation *hiérarchique* parce qu'elle reste la plus communément rencontrée dans les textes. Notre algorithme de structuration reprend le principe des algorithmes de Marcu (1999a) et Choi (2002). Nous l'avons adapté afin de tenir compte de la relation de coordination. Cet algorithme construit une structure hiérarchique du discours dont les arcs sont orientées vers les énoncés entrants toujours attachés sur la frontière droite de l'arbre. Un énoncé entrant coordonné à un énoncé de la structure est considéré comme étant subordonné au même énoncé que l'énoncé auquel il est coordonné. Un nœud factice joue le rôle de père de tous les nœuds.

L'algorithme utilise deux structures de données : une pile qui stocke la branche “frontière droite” de l'arbre en cours de construction (le dernier élément empilé est le point d'attache le plus prioritaire), et une file qui contient la liste des énoncés tels qu'ils sont ordonnés dans le texte et analysés successivement. La pile joue un rôle de mémoire dont chaque élément correspond à une granularité inférieure obtenue dans la structure du discours. L'objectif est d'identifier les énoncés qui sont liés et les relations qu'ils entretiennent.

Algorithme :

1. Si la pile est vide, on défile la file et empile la pile (état initial).
2. Tant que la pile et la file ne sont pas vides, calcul de la relation entre l'élément au sommet de la pile et le premier élément de la file.
 - Si une relation de subordination est détectée, alors l'élément de la file est défilé et empilé (on descend dans la granularité du texte) ;
 - Si une relation de coordination est détectée, alors l'élément au sommet de la pile est défilé et remplacé par l'élément de la file ;
 - Sinon (aucune relation) l'élément au sommet de la pile est défilé et écarté (l'idée étant de remonter jusqu'au niveau de dépendance lié à l'élément en tête de file).

4 Indices discursifs

La reconnaissance des relations discursives entre deux énoncés est fondée sur la présence d'indices significatifs dans les textes scientifiques : relations lexicales, expressions clefs et parallélisme de construction.

4.1 Relations lexicales

Les relations lexicales entre deux énoncés sont envisagées selon leur nature sémantique et selon les parties des énoncés concernées (thème ou rhème). Nous utilisons un module de Construction de Chaînes Lexicales (CCL) pour les repérer. Celui-ci est fondé sur une variante de l'algorithme de (Barzilay & Elhadad, 1997). CCL recherche les relations entre les lemmes associés aux paires de mots étudiés en tenant compte de la distance sémantique entre ces mots ainsi que de leur distance dans le texte. Le mot le plus fréquent au sein d'une chaîne constitue son élément représentatif. Nous considérons :

- Les relations morphologiques : deux mots appartenant à la même famille morphologique⁴, indépendamment de leur catégorie lexicale ;
- Les relations utilisées pour référer à un même objet du discours, telles que la synonymie, l'hyponymie et l'hyperonymie, la méronymie et l'holonymie, trouvées dans WordNet ;
- Les relations d'antonymie trouvées grâce à WordNet ou à la présence de préfixes tels que *dis-*, *in-*, *un-*, *non-*, *under-*, *im-*, *a-*, *de-*, *ir-*, *anti-* sur les mêmes lemmes. Nous construisons des chaînes lexicales spécifiques à ce type de relation.

Etant données deux phrases constituant le contexte d'études, des chaînes lexicales sont calculées entre les deux phrases, globalement, et entre les différentes combinaisons des parties constituant le thème et le rhème des deux phrases. La distinction entre les parties thématique et rhématique d'une phrase est réalisée selon une heuristique robuste de découpage de la phrase en deux par rapport au verbe le plus proche de son milieu.

La présence d'un lien lexical entre les rhèmes de deux énoncés traduit généralement une subordination du deuxième énoncé vis-à-vis de l'énoncé précédent (e.g. une élaboration ou une reformulation). Une progression linéaire, de rhème en thème, correspond aussi au même type de subordination (e.g. une annonce thématique). Une relation contrastive peut dénoter une coordination. Dans tous les autres cas, la présence ou l'absence d'une relation lexicale constitue un indice supplémentaire qui pourra se combiner avec les suivants.

4.2 Expressions clefs (essentiellement des connecteurs)

Notre liste d'expressions clefs est issue en partie de la liste de méta-descripteurs acquise automatiquement par Hernandez & Grau (2003), et de l'analyse de notre corpus. Nous l'avons aussi complétée à partir des mots clefs fournis par Choi (2002) pour lesquels nous réassignons la relation (subordination ou coordination) en fonction de nos observations personnelles.

En raison du nombre d'exemples que compte notre corpus (1190 couples de phrases) par rapport au nombre de marques que nous avons retenu (178), nous n'avons pas choisi de considérer chacune des marques comme une caractéristique distincte, au contraire de Choi dont le modèle

⁴Nous utilisons la base CELEX (www ldc.upenn.edu/readme_files/celex.readme.html).

compte 29 marques. Nous avons opté pour une pré-classification de celles-ci en 5 classes en fonction de leur comportement pour structurer le discours et réduit ainsi la complexité du nombre d'indices. Les classes rassemblent des marques ayant un même "comportement" structurel vis-à-vis de la subordination et de la coordination entre deux énoncés. Les classes que nous avons définies sont les suivantes :

- **Initie** : marque le premier item d'une liste d'items (suppose une coordination) : "*former, first, on the one hand, 'l.', 'a)', begin, start*" ;
- **Continue** : Coordonne mais n'initie pas une liste et n'en termine pas forcément une : "*second, another, other, also, and, or, however, but, then, in addition, although, etc.*" ;
- **Termine** : Marque le dernier item d'une liste (suppose une coordination) : "*on the other hand, last, finally, to conclude, to sum up, end, finish, latter, in conclusion, result*" ;
- **Subordonné** : Apparaît en début d'un énoncé subordonné : "*so, the, this, these, it, he, by this, consequently, for example, for instance, therefore, thus, note that, such, etc.*" ;
- **Subordonnant** : Apparaît en fin d'un énoncé subordonnant (i.e. en général une annonce) : "*such as, follow, as follow, see below, below, and, :, ?, etc.*" .

Une marque est dans une seule classe sauf si cette dernière permet de la distinguer dans sa définition (e.g. la position dans la phrase pour différencier les marques *subordonnées* des marques *subordonnantes*). Les notions de début et de fin sont relatives à chaque énoncé et correspondent à une distance en nombre de mots exprimée en pourcentage (respectivement fixée à 40% du début ou de la fin). La taille maximale est fixée à 10 tokens.

En plus de ces classes de marques discursives, nous rajoutons une classe de marques désignant la négation (e.g. *aren't, can't, nothing, nobody, rarely, etc.*) et afin de prendre en compte les formes passives, et l'inversion des parties thématiques et rhématiques qui en découle, nous considérons la présence du verbe "être" suivi directement d'un autre verbe comme une caractéristique. Par la suite, nous appellerons ces dernières caractéristiques les *indices syntaxiques*. Au final, la distribution de nos 178 marques se répartit ainsi : 7 marques pour la classe *Initie*, 38 pour la classe *Continue*, 9 pour la classe *Termine*, 62 pour la classe *Subordonnée*, 30 pour la classe *Subordonnant*, 31 marques de négation et 1 marque du verbe *être*.

4.3 Parallélisme

Le parallélisme de construction entre deux énoncés rend compte d'une importance égale (lien de coordination) (Hernandez, 2004). Il se traduit par a) des similarités des constituants à différents niveaux paradigmatiques (lemme, trait sémantique, catégorie grammaticale, fonction syntaxique) ; b) une similarité syntagmatique qui s'exprime à la fois par une similarité dans l'ordre des constituants parallèles et par une similarité dans les écarts de distance entre ces mêmes constituants.

Afin de calculer le *degré de parallélisme* entre deux énoncés, nous réduisons la complexité du problème d'abord en homogénéisant les entités du discours (chaque mot est remplacé par l'élément représentatif de la chaîne lexicale à laquelle il appartient). Ensuite, chaque structure syntaxique hiérarchique est remplacée par une liste plate, qui correspond à une notation préfixée de l'arbre (les nœuds internes, qui sont des étiquettes, sont placés avant les feuilles, qui sont les lemmes). Cette liste est obtenue à partir du résultat d'analyse fourni par l'analyseur statistique de Charniak (1997)⁵, en supprimant les niveaux de parenthèses.

⁵Nous utilisons la version 2001, développée pour l'anglais à l'université de Brown.

Pour tout couple de phrases donné, le système calcule un degré de parallélisme entre toutes les séquences extraites de chacune des phrases, comportant le même nombre d'items similaires, au minimum deux, différents ou non, placés dans leur ordre d'apparition dans les phrases. Par exemple, les phrases *cabcad* et *acba* partagent 4 constituants : *c*, *a* deux fois et *b*. Une fois supprimés les constituants non similaires (i.e. *d*), on extrait de la première phrase *caba* et *abca*, et de la deuxième *acba*. On ne tient pas compte des éléments différents, qui peuvent être insérés n'importe où dans les phrases. Le parallélisme est fondé sur des constructions similaires d'éléments similaires. La mesure que nous avons définie s'inspire des mesures de distances d'édition entre des séquences de caractères. Chaque constituant est identifié de manière unique par sa position. Plus un constituant est distant de son symétrique dans l'autre séquence, plus les séquences comparées diffèrent. Elle est définie par la formule suivante :

$$\text{degreDeParallelisme}(s_m, s_n) = \sum_{i=1}^{l(s)} \left(\frac{D(s) - d(x_i)}{D(s)} \right)$$

avec x_i , le i^{eme} constituant de la séquence s_m , $l(s)$, la longueur des séquences comparées, $D(s)$, la distance maximale possible entre un constituant d'une séquence s et son constituant parallèle i.e. $D(s) = l(s) - 1$, et d , la distance effective d'un constituant courant de la séquence s_m et son constituant parallèle. Le degré de parallélisme d'un couple d'énoncés correspond au degré maximal obtenu pour les séquences extraites de ces énoncés.

Nous évaluons cette mesure à l'aide d'une approche de base qui calcule simplement le nombre de mots communs entre deux énoncés.

5 Apprentissage des relations discursives

Afin de reconnaître les relations discursives, nous avons décidé d'opter, de même que Marcu (1999a), pour un apprentissage par arbre de décision qui possède l'avantage d'être compréhensible par tout utilisateur et d'avoir une traduction immédiate en terme de règles de décision. Nous avons utilisé le classifieur C4.5 fourni dans le logiciel WEKA⁶.

5.1 Données

Afin de constituer un ensemble de couples de phrases et de relations correspondantes, nous avons manuellement annoté un corpus de 5 documents anglais appartenant au domaine de la linguistique informatique. Ils font tous entre 8 et 10 pages et sont au format pdf. L'un d'eux est en simple colonne. De fait ils couvrent la période 1998 et 1999 et aucun d'eux ne partage de références communes. Ces articles sont Mitkov (COLING-ACL'98), Kan et al. (WVLC'98), Green (ACL'98), Sanderson et al. (SIGIR'99) et Oakes et al. (IRSG'99).

L'annotation a consisté à indiquer pour chaque phrase du texte les relations de subordination et de coordination explicite existantes avec une phrase se trouvant en amont dans le texte ; ces deux types de relations pouvant exister pour une même phrase. Le principe de dépendance que nous avons suivi consiste à toujours resituer un énoncé vis-à-vis de la thématique globale puis d'analyser si localement il n'y a pas des dépendances plus fortes. Chaque couple d'énoncés que nous avons liés est décrit par une décision, D , concernant le type de relation qui les unit. Ces

⁶Cette boîte à outils est disponible à l'URL suivante www.cs.waikato.ac.nz/ml/weka.

couples sont ensuite représentés par l'ensemble des caractéristiques discursives, C , que nous avons précédemment définies. Sur un total de 1038 phrases⁷, 1190 couples exemples, (C, D) , ont été constitués. Ils se répartissent en 632 couples liées par une relation de subordination, 285 instances de coordination et 273 instances décrivant une absence de relation. Les instances décrivant une absence de relation ont été engendrées automatiquement en considérant les couples d'énoncés contigus ne possédant pas de relation entre eux. En comparaison Choi utilise un corpus d'apprentissage de 754 couples de phrases et Marcu utilise un corpus de 99360 mots (le notre est de 25000 mots environ soit un rapport d'1/4).

Nos couples exemples sont décrits automatiquement par les 22 caractéristiques présentées à la section 4, à savoir : la présence de cohésion lexicale entre les énoncés (1 caract.), les patrons de progression thématique (4), l'antonymie (1), la présence des connecteurs selon nos 5 classes et selon l'énoncé du couple (2*5), la présence du verbe être (2*1), de la négation (2*1), le degré de parallélisme (1) et le nombre de mots communs (1). En général, malgré des familles d'indices communes (relation lexicale, connecteur, typographie), on ne retrouve pas exactement les mêmes caractéristiques d'un système à un autre. Le modèle de Marcu compte 2789 caractéristiques (parmi elles, les relations sémantiques de WordNet, des indices contextuels et structurels). Le modèle de Choi compte un indice de cohésion lexicale, des connecteurs et des indices de typographie.

5.2 Résultats

De part la quantité de nos données d'apprentissage, nous adoptons une technique d'évaluation par validation croisée sur 10 partitions. Son principe consiste à partitionner le corpus d'apprentissage en un certain nombre de parts égales et d'utiliser tour à tour une partie comme ensemble d'exemples de test et les autres comme ensemble d'exemples d'entraînement. La moyenne des taux d'erreur correspond au *taux d'erreur global*.

Nous avons réalisé deux jeux d'expérience : le premier en considérant toutes les relations de notre modèle (subordination, coordination et absence de relation), le deuxième en ne considérant plus que la relation de subordination et l'absence de relation. Ce dernier jeu d'expériences nous permet de comparer nos résultats avec ceux de Choi (2002). Pour simplifier la présentation de ces jeux d'expériences par la suite, nous omettrons la relation "absence de relation" dans leur désignation. Pour chacun des jeux nous proposons de comparer les résultats sur deux familles d'indices auxquelles on ajoute tour à tour telle ou telle caractéristique pour observer les améliorations éventuelles. Les performances des deux familles combinées sont aussi considérées. Ces deux *familles d'indices* sont : 1) les caractéristiques décrivant la *progression thématique* (de thème en thème, de thème en rhème, de rhème en thème et de rhème en rhème) ; 2) les caractéristiques fondées sur les *expressions clefs* : les classes Initie, Termine, Continue, Subordonné et Subordonnant. Les caractéristiques individuelles que nous ajoutons sont : les liens lexicaux autre que d'antonymie (appelés par la suite "cohésion lexicale"), les liens lexicaux d'antonymie, les indices syntaxiques (*be* et négation), le degré de parallélisme et le nombre de mots communs.

Afin de positionner l'apport des différents apprentissages, nous comparons leur performance vis-à-vis d'une *approche de base* qui correspond à la prédiction de la classe majoritaire dans le corpus d'apprentissage (c'est-à-dire qu'elle correspond au taux d'erreur si l'on assigne tous les exemples à cette classe).

⁷Les phrases ont été détectées à l'aide des caractères de ponctuation puis corrigées manuellement.

La table 1 décrit les résultats que nous obtenons respectivement lorsque l'on considère les relations de coordination et de subordination, puis lorsque l'on ne considère plus que la relation de subordination. Les valeurs en gras correspondent à des précisions maximales.

coordination_et_subordination : approche de base de 53,10%

		Familles d'indices		Familles combinées
		Progression thématique	Expressions clefs	
Ensemble de base		52,68%	57,31%	56,13%
Caractéristique ajoutée	<i>cohésion lexicale</i>	52,68%	57,31%	57,05%
	<i>antonymie</i>	52,43%	56,89%	55,79%
	<i>indices syntaxiques</i>	54,70%	58,57%	55,71%
	<i># de mots communs</i>	52,43%	57,31%	56,47%
	<i>degré de parallélisme</i>	52,43%	56,97%	55,12%
Toutes les caractéristiques		54,62%	57,05%	55,21%

seulement_la_subordination : approche de base de 69,83%

Ensemble de base		69,83%	73,14%	73,59%
Caractéristique ajoutée	<i>cohésion lexicale</i>	69,83%	72,26%	73,70%
	<i>antonymie</i>	69,83%	73,14%	73,70%
	<i>indices syntaxiques</i>	72,48%	76,35%	75,02%
	<i># de mots communs</i>	69,83%	72,81%	74,03%
	<i>degré de parallélisme</i>	69,83%	73,59%	74,25%
Toutes les caractéristiques		70,16%	75,13%	75,02%

Table 1: Précisions de DST dans la prédiction de relation

Par rapport aux approches de base, les meilleurs sous-ensembles de caractéristiques augmentent la précision de plus de 5% pour chacun des jeux d'expériences. Par rapport à l'existant, nous obtenons *des résultats supérieurs à ceux obtenus par Choi* dans des configurations expérimentales similaires : 76,35% contre 73,61% pour nos meilleures performances de précision respectives. Les résultats rapportées par Marcu (1999a) sont assez similaires à certaines de nos évaluations (autour de 60%) mais nos objets d'apprentissage diffèrent. Le système de Marcu se situe à un degré supérieur de complexité dans le sens où il cherche à reconnaître l'opération de structuration à réaliser en fonction du contexte et de la configuration structurelle en cours. Kruijff-Korbayová & Kruijff et Polanyi *et al.* ne fournissent pas d'évaluation de leur système.

En terme de comparaison de familles d'indices, *les meilleurs résultats que nous obtenons le sont à partir de caractéristiques fondées sur les expressions clefs*. Ces expériences mettent en avant la pertinence des classes discursives que nous avons définies pour factoriser les connecteurs. Par ailleurs, *les résultats obtenus avec les caractéristiques fondées sur des liens lexicaux quels qu'ils soient, combinées ou non, sont bien en dessous de ceux que l'on pouvait espérer*. Ces résultats montrent que des heuristiques de repérage robustes ne suffisent pas pour repérer des phénomènes discursifs tels que la progression thématique.

Enfin, *notre évaluation du "degré de parallélisme" par rapport au "nombre de mots pleins communs" n'apparaît pas comme concluant et requiert d'être approfondie*. Alors que le degré de parallélisme apporte une progression de la précision supérieure pour l'expérience de la détection de la subordination seule, il n'en est pas le cas pour l'autre expérience.

Nos expérimentations constituent une évaluation globale de notre système. Les résultats sont très encourageants et montrent les aspects de la structuration où nous devons travailler.

6 Conclusion

En considérant en plus la relation de coordination, notre approche enrichit le modèle de Choi (2002), ce qui nous permet de modéliser plus finement le discours. En distinction avec Marcu (1999a), le fait de dissocier le modèle de dépendance de la structuration nous permet de fixer indépendamment les contraintes de structuration, et par là d’appréhender plus largement différentes structures du discours.

En utilisant l’algorithme “*shift and reduce*”, nous obtenons une structure hiérarchique proche de celle décrite par une analyse RST. La différence majeure concerne la nucléarité des relations unissant les énoncés.

Parmi nos perspectives nous envisageons d’enrichir notre modèle avec la relation de subordination dirigée vers l’aval du texte, ainsi que de nouveaux indices (comme ceux de mis en forme visuelle) qu’ils se trouvent dans les énoncés considérés ou dans leur contexte.

Références

- Asher N., Lascarides A. (1994), Intentions and information in discourse.
- Barzilay R., Elhadad M. (1997), Using lexical chains for text summarization, In *ACL’97/EACL’97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain.
- Charniak E. (1997), Statistical parsing with a context-free grammar and word statistics, In *Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI*, Menlo Park, MIT Press.
- Choi F. Y., (2002), *Content-based Text Navigation*, PhD thesis, Computer Science, Univ. of Manchester.
- Couto J., Ferret O., Grau B., Hernandez N., Jackiewicz A., Minel J-L., Porhiel S., (2004), RÉgal, un système pour la visualisation sélective de documents, *La présentation d’information sur mesure, Numéro Spécial de RIA*, pages 481–514.
- Hearst M. A., (1997), Texttiling: Segmenting text into multi-paragraph subtopic passages, *Computational Linguistics*, 23(1):33–64.
- Hernandez N., Grau B., (2003), Automatic extraction of meta-descriptors for text description, In *RANLP*, Borovets, Bulgaria.
- Hernandez N., (2003), Un indice de structuration de texte combinant finesse et disponibilité au niveau global et local, In *ATALA*, La Rochelle, France. (1996),
- Kruijff-Korbayová I., Kruijff G-J. M., Identification of topic-focus chains, In S. Botley, J. Glass, T. McEnery, et A. Wilson, editors, *DAARC96*, volume 8, pages 165–179.
- Mann W. C., Thompson S. A., (1987), Rhetorical structure theory: A theory of text organisation, Technical report isi/rs-87-190, Information Sciences Intitute.
- Marcu D., (1999), A decision-based approach to rhetorical parsing, In *The 37th Annual Meeting of the Association for Computational Linguistics (ACL’99)*, pages 365–372, Maryland.
- Moens M.-F., De Busser R., (2001), Generic topic segmentation of document texts, In *ACM SIGIR*, pages 418–419, New York.
- Polanyi L., Culy C., van den Berg M., Thione G. L., Ahn D., (2004), A rule based approach to discourse parsing, In *5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL*, Boston, USA.
- Polanyi L., (1988), A formal model of the structure of discourse, *Journal of Pragmatics*, 12:601–638.
- Virbel J., (1989), The contribution of linguistic knowledge to the interpretation of text structure, In J. André, V. Quint, et R. Furuta, editors, *Structured Documents*, pages 161–181. Cambridge University.