

L'identification de la macrostructure des textes : computation et cognition

Claude St-Jacques

Groupe de technologies langagières interactives
Centre National de Recherche du Canada
Gatineau, (Québec) K1A 0R6
Claude.St-Jacques@nrc-cnrc.gc.ca

Lise Duquette

Institut des langues secondes
Université d'Ottawa
Ottawa, (Ontario) K1N 6N5
lduquett@uottawa.ca

Résumé

L'identification de la macrostructure d'un texte facilite, auprès des lecteurs experts, la rétention des idées principales du scripteur et donc la compréhension générale. La linguistique computationnelle peut-elle, à cet égard, appuyer le travail des didacticiens de l'ELAO (Enseignement des Langues Assisté par Ordinateur) ? Dans cet article, nous présentons les diverses étapes de nos travaux exploratoires visant l'informatisation de cette tâche afin de fournir, aux lecteurs en français langue seconde, un appui pédagogique.

1 Introduction

Une équipe pluridisciplinaire¹, composée d'une didacticienne de la langue, d'un psycholinguiste, d'un spécialiste en mesure et évaluation, de lexicographes, d'un logicien et d'ingénieurs de la langue travaillent à l'élaboration de *DidaLect*, un didacticiel «intelligent» pour l'auto-apprentissage de la lecture en français langue seconde. Dans un contexte d'apprentissage en ligne, le public cible de ce didacticiel est un apprenant ayant un niveau de compétence intermédiaire ou avancé en français langue seconde ou étrangère². Ainsi, parmi les volets de recherche étudiés dans le didacticiel de lecture (dorénavant *DidaLect*), on peut citer

l'harmonisation entre le niveau de difficulté des textes et le niveau de compétence du lecteur, le classement des questions de compréhension de manière à s'assurer qu'il y ait pour chacun des textes des questions explicites et implicites, ainsi que la correction automatique des réponses ouvertes aux questions de compréhension. En effet, il s'agit de déterminer dans quel(s) contexte(s) les réponses aux questions peuvent être corrigées par le système.

Dans cet article, nous allons étudier le moyen d'aider un apprenant à reconnaître la structure des textes pour en mieux saisir les idées importantes. Notre hypothèse de travail est basée sur la théorie de Meyer et al. (1980), considérant qu'un apprenant qui est capable d'identifier la macrostructure d'un texte a une meilleure rétention des idées exprimées par le scripteur. Ainsi, nous présentons les diverses étapes de nos travaux exploratoires visant la pédagogisation informatisée de l'auto-apprentissage de cet aspect d'une compétence linguistique à acquérir afin, qu'éventuellement, le système puisse appuyer automatiquement l'apprenant dans cette tâche.

Dans le cadre de l'élaboration du *DidaLect*, l'équipe de recherche a privilégié certains appuis à la lecture étant donné que les apprenants doivent apprendre dans un contexte d'autonomie. D'abord, il y a le dictionnaire d'apprentissage pour le français langue étrangère et seconde (DAFLES³) qui, grâce à l'enrichissement des connaissances lexicales, facilite la compréhension des textes⁴. De plus, il y a les types de questions de compréhension qui, comme elles sont soit textuelles explicites, textuelles implicites ou scripts implicites, permettent à l'apprenant d'évaluer son

¹ Cette étude s'inscrit dans le projet d'élaboration de *DidaLect* qui bénéficie d'une subvention du CRSH dans le cadre du programme l'Initiative de la Nouvelle Économie (no 820-2001-00130) dans laquelle Lise Duquette agit à titre de chercheure principale.

² Pour avoir accès au programme *DidaLect*, il faut que l'apprenant puisse se classer au niveau intermédiaire ou avancé dans le test de classement French CAPT (cf. Laurier, M. (1993). *L'informatisation d'un test de classement en langue seconde*. Québec, Centre Internationale de Recherche en Aménagement Linguistique (CIRAL/ICRLP). Publication b-190.

³ www.kuleuven.ac.be/dafles

⁴ Il y a consensus chez les spécialistes en apprentissage du vocabulaire à l'effet qu'il y ait une corrélation entre le nombre de mots connus et la compréhension générale d'un texte (voir Duquette 1993).

niveau de compréhension des idées importantes de chaque texte (Graesser et al., 1994). Enfin, il y a l'importance du schème rhétorique de la macrostructure d'un texte (Meyer, 1975) qui permet, entre autres, l'identification de l'idée générale et la rédaction du résumé du texte⁵.

L'objectif de cet article est de présenter les étapes qui nous ont menés à l'utilisation de l'organisation rhétorique de textes comme stratégie permettant d'aider le lecteur à comprendre les idées exprimées dans des documents authentiques⁶. Nous présentons d'abord dans la prochaine section les choix théoriques qui ont motivé la sélection des textes de *DidaLect*. Ensuite, nous allons décrire la méthodologie utilisée pour automatiser l'identification de la catégorie de chaque texte afin de se servir de la structure rhétorique à des fins pédagogiques. La difficulté rencontrée à catégoriser des textes authentiques à partir des connecteurs devant caractériser leur macrostructure, nous amène à explorer une nomenclature de ces marqueurs basée sur la théorie de la structure rhétorique ou la RST (Mann and Thompson, 1988). À la lumière des résultats que nous obtenons, nous proposons une révision à la théorie des patrons rhétoriques de Meyer afin de l'adapter à la réalité des textes authentiques. Finalement, nous concluons en présentant un exemple de l'instrumentation informatisée que nous nous proposons de construire dans *DidaLect* comme aide à la modélisation pour faciliter la compréhension de textes authentiques.

2 Sélection des textes et travaux antérieurs

Nous avons sélectionné 96 textes informatifs, provenant de différentes revues de la Francophonie, particulièrement de France et du Québec. Nous avons été inspirés, pour la sélection des textes, par la classification de Meyer (1975). Selon le modèle de cette dernière, la structure du contenu d'un texte est organisée d'une manière hiérarchique suivant les relations rhétoriques liant les différentes parties du discours d'un scripteur. Meyer nomme cette relation dominante d'un texte, la macrostructure, qu'elle associe à différents types de texte. Ainsi, nous avons opté, dans notre

sélection, pour quatre des catégories de textes recensées par Meyer (1975): problème-solution (ci-après PS), cause-effet (ci-après CE), comparaison (ci-après C) et description (ci-après D).

Plusieurs études empiriques (Carrell 1984; Goh 1990; McGee 1982; Meyer et al. 1980; Raymond 1993; Richgels et al. 1987) ayant montré que la connaissance de l'organisation d'un texte aide un apprenant à se remémorer les idées exprimées, nous avons voulu, dans un premier temps, identifier automatiquement la catégorie de nos textes authentiques⁷. Avant de présenter le résultat de notre recherche sur la catégorisation des textes dans *DidaLect*, nous tenons à rappeler les résultats de quelques travaux, effectués avant nous par d'autres membres de notre équipe, et qui ont inspiré nos choix méthodologiques.

Une première exploration de la catégorisation automatique de nos textes effectuée par Barrière et Agbago (2002) a montré que l'utilisation simple des connecteurs ne peut suffire à identifier les types de textes authentiques selon la taxonomie de Meyer (1975). En effet, deux essais ont été effectués sur deux groupes de textes à l'aide d'algorithmes de classification automatique d'Apprentissage Machine (Machine Learning). Le premier essai est basé sur 69 textes de 7 catégories (collection, cause-effet, comparaison, description, énumération, hybride, problème-solution) en nombres inégaux et effectué à partir d'une liste de 203 connecteurs regroupés en 35 catégories. Dans le deuxième essai, 48 textes de 4 catégories seulement (CE, C, D, PS) ont été utilisés avec la première liste de connecteurs ainsi qu'une autre de 219 connecteurs regroupés en 29 catégories hiérarchisées en 6 catégories racines. De cette série d'essais, le meilleur classement ne dépasse pas 48 %.

Suite à cette première expérimentation, Hermet et Matwin (2005) ont tenté de modifier la caractérisation des textes en réduisant manuellement la liste des mots signalétiques à seulement 70 marqueurs et en y incluant certains mots de vocabulaire qui apparaissent pertinents (Meyer et al. 1980; Raymond 1993). De plus, ces

⁵ Dans le *Bulletin de Psychologie* (no 371, 1985), J. Caron insiste sur le rôle des marques argumentatives dans le rappel d'un texte.

⁶ Alors que le texte pédagogique est conçu à des fins spécifiques, le document authentique provient de revues ou de livres par exemple et est conçu pour des locuteurs natifs.

⁷ Cette tentative d'automatisation de l'identification de la catégorie d'un texte avait deux objectifs : valider l'étiquette accolée à un texte lors de sa sélection et utiliser ensuite cette identification automatisée de la macrostructure d'un texte pour amener l'apprenant à la découvrir. Cette stratégie devrait faciliter la compréhension en permettant l'identification de l'idée principale et de rédiger le résumé d'un texte.

chercheurs ont utilisé une méthode mathématique permettant de mesurer à la fois l'appartenance et la non appartenance d'un item à une classe afin d'optimiser des résultats pouvant être sujets à ce genre d'erreur possible (ECOC method or Error-Correcting-Output-Code). Malgré tous ces efforts, aucun résultat concluant n'a pu être obtenu encore une fois.

3 La catégorisation des textes

Des résultats de ces premières explorations d'une catégorisation automatisée de textes, nous tirons trois pistes d'analyse : préciser le type de mots signalétiques pouvant marquer la catégorie d'un texte, limiter l'étude à des catégories non hybrides de textes authentiques et explorer les mesures d'appartenance ainsi que l'émergence des règles de mots signalétiques associées à une catégorie.

La première piste débouche sur une myriade d'approches dont il est difficile de distinguer la plus pertinente pour l'identification automatisée de la catégorie d'un texte. En effet, Degand et Sanders (2002) rapportent qu'il n'y a pas consensus chez les psycholinguistes quant au rôle explicite des connecteurs et des marqueurs de cohérence facilitant le processus de lecture d'un texte. Selon eux, certaines études empiriques montrent que la saisie de ces mots signalétiques accélère la lecture du segment qui suit, d'autres soutiennent que les marqueurs de cohérence facilitent plutôt la représentation mentale du texte et amènent une réponse plus rapide à une question de compréhension, alors que d'autres montrent au contraire que les marqueurs linguistiques n'ont pas d'effet véritable sur le rappel de l'information ou sur les réponses à des questions de compréhension et certains vont même jusqu'à affirmer que les connecteurs ont un impact négatif sur la compréhension. L'une des sources de contradiction selon Degand et Sanders (2002) vient du fait que la catégorie des marqueurs linguistiques utilisée est mal définie; parfois on utilise des mots signalant l'importance d'un segment de texte et, à d'autres moments, on a recours aux mots marquant la relation entre deux parties d'un texte.

Faut-il alors, comme le proposent Fry et al. (1984), établir des listes de mots signalétiques faites de connecteurs et de syntagmes marquant dans le discours l'importance de la continuité, du changement de direction, de la séquence, de

l'illustration, de la mise en relief de la causalité ou, se limiter à ce que Moeschler et Reboul (1994) définissent comme étant un connecteur, c'est-à-dire « un foncteur qui a pour argument une paire ordonnée de propositions » (p. 179) ? Les connecteurs linguistiques à eux seuls font l'objet d'une littérature abondante dont il est souvent difficile de distinguer leurs sens exacts. Parfois nommé connecteur pragmatique, sémantique, argumentatif, discursif ou interactif selon la perspective linguistique envisagée, il n'en demeure pas moins pour Touratier (2000) qu'une double distinction peut être faite : les connecteurs sont pris soit dans un sens étroit ou soit dans un sens large. Le premier sens vient de ce qu'on appelle généralement les connecteurs logiques permettant les opérations de conjonction, de disjonction, de conditionnalité et de biconditionnalité entre des segments de discours, que l'on distingue des autres non logiques. Pour Touratier (2000), la grammaire d'aujourd'hui a retenu ce sens étroit des connecteurs non logiques exerçant ce que Moeschler et Reboul (2000, p. 465) appellent la « connexité transphrastique ». Le second sens d'un connecteur est pris dans un sens large incluant à la fois la mise en relation d'une phrase à une autre aussi bien que celle d'« un groupe nominal ou adverbial à un groupe syntaxique » (Krause, 2000, p. 213). Voilà pourquoi « le terme même de « connecteur » n'est pas reconnu par la tradition grammaticale » souligne Guimier (2000, p. 11).

Plutôt que de s'ajuster au dilettantisme d'un jugement anomal, là où même les experts en sémantique grammaticale ne peuvent trancher, nous avons opté pour une définition extensive des mots signalétiques dans un premier temps, quitte à revenir à un sens étroit dans un deuxième. Nous avons donc demandé à un grammairien de recenser tous les mots signalétiques, éléments de mise en relation pouvant servir, dans les 48 textes (CE, C, D, PS) déjà utilisés par nos collègues Barrière et Agbago (2002), à identifier leurs catégories. Ce dernier a recensé pas moins de 1366 marqueurs sémantiques⁸ distincts dans les 48 textes pour un grand total de 28147 occurrences de marques signalétiques. De plus, le grammairien a classé les

⁸ L'usage que nous faisons de l'expression « marqueur sémantique » se situe à la croisée de ce que Guimier (2000) qualifie de « sémantique grammaticale » des connecteurs et de « démarche sémasiologique conduisant à l'identification de leur sens.

connecteurs sous 49 rubriques distinctes (ex. addition, adjonction, anaphore, but).

Notre deuxième piste d'analyse s'impose par elle-même. Afin d'éviter, dans un premier temps, des embûches liées à un chevauchement de catégorie, nous avons d'abord choisi de limiter l'automatisation de l'identification des schèmes rhétoriques à des catégories non hybrides de textes authentiques. Ainsi, nous pouvons nous attendre à l'obtention d'un schème rhétorique typique correspondant à la macrostructure de chaque catégorie de textes. La Figure 1 donne un exemple de ce que nous sommes en droit de nous attendre d'une identification automatique de la macrostructure d'un texte de la catégorie « D », c'est-à-dire une identification de la structuration énonciative d'un texte induite à partir des marqueurs sémantiques recensés.

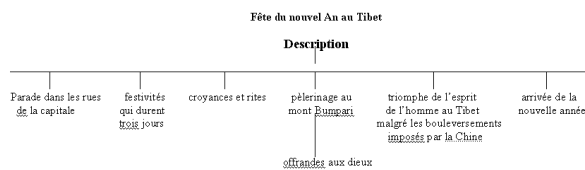


Figure 1. Macrostructure d'un texte

En ce qui concerne notre troisième piste d'analyse, elle s'inspire de l'idée explorée par Hermet et Matwin (2005) d'optimiser la catégorisation des textes en utilisant des modèles computationnels tenant compte à la fois de l'appartenance et de la non appartenance d'un objet à une classe. La logique floue a développé plusieurs outils algorithmiques permettant d'exploiter le degré d'appartenance d'un objet à une classe⁹. Nous utilisons deux des outils de la logique floue : soit un logiciel libre¹⁰ de regroupement flou basé sur la fonction objective de Bezdek (1981) qui permet d'optimiser la partition de données. De plus, nous utilisons un classificateur flou¹¹, aussi un logiciel libre, qui fait émerger les règles ou la fonction caractérisant les données. Ce dernier outil est un amalgame des

⁹ La théorie récente de la « computation par les mots » (Zadeh et Kacprzyk, 1999) de l'auteur de la logique floue Lotfi A. Zadeh a inspirée plusieurs modélisations flexibles telle celle de la sémantique du temps ou de la granularité de l'information en langue naturelle (voir opus cité).

¹⁰ Ce logiciel a été conçu par F. Höppner (2000). Fuzzy Clustering Algorithms - A Tool Library: User's Manual.

¹¹ Nous utilisons le logiciel libre NEFCLASS-J : U. Nauck, Design and implementation of neuro-fuzzy data analysis tool in java. Diplomarbeit, Technische Universität Braunschweig, 1999.

technologies de contrôleur flou et de réseau de neurones flous.

Concernant la préparation des données pour les tests de classification automatique, nous avons procédé comme suit : chacun des 48 textes est représenté par le vecteur de ses marqueurs sémantiques. La première expérimentation a été divisée en deux séries de données. En guise de première série, nous avons utilisé un simple comptage des occurrences des marqueurs pour définir leur poids dans chacun des vecteurs de la matrice. Afin de comparer l'effet d'un regroupement de marqueurs sous un même type de relation, nous avons utilisé, dans une deuxième série des catégories identifiées par notre grammairien, comme attribut dont le poids est déterminé par un comptage des occurrences de ces marqueurs sous la même rubrique.

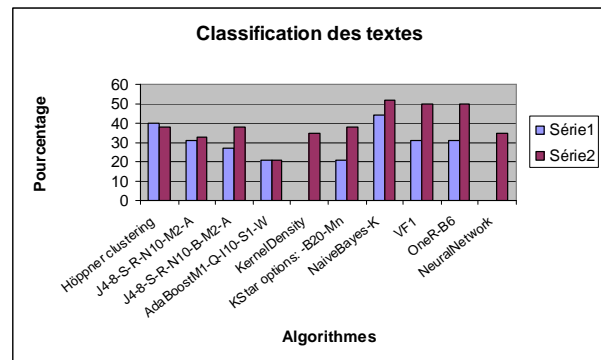


Figure 2. Classification avec 1366 connecteurs

Cette première expérimentation a été menée à l'aide du logiciel de classification Weka 3.3, à partir duquel nous avons fait l'essai de plusieurs algorithmes, ainsi qu'avec le logiciel de regroupement flou de Höppner (2000). La Figure 2 montre les résultats de cette première expérimentation. Le constat est encore très décevant ; pas beaucoup plus que 50 % des textes sont bien classés par trois algorithmes (Naive Bayes, VFI, OneR-B6¹²). Nous pouvons observer, à partir de cette même figure, une classification quelque peu supérieure pour la série (2) de données regroupant les marqueurs en catégories mais rien de véritablement significatif.

Dans une deuxième expérimentation, nous avons voulu vérifier si la longueur des textes

¹² L'algorithme Naive Bayes est une méthode de classification utilisant le théorème de Bayes, celui de VFI (Voting Feature Intervals) utilise une méthode de classification par attribut-discrétisation et l'algorithme OneR (One Rule) est un classificateur basé sur un système de règles.

(variant de 221 à 1712 mots) avait un effet sur le poids des marqueurs dans un texte. Nous avons donc utilisé une troisième série de données où le poids d'une catégorie de marqueurs a été normalisé par le nombre de mots dans un texte. Nous comparons à la Figure 3 la série précédente (2) de données établies par simple regroupement de marqueurs avec la présente série (3) normalisée. Or, nous avons un résultat de classification semblable au précédent sans amélioration de la classification due à la normalisation.

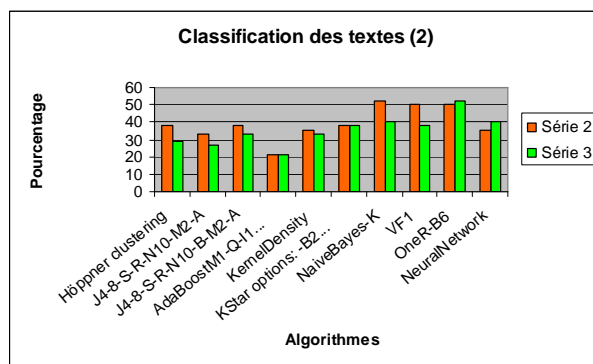


Figure 3. Classification normalisée

À cette étape de notre étude, nous pouvons déjà conclure que l'utilisation des connecteurs pris dans un sens large, incluant des marqueurs mettant en relation des groupes syntaxiques nominaux, adverbiaux ou autres, n'apporte aucune amélioration par rapport aux expérimentations de nos collègues. Par ailleurs, l'utilisation du logiciel Nefclass nous amène à constater que seulement trois règles/classes émergent de ces données alors que l'outil de regroupement flou de Höppner en mode non supervisé produit 3 regroupements plutôt que les 4 identifiés aux types de textes (CE, C, D, PS) du départ. Les informations fournies par le logiciel de regroupement flou à l'effet que dans plusieurs cas le degré d'appartenance d'un texte à un regroupement plutôt qu'à un autre varie à peine, ces constats font surgir deux questions. D'abord se peut-il que deux types de textes se confondent du point de vue des marqueurs sémantiques en un seul ? De plus, la difficulté montrée par l'outil de regroupement flou à partitionner distinctivement deux groupes de textes ne viendrait-elle pas du fait que les textes authentiques n'ont pas toujours une catégorie dominante de marqueur sémantique ? Nous poussons donc un cran plus loin cette recherche pour tenter de répondre à ces questions.

4 Les connecteurs au sens étroit

Afin de vérifier l'hypothèse à l'effet que les textes authentiques, que les didacticiennes ont sélectionnés, soient souvent hybrides, nous avons retenu 12 des 48 textes de départ dont la catégorie semblait la plus évidente et nous avons demandé à deux expertes¹³ en linguistique textuelle de comparer leur identification des textes. Dans 3 cas sur 12 les cojuges ont constaté qu'il n'y avait pas facilement consensus. Toutefois, contrairement à certaines études sur l'organisation des textes, par exemple celles de Meyer (1975) ou Carrell (1984), nous n'avons sélectionné que des documents authentiques provenant des journaux si bien que dans un texte de type C, il y a souvent de la description et dans un texte de type PS, on présente souvent la cause du problème et ses effets avant de présenter la solution.

Dans une ultime tentative de classification automatique, nous faisons un pas en arrière et demandons à l'une de nos expertes linguistes de recenser uniquement les connecteurs interphrastiques dans nos 12 textes à catégorie dominante. Nous avons divisé cette deuxième expérimentation en deux étapes. D'abord dans la première, nous utilisons deux séries de données.

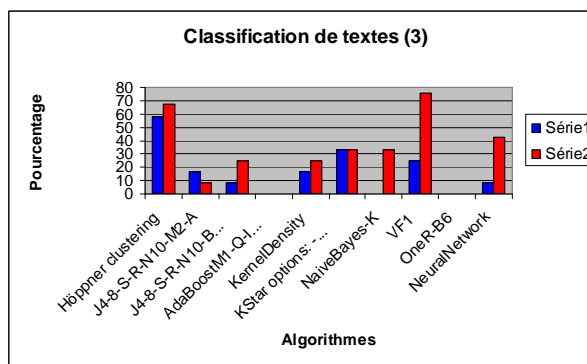


Figure 4. Classification en 4 catégories

Dans la première série de données, les connecteurs sont regroupés suivant les catégories préalablement utilisées par notre grammairien. Puis dans la deuxième série, nous regroupons les connecteurs suivant la définition des relations de la RST de Mann et Thompson (1988) en identifiant contextuellement la nature de cette signalisation.

¹³ Nous tenons à remercier les deux linguistes, France Lemonnier de l'Université Laval, à Québec et Odette Gagnon de l'Université du Québec à Chicoutimi, pour l'aide qu'elles nous ont apportée dans l'analyse de la structure de certains de nos textes.

La Figure 4 nous montre pour la première fois des résultats soulevant un début d'intérêt. Deux algorithmes classifient correctement plus de 60% des textes : 67% de bons résultats pour le « Höppner clustering » et 75% pour le « VFI » en utilisant les connecteurs regroupés suivant les relations définies par la RST.

À la deuxième étape de cette expérimentation, nous avons voulu vérifier l'hypothèse d'un chevauchement des connecteurs pour certaines catégories de textes comme CE et PS ainsi que C et D. Nous avons utilisé les mêmes données que celle de la série 2 de la Figure 4, c'est-à-dire le regroupement des connecteurs selon les relations de la RST¹⁴, mais en forçant une catégorisation des textes en trois et deux classes.

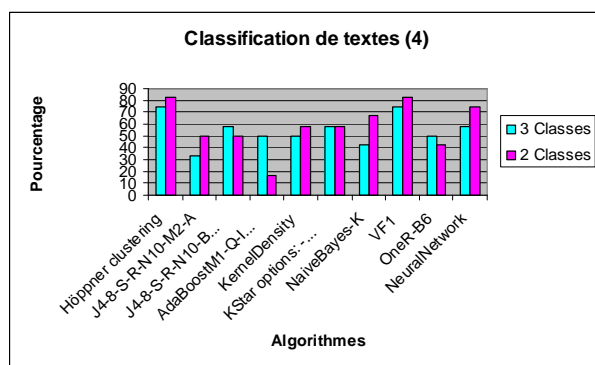


Figure 5. Classification en 2 et 3 catégories

La Figure 5 nous montre que lorsque nous confondons les catégories de textes CE et PS en réduisant à trois notre classification des textes, nous obtenons 75% de bons résultats avec deux algorithmes (Höppner clustering et VFI). Le graphique nous indique ensuite qu'en réduisant à deux classes (CE-PS et C-D), la justesse de la catégorisation augmente à 83% avec deux algorithmes (Höppner clustering et VFI) et à 75% avec un autre (Neural Network¹⁵).

5 Discussion

Nous devons nous rendre à l'évidence que la majorité des textes authentiques sont hybrides, au sens donné par Meyer (1975) aux types de textes de CE, C, D et PS. Ainsi, à partir des exemples que cette dernière fournit dans ses articles, il semble

qu'elle ait effectué ses études uniquement à partir de textes élaborés à des fins pédagogiques. Par ailleurs, lorsque l'identification de la macrostructure des textes se limite à une distinction entre des énoncés performatifs et dénotatifs¹⁶, nous devons nous demander en quoi ce simple constat peut aider le lecteur à se remémorer les idées essentielles d'un texte authentique.

Si la stratégie de Meyer et al. (1980) fonctionne bien pour des textes ayant une signalisation adaptée à la typologie anticipée par cette théorie, nous devons poser certaines restrictions pour les textes authentiques. D'abord, seul un nombre restreint de textes authentiques préalablement identifiés non hybrides peut être utilisé pour l'identification de la macrostructure du texte et pour sa meilleure compréhension puisque la majorité des textes authentiques sont hybrides.

D'une part, le constat le plus important que nous devons faire tient beaucoup plus au fait que nos résultats concernant la signalisation de la sémantique grammaticale pointent beaucoup plus vers une taxonomie énonciative de l'organisation rhétorique comme nous venons de le constater que vers celle d'une typologie de textes (CE, C, D, PS). Lee (2001) arrive à une conclusion semblable lorsqu'il examine à quel aspect du langage nous nous intéressons en catégorisant des textes par genre, registre, domaine, style, type de texte, etc. À son avis, il existe une confusion totale dans l'usage de cette terminologie. Par exemple, ce que Meyer appelle, comme bien d'autres d'ailleurs, des « types de texte », vient en contradiction, au dire de Lee (2001), avec le sens utilisé par les linguistes de ce qu'est la structure rhétorique ou discursive d'un texte.

D'autre part, le concept de « type de texte » est tellement évasif nous dit Lee (2001) que certains comme Meyer l'utilise au sens de la distinction traditionnelle des quatre catégories de discours en rhétorique, soit narratif, descriptif, explicatif et

¹⁴ Selon la théorie de la structure rhétorique (Rhetorical Structure Theory) de Mann et Thompson (1988).

¹⁵ Il s'agit d'un classificateur utilisant un réseau de neurones artificiels à rétropropagation.

¹⁶ Nous reprenons cette distinction d'abord faite par J. L. Austin. 1982. *How to do things with words*. 2nd ed. Oxford University Press, Oxford: New York. Lyotard (1979 ; p. 21) la résume en soulignant que le performatif a pris un sens bien précis en théorie du langage en l'associant à ce qu'il appelle un « rapport de input/output » d'un système par exemple, ce qui se rapproche à notre avis de la cause et de l'effet ainsi que du problème et de la solution. Tandis que l'énoncé dénotatif correspond à ce que Austin préfère appeler le constatif plutôt que descriptif puisqu'il situe, dans un sens plus large, ce constat d'un état de chose dans le monde que nous considérons ici comme une comparaison aussi bien qu'une description.

argumentatif¹⁷. Alors que d'autres réfèrent à ces mêmes catégories comme à des « types de discours » en les nommant alternativement « type de texte » et « genre » (Lee, 2000 ; p. 41).

En somme, la linguistique computationnelle utilise couramment en ingénierie de la langue divers processus permettant d'identifier un sous-ensemble langagier tel que des collocations, des thésaurus spécialisés, des concordances et diverses structures linguistiques. Par ailleurs, la didactique des langues peut bénéficier de ces technologies à condition bien entendu de pouvoir identifier avec le plus de précision possible la nature de l'objet devant être saisi.

6 Conclusion

Rappelons que nous avons comme objectif au départ l'informatisation de la tâche d'identification de la macrostructure d'un texte afin de l'utiliser ensuite pour faciliter la compréhension de l'apprenant. Si l'exercice computationnel que nous venons de mener dans le présent travail ne permet pas de montrer, pour des textes authentiques, l'utilité de l'association faite par Meyer (1975) entre la typologie versus la signalisation d'un texte, celui-ci nous a toutefois permis d'identifier une autre voie. En effet, la théorie de la structure rhétorique de Mann et Thompson (1988) nous permet de répondre à notre besoin en didactique des langues même si cette dernière ne permet pas la catégorisation des textes selon la typologie de Meyer.

À dire vrai, cet exercice de simulation du processus d'apprentissage nous amène à conclure que, ce qui importe pour la rétention du contenu d'un texte par un lecteur, ce n'est pas l'étiquette qu'on lui affuble nommément, mais plutôt ce qui forme le noyau du texte. Il faut accorder à Meyer (1975) le crédit d'avoir reconnu l'importance de la macrostructure d'un texte pour la « pédagogisation » de cet aspect de compétence linguistique à acquérir, qu'est l'identification des

idées principales, à partir d'un schème rhétorique. Cependant, l'organisation de la structure est beaucoup plus expressive sur le plan sémantique lorsqu'on l'examine, comme le proposent Mann et Thompson (1988), à partir des relations liant deux segments de texte.

À la Figure 6, nous donnons un exemple de l'application de la théorie de Mann et Thompson (1988) à l'un des textes de *DidactLect*. Nous utilisons l'outil informatique de O'Donnell (1997) pour montrer comment l'identification de la relation d'Arrière-plan¹⁸ qui lie le segment-noyau, « Dès le début de la grossesse, il est important de surveiller le taux de fer chez la femme » à deux segments-satellites, « Voilà qui est bien connu » et « Pendant la grossesse, rappelle Marta Santuré, étudiante qui s'est intéressée à la question dans le cadre de sa maîtrise en nutrition, le développement du bébé demande beaucoup de fer », permet de reconnaître l'idée centrale exprimée dans cet extrait par ce que la RST identifie comme un noyau.

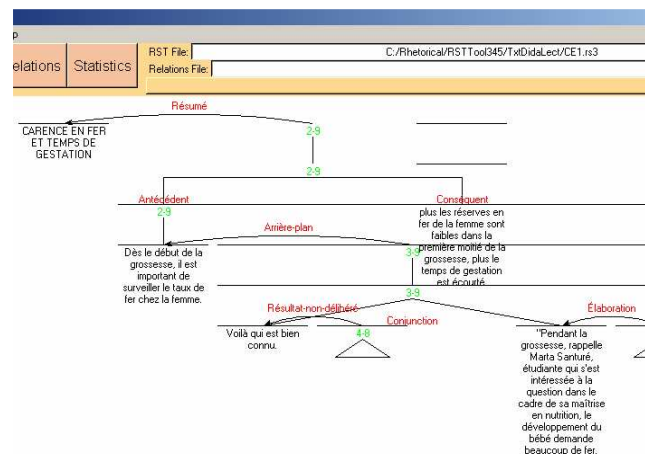


Figure 6. La relation d'Arrière-plan

Dans de futurs travaux, nous comptons explorer l'avenue offerte à l'ELAO par la théorie RST de Mann et Thompson (1988) puisque qu'elle s'avère applicable en linguistique computationnelle à la création automatisée de résumés de texte (Marcu, 2000). Sur le plan cognitif, il n'y a qu'un pas à franchir entre la capacité d'un apprenant à retenir les idées principales d'un texte et celle de résumer un texte. C'est donc dire que si la stratégie computationnelle de Marcu (2000) permet de

¹⁷ Nous soulignons que la typologie des textes varie elle aussi dépendamment si l'on réfère à celle de Egon Werlich. 1976. *A Text Grammar of English*. Quelle und Meyer, Heidelberg ou de Jean-Michel Adam. 1987. *Types de séquences textuelles élémentaires. Pratiques*, 56 (décembre), pp. 54-79 ou autres. La typologie de Werlich distingue 5 types de textes : descriptif, narratif, expositif, argumentatif et instructif. Adam en distingue d'abord 8 types : narratif, descriptif, explicatif, argumentatif, injonctif, prédictif, conversationnel et rhétorique. Récemment, il revoit à la baisse ce nombre pour ne retenir que les 5 types énumérés dans le titre de cet ouvrage : Jean-Michel Adam. 2001. *Les textes : types et prototypes. Récit, description, argumentation, explication et dialogue*. Nathan, Paris.

¹⁸ La relation d'Arrière-plan correspond aux connaissances préalables du lecteur.

résumer un texte à partir des segments jouant le rôle de noyau dans des relations interphrastiques complexes, alors il devient possible d'aider un apprenant à rédiger le résumé d'un texte au moyen de l'identification automatique des noyaux. Les marqueurs sémantiques sont ici réutilisables.

References

- C. Barrière & A. Agbago. 2002. Classification automatique selon quatre catégories du discours. Université d'Ottawa – GRIL. *Rapport*.
- J. C. Bezdek. 1981. *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York, NY.
- P. L. Carrell. 1984. The effects of rhetorical organization on ESL readers. *TESOL Quarterly*, 18(3):441-469.
- L. Degand & T. Sanders. 2002. The impact of relational markers on expository text comprehension in L1 and L2. *Reading and Writing: An Interdisciplinary Journal*, 15(7-8):739-758.
- L. Duquette. 1993. *L'étude de l'apprentissage du vocabulaire en contexte par l'écoute d'un dialogue scénarisé en français langue seconde*. Centre international de recherche en aménagement linguistique (publication B-187), Québec.
- E. B. Fry, D. L. Fountoukidis & J. Kress Polk. 1985. *The new reading teacher's book of lists*. 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.
- S. T. Goh. 1990. The effects of rhetorical organization in expository prose on ESL readers in Singapore. *RELC Journal*, 21:1-13.
- A. C. Graesser, C. L. McMahan & B. K. Johnson. 1994. Question Asking and Answering. In *The Handbook of Psycholinguistics*. ed. Morton A. Gernsbacher. Academic Press, San Diego, CA.
- C. Guimier (éd.). 2000. *Connecteurs et marqueurs de connexions*. Presse Universitaire de Caen, Caen, France.
- M. Hermet & S. Matwin. (à paraître 2005). Classification automatique de textes d'après la catégorie du discours. In *Technologies langagières et apprentissage des langues*. Collectif dir. Lise Duquette et Claude St-Jacques. Les Cahiers Scientifiques de l'ACFAS, Montréal.
- M. Krause. 2000. Vers une grammaire des prépositions, substituts et particules verbales de l'allemand : bilan et perspectives. In *Connecteurs et marqueurs de connexions*. Collectif dir. Claude Guimier. Presse Universitaire de Caen, Caen, France.
- D. YW. Lee. 2001. Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle. *Language, Learning & Technology*, 5(3):37-72.
- J.-F. Lyotard. 1979. *La condition postmoderne*. Les Éditions de Minuit, Paris.
- W. C. Mann & S. A. Thompson. 2001. Deux perspectives sur la théorie de la structure rhétorique (RST). *Verbum*, 1:9-29.
- W. C. Mann & S. A. Thompson. 1988. Rhetorical Structure Theory : Toward a functional theory of text organization. *Text*, 8(3):243-281.
- D. Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.
- L. M. McGee. 1982. Awareness of text structure: Effects on children's recall expository text. *Reading Research Quarterly*, 17:581-590.
- B. J. F. Meyer. 1975. *The organization of prose and its effect on memory*. North Holland Publishing Company, Amsterdam, Netherlands.
- B. J. F. Meyer, David M. Brandt & George J. Bluth. 1980. Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly*, 16:72-103.
- J. Moeschler & A. Reboul. 1994. *Dictionnaire encyclopédique de pragmatique*. Seuil, Paris.
- M. O'Donnell. 1997. RST-Tool: An Analysis Tool. *Proc. 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duisburg, Germany.
- P. M. Raymond. 1993. *The effect of structure strategy training on the recall of expository prose for university students reading French as a second language*. International Center for Research on Language Planning = Centre international de recherche en aménagement linguistique, Québec.
- D. L. Richgels, L. M. McGee, R. G. Loman & C. Sheard. 1987. Awareness of four text structures: Effects on recall of expository text. *Reading Research Quarterly*, 22:177-196.
- C. Touratier. 2000. *La sémantique*. Colin, Paris.
- L. A. Zadeh & Janusz Kacprzyk (eds.). *Computing with Words in Information/Intelligent Systems 1-2*. Heidelberg; New York : Physica-Verlag, 1999.