De l'analyse syntaxique automatique à l'analyse automatique de discours dans les collections multilingues de documents numériques composites

Emmanuel Giguet

De l'analyse syntaxique automatique à l'analyse automatique de discours dans les collections multilingues de documents numériques composites

Emmanuel Giguet

13 septembre 2011

Mémoire d'habilitation à diriger des recherches

Université de Caen Basse-Normandie

Jury :

Jean-Gabriel Ganascia, Professeur d'informatique, LIP6, Université de Paris 6, rapporteur

Éric Gaussier, Professeur d'informatique, LIG, Université de Grenoble 1, rapporteur

Mathieu Valette, Professeur de linguistique, ERTIM, INALCO, rapporteur

Éric Bruillard, Professeur d'informatique, STEF, ENS Cachan / INRP

Patrick Constant, Président et Fondateur de PERTIMM

Pierre Zweigenbaum, Directeur de recherche, LIMSI, CNRS

Nadine Lucas, Chargé de recherche, GREYC, CNRS

Jacques Vergne, Professeur d'informatique, GREYC, Université de Caen Basse-Normandie

Remerciements

Si pour moi la recherche est avant tout un cheminement personnel, c'est une aventure que je n'aurais pleinement su vivre isolé. Aussi ne puis-je commencer ce document sans rendre hommage à toutes celles et ceux qui ont contribué à ma formation, à toutes celles et ceux qui ont eu une influence sur mon parcours scientifique, qui me permettent de vivre intensément ma passion pour les langues.

Je voudrais tout d'abord exprimer ma sincère gratitude envers Jacques Vergne pour son accompagnement et son soutien indéfectible depuis vingt ans. Jacques Vergne m'a enseigné l'analyse syntaxique automatique et le traitement des langues, un traitement des langues atypique à l'écart des grands courants de pensée, un traitement des langues où les relations entre termes priment sur les termes eux-mêmes. Son stoïcisme face à l'évolution des institutions ou l'incompréhension de ses confrères est en soi une leçon. Je ne le remercierai en outre jamais assez de m'avoir présenté Nadine Lucas.

Je remercie spécialement Anne Nicolle qui m'a accueilli au sein de son équipe de recherche et qui a suivi attentivement l'écriture de mon mémoire. Je souhaite témoigner ma reconnaissance aux collègues de l'équipe ISLanD qui ont tous contribué à ma formation par la richesse de nos échanges interdisciplinaires. Les interventions de Pierre Beust, Serge Mauger, Bernard Morand, Luigi Lancieri, Yves Lepage, Éric Bruillard, et de nos doctorants m'ont notamment conforté quant à l'intérêt de ma démarche.

Je remercie également mes collègues du laboratoire GREYC, qu'ils soient algorithmiciens, biométriciens, traiteurs d'images ou fouilleurs de données. Je pense en particulier à Jacques Madelaine pour nos conversations sur l'élégance de la modélisation informatique, à Frédérique Loew-Turbout pour ses explications concernant la création de représentations cartographiques tout aussi esthétiques qu'informatives, à Hervé Le Crosnier pour le partage de sa vision du monde numérique. Mes recherches n'auraient pu connaître leur développement sans un environnement informatique propice à l'expérimentation et un soutien administratif de qualité. À ce titre, je remercie les membres des services informatique et administratif du GREYC.

Je remercie les personnes qui ont suivi mes travaux ou ont une influence sur mon parcours professionnel, souvent au-delà des disciplines et des différences méthodologiques, en particulier Bernard Victorri, Catherine Fuchs, Brigitte Vallée, Christian Boitet, Gérard Sabah, Dominique Dutoit, Jacques Chauché, Pascal Buléon et Pierre Le Goffic.

Je remercie très chaleureusement Jean-Marie Mélé, administrateur de crashdump.net, dont la compétence et la célérité permettent la valorisation de mes travaux sur internet.

Je remercie vivement Hervé Déjean qui a relu avec sagacité mon mémoire, soulevant les questions les plus difficiles pour mieux approfondir la connaissance.

Je voudrais ici relater le privilège d'avoir rencontré Nadine Lucas à Coling en 1992. Enseignante, elle a guidé mes pas vers l'analyse automatique du discours et les traitements multilingues. Collègues, nous avons passé des jours et des nuits à discuter de nos approches, à affiner à la même table, elle, ses modèles linguistiques, moi, mes modèles informatiques, toujours animés par la même nécessité de comprendre. Nous avons partagé la joie des expérimentations réussies et la déception des articles refusés. Puissions-nous longtemps encore ensemble transcender nos connaissances.

Mes sincères remerciements s'adressent à Jean-Gabriel Ganascia, Éric Gaussier et Mathieu Valette pour avoir accepté de rapporter sur ce mémoire d'habilitation dont je ne sais s'il traite d'informatique tout en étant persuadé qu'il ne traite pas de linguistique. Des remerciements tout aussi sincères vont également à Éric Bruillard, Patrick Constant et Pierre Zweigenbaum, pour avoir accepté de participer à mon jury.

Mes remerciements s'adressent enfin à ma famille, à mes parents et tout spécialement à ma femme Laurence, pour leurs encouragements, leur disponibilité, leur compréhension, leur patience et leur confiance. Une tendre pensée vole vers mes filles dont les yeux rieurs et les sourires malicieux ont été pour moi un moteur. Je mesure combien mon indisponibilité pendant la période d'écriture les a affectées.

Table des matières

1 Introduction....7
2 Axes de recherche....9
2.1 La variation en langues....10
2.2 La prise en compte du genre des textes....10
2.3 La gestion des ordres de grandeur de documents....11
3 Synthèse des travaux sur l'alignement automatique multilingue....13
3.1 État de l'art....13
3.2 Émergence de la thématique, encadrements et collaborations....14
3.3 Principales contributions....15
3.3.1 Une approche distributionnelle de l'alignement....15
3.3.2 Une approche multilingue de l'alignement....17
3.3.3 Une approche de l'alignement à grain d'analyse paramétrable....19
3.3.4 Une approche structurelle et typo-dispositionnelle de l'alignement....21
3.4 Méthode d'alignement multilingue proposée....22
3.4.1 Spécificités....22
3.4.2 Éléments d'implémentation....22
3.5 Prospective....23
3.5.1 Vers l'alignement de structures de documents non parallèles....23
3.5.2 Vers la sélection automatique du grain d'analyse et de la fenêtre d'observation....23
3.5.3 Vers un alignement massivement multilingue généralisé....24
3.5.4 Vers un alignement automatique des chaînes de coréférence....24
3.6 Conclusion....25
3.7 Publications liées....26
4 Synthèse des travaux sur la structuration automatique des documents....29
4.1 Enjeux de la structuration automatique des documents....29
4.1.1 Enjeux applicatifs....29
4.1.2 Enjeux scientifiques....30
4.2 État de l'art de la structuration automatique des documents....31
4.3 Émergence et développement de la thématique....32
4.4 Positionnement....33
4.5 Les structures du document considérées....34
4.5.1 La structure physique du document....34
4.5.2 La structure logique du document....34
4.6 Contribution à la structuration physique du document....34
4.7 Contribution à la structuration logique automatique....37
4.8 Prospective....39
4.8.1 Vers un modèle cognitif de la structure logique du document....39
4.8.2 Réflexions sur les format de documents, la perception par le lecteur, et les conséquences pour le calcul de la structure....40
4.9 Conclusion....42
4.10 Publications liées....42
5 Synthèse des travaux sur l'analyse des forums de discussion....43
5.1 Les forums de discussion....43
5.2 État de l'art....44
5.3 Émergence de la thématique....44
5.4 Positionnement....44
5.5 Principales contributions....45
5.5.1 Contribution à la prise en compte des ordres de grandeur....45

5.5.2 Contribution à la prise en compte du style....48
5.5.3 Contribution à la prise en compte de la variation en langues....48
5.5.4 Réflexions sur la prise en compte des ordres de grandeur....48
5.6 Prospective....50
5.6.1 Vers une analyse sémiotique des forums de discussions....50
5.6.2 Réflexions sur la pertinence des représentations synthétiques textuelles....50
5.7 Conclusion....50
5.8 Publications liées....51
6 Les cadres d'expérimentation et de diffusion scientifique....53
6.1 Le cadre d'expérimentation scientifique....53
6.2 Le cadre de diffusion scientifique....53
6.3 Publications liées....54
7 Regard épistémologique sur mes recherches....55
7.1 Le modèle linguistique au centre de la gestion des ordres de grandeur....55
7.2 Le statut incertain de constituants non linguistiques....56
7.3 La délimitation et l'interprétation des constituants opératoires....57
7.4 Les constituants opératoires au centre du modèle opératoire....58
7.5 La sélection des marques pour la structuration en constituants opératoires....59
7.5.1 De la dissymétrie du marquage des frontières gauche et droite....59
7.5.2 De la fiabilité des marques dans la structuration en constituants opératoires....61
7.5.3 De la dissociation des marques frontières et des marques constitutives....61
7.5.4 De la détermination du local par le global....64
7.5.5 Du statut des ressources lexicales dans la structuration en constituants opératoires....65
7.5.6 De l'efficacité des marques dans la structuration en constituants opératoires....67
7.6 Du principe même de structuration fondée sur les constituants opératoires....68
8 Prospective....69
8.1 Retour sur la méthode de structuration ascendante basée sur l'identification manuelle des marques....69
8.2 Vers une méthode de structuration descendante guidée par le modèle et l'induction automatique des marques....69
8.2.1 Une méthode de structuration guidée par le modèle....69
8.2.2 Une méthode reposant sur l'analyse distributionnelle....70
8.2.3 Une méthode systématisant l'analyse distributionnelle....70
8.3 Conclusion....71
Références bibliographiques....73
Publications et travaux encadrés....81
Production logicielle....85

1 Introduction

Sans excellents linguistes, il n'est de traitement des langues qui vaille.

Sans excellents sémioticiens, il n'est de traitement des langues qui vaille.

Le *traitement des langues* est un domaine interdisciplinaire par nature. Il n'appartient ni à la linguistique, ni à la sémiotique, ni à l'informatique. En cela, il ne trouve pleinement sa place dans une structuration de la recherche en grands domaines scientifiques. Ni dans les sciences de l'homme et de la société, ni dans les sciences informatiques.

Appelé *linguistique informatique*, *linguistique computationnelle,* ou *informatique linguistique* au gré des regards disciplinaires, c'est sous la dénomination *traitement des langues* que le domaine trouve un équilibre propre à son développement, conviant les disciplines sur le terrain neutre du rapport à l'objet, les invitant à la collaboration, au-delà de tout clivage. Comme dans tout domaine interdisciplinaire, chaque acteur se doit de conserver sa propre culture tout en s'ouvrant à celle de l'autre, pour que le dialogue s'établisse, pour que la collaboration s'installe, pour que l'union fasse plus que la somme.

Une collaboration interdisciplinaire en traitement des langues ne saurait être qu'une juxtaposition des compétences de chaque discipline. Elle ne saurait réduire la contribution du linguiste à l'annotation de corpus, à la création de lexiques, ou à l'écriture de grammaires, tout comme elle ne saurait cantonner l'informaticien à son savoir-faire ingénierique pour outiller sur commande la linguistique avec des interpréteurs ou des compilateurs de grammaires, des plateformes d'annotation, d'expérimentation, ou encore d'évaluation.

Pour qu'il y ait collaboration fructueuse, il faut qu'il y ait acceptation que les résultats des recherches en informatique, en logique, ou en mathématiques ne sont pas immédiatement applicables à une langue, que les résultats des recherches en linguistique ne sont pas immédiatement transférables vers un traitement automatique, et que ce n'est que dans la cadre d'une relation interdisciplinaire engagée que pourront être mises au point les méthodes d'analyse automatique des langues.

Bien entendu, il convient également d'accepter que des connaissances scolaires en linguistique ne seront jamais suffisantes, et que l'informaticien, le logicien ou le mathématicien doté de compétences de locuteur, doit résister à cette tentation insidieuse d'appréhender l'objet seul, de faire valoir un statut d'expert du domaine pour réclamer une quelconque légitimité à construire seul un système de traitement des langues et ainsi s'affranchir des contraintes et des exigences d'une recherche interdisciplinaire engagée.

S'il est de mon souci de préserver cette relation fragile entre disciplines, je n'ai de cesse de m'interroger sur les forces et les faiblesses des modèles conçus ou utilisés en traitement des langues pour faire face aux véritables défis que représentent aujourd'hui le multilinguisme, la diversité des pratiques sociales, et la nécessaire efficacité des traitements. Par nécessaire efficacité, il convient d'entendre la capacité à gérer des volumes documentaires chaque jour plus importants, mais il faut surtout comprendre la capacité à produire des analyses interprétables, des analyses qui *font sens,* de documents devenus des objets sémiotiques complexes.

C'est au travers d'une réflexion approfondie sur les méthodes de résolution et dans la recherche de solutions innovantes au-delà de l'état de l'art, que je situe mon travail. C'est une recherche collaborative, car menée avec des chercheurs au profil complémentaire partageant une même vision du traitement des langues, notamment au sein de mon équipe de recherche. C'est une recherche originale qui a pour objectif d'améliorer les applications de traitements des langues par l'intermédiaire d'un cadre de résolution plus puissant.

Dans ce document, il s'agit de réaliser la synthèse de mes travaux de recherche. Ce sera le regard de l'informaticien sur son activité de recherche interdisciplinaire en traitement des langues. Il s'agit d'un travail d'ouverture qui a pour ambition d'inviter le lecteur à la réflexion : un travail d'ouverture à la macrostructuration linguistique et sémiotique pour faire face aux enjeux du multilinguisme et des nouveaux usages, un travail d'ouverture méthodologique pour gérer efficacement et au sein d'un cadre opératoire unifié la variation en langues, la prise en compte du genre, et les ordres de grandeur.

Le mémoire débute par une présentation de mes axes de recherches : variation en langues, prise en compte du genre des textes, et gestion des ordres de grandeur. Viennent ensuite trois synthèses de mes travaux les plus récents : l'alignement automatique multilingue en chapitre 3, la structuration automatique du document en chapitre 4, et l'analyse des forums de discussion en chapitre 5. Le chapitre 6 motive la conception de deux plateformes, la première servant de cadre pour mes expérimentations, la seconde pour la diffusion scientifique des résultats obtenus. Le chapitre 7 porte un regard épistémologique sur mes recherches. Le chapitre 8, consacré à la prospective, a pour objet de montrer l'unité de mon travail de recherche et d'en esquisser les suites.

Notes au lecteur :

La linéarité du document, imposée par le support, ne doit pas masquer le fait qu'après lecture des axes de recherche, deux parcours du document sont envisageables, l'un commençant par les synthèses et se poursuivant par le chapitre *Regard épistémologique*, l'autre commençant par le chapitre *Regard épistémologique* et se poursuivant par les synthèses.

Les références bibliographiques apparaissent en fin de document. Les références à mes propres publications et aux publications des étudiants que j'ai encadrés apparaissent en fin de chapitre.

2 Axes de recherche

Mes recherches en traitement des langues portent sur la *structuration du discours inscrit*. Elles ont pour objet la modélisation de la construction du sens dans le cadre d'une interaction médiatisée. Ce travail prend appui sur une entrée tangible et observable, support de l'interaction, en l'occurrence le document numérique. Mon ambition est de proposer une méthode automatique de structuration linguistique cognitivement plausible, c'est-à-dire efficace tant sur le plan calculatoire qu'en terme des structures de discours facilement interprétables. Mes recherches m'ont conduit à aborder des sujets traitant de trois axes principaux :

- la variation en langues, ou comment mettre au point des méthodes d'analyse indépendante des langues,
- la prise en compte du genre des textes, ou comment détecter et tenir compte de régularités de forme ou de style dans l'analyse,
- la gestion des ordres de grandeur, ou comment analyser de manière raisonnée des documents de taille variée.

Ces trois axes sont étudiés au travers d'applications informatiques, à savoir des logiciels de validation de concepts. Leurs objectifs sont variés mais un même mode de raisonnement les unit, dirigé par des principes d'économie de ressources et d'efficacité des calculs. Quelques exemples : l'identification de la structure des articles scientifiques et des livres, la constitution automatique de lexiques multilingues, le suivi des propos des locuteurs dans la presse internationale, l'analyse de nouvelles formes de communication comme les forums de discussion.

Les trois axes présentés ne sont jamais étudiés simultanément dans un même cadre applicatif mais sélectionnés en fonction des traits communs de la collection de documents à traiter. Ces traits communs, en terme de langue, de genre, de taille ou d'ordre de grandeur, constituent des invariants sur lesquels la méthode de résolution s'appuie. Ce sont en quelque sorte les paramètres d'un système d'équations, fixés pour les besoins de l'expérimentation.

Mes travaux se font dans le cadre de collaborations institutionnelles et industrielles, aussi bien locales, régionales, que nationales. Ces collaborations qui s'inscrivent dans la durée prennent la forme de projets interdisciplinaires – en linguistique, en sciences de l'éducation, et en traductologie – et de co-encadrements de thèses ou stages de master recherche. Elles impliquent des chercheurs au profil complémentaire, particulièrement Nadine Lucas, chargée de recherche en linguistique, et Jacques Vergne, professeur en informatique, avec lesquels je partage une même vision du traitement des langues au sein de l'équipe DLU du GREYC, Éric Bruillard, professeur d'informatique à l'ENS Cachan et à l'INRP, Christine Durieux, professeure en sciences de la traduction à l'UCBN, Hervé Déjean, chercheur de la société XRCE, Patrick Constant, directeur de la société Pertimm ou Dominique Dutoit, directeur des sociétés Memodata et Sensegates.

La visibilité donnée à mes travaux se traduit par une activité de publication internationale, principalement dans des conférences internationales avec comité de lecture et plus récemment en revues internationales avec comité de lecture. Elle se concrétise également par la conception, la réalisation et la maintenance de plateformes accessibles par Internet et permettant la présentation interactive et permanente des outils d'analyse et de visualisation que je réalise seul ou en équipe.

2.1 La variation en langues

Comment mettre au point des méthodes d'analyse indépendante des langues ? Par indépendante des langues, il faut entendre des méthodes automatiques, qui une fois implémentées, sont capables d'analyser un texte ou un document quelle que soit sa langue, sans aucun paramétrage manuel. Ce positionnement entre en résonance avec la préoccupation d'organisations internationales comme l'Union Européenne ou l'ONU qui soutiennent un traitement égalitaire de toutes les langues et préconisent la prise en compte de toutes les langues dans les traitements linguistiques, y compris celles qui ne présentent pas un intérêt économique ou stratégique immédiat.

En traitement des langues, la manière classique d'envisager le multilinguisme est de proposer un algorithme, ou un système informatique générique, paramétré par des ressources linguistiques spécifiques à chaque langue. Dans ce contexte, pour être en mesure d'analyser tel ou tel ensemble de langues, deux solutions existent : faire appel à des locuteurs de chacune des langues ayant capacité à paramétrer l'algorithme ou le système informatique, ou bien construire automatiquement les dites ressources à partir de corpus. Lorsque l'on travaille dans un contexte fortement multilingue ou lorsque l'on s'intéresse à des phénomènes ayant une couverture mondiale, l'épidémiologie par exemple, aucune de ces solutions n'est en fait viable.

Depuis la fin de ma thèse, et en rupture avec la communauté scientifique, je m'intéresse à mettre au point une troisième voie : celle des méthodes de résolution indépendantes des langues mais dépendantes des invariants de la tâche. Il s'agit de modéliser des propriétés linguistiques (Vergne, 2002, 2005 et 2009), donc de faire davantage de place aux théories favorisant la situation ou les contraintes de perception, comme le souligne Lucas (2009). Nous verrons que ce positionnement conduit indirectement à faire place à des modèles intégrant des propriétés sémiotiques, dans la lignée des travaux de Valette et Rastier (2008). Il s'agit à mon sens d'une innovation majeure en traitement des langues.

Plusieurs applications illustrent ces principes. L'identification automatique de la structure de documents est un exemple de traitement multilingue. Il permet la fabrication automatique de tables de matières de documents numérisés, quelle que soit leur langue, et sans aucun paramétrage linguistique. La segmentation thématique de forums de discussion est un autre exemple de traitement multilingue. Elle permet la production de représentations compactes de forums de discussion en toute langue. Nous l'avons notamment appliquée à des langues aussi diverses que le français, l'anglais, le grec, l'hébreu ou le vietnamien. Le système fonctionne sans retouche, sans paramétrage, pour n'importe quel forum écrit dans une autre langue.

2.2 La prise en compte du genre des textes

Comment détecter et tenir compte de régularités de forme ou de style propres aux articles de presse par exemple ? Avec l'avènement des nouvelles formes de communication médiatisée, comme les conversations en ligne ou les forums de discussion, il est aujourd'hui possible de s'exprimer beaucoup plus librement sur Internet. L'expression n'est plus purement textuelle. On a tout d'abord vu apparaître les smileys en même temps que les règles de grammaire étaient mises à mal. Les formes propres aux communautés sociales se sont développées, chacune avec leur propre code, une mise en forme matérielle beaucoup plus riche, accompagnés de smileys « nouvelle génération » puisés dans des bibliothèques, la personnalisation des images de fond, dans les réseaux sociaux.

Sans se référer à ces situations de communication, parfois jugées « extrêmes », on sait également que les performances d'un même analyseur de langue se dégradent nettement lorsque l'on change de domaine d'application. Par exemple, des systèmes d'analyse conçus pour la presse sont peu performants pour analyser des romans ou des articles scientifiques. Les travaux concernant l'analyse

du langage SMS (Fairon *et al.*, 2006) ou l'analyse d'oral retranscrit témoignent également des limites des approches actuelles.

Le traitement des langues s'est longtemps focalisé sur l'analyse d'écrits propres et lissés, débarrassés de toute scorie qui pourrait interférer avec les attendus des divers algorithmes. La stratégie dominante consiste à nettoyer le texte avant de l'analyser, à le débarrasser de sa mise en forme, à corriger ses fautes d'orthographe, à « traduire » ses sigles, ses abréviations, à le réduire à un français codifié, et plus généralement à une langue codifiée. L'attente forte sur la qualité du texte à analyser explique la dégradation des résultats lorsque l'on change de genre, et l'impossibilité de traiter les nouvelles formes de communication plus libres et non corrigibles.

Depuis la fin de ma thèse et plus particulièrement depuis mon entrée au laboratoire GREYC en 2005, je m'attache à mettre au point des solutions logicielles innovantes pour analyser les documents tels qu'ils sont, avec leur dimension textuelle, figurative, typographique et dispositionnelle. On ne peut présager des modalités utiles pour un traitement avant de connaître le contexte applicatif. Les critères qui vont marquer une structure, en fonction d'une langue ou d'un ordre de grandeur, doivent être captés par le processus de structuration automatique. Je cherche donc à mettre au point des principes méthodologiques qui permettent la prise en compte de chaque modalité lors du calcul de structure.

Plusieurs applications mettent en avant la prise en compte du genre. Dans l'analyse thématique des forums de discussions, les smileys, la ponctuation, la mise en forme entrent dans le calcul de la structure au même titre que le texte, texte qui est par ailleurs analysé « en l'état », sans correction de l'orthographe. Dans les récents développements de mes travaux sur l'alignement automatique multilingue, la typographie, les images, les sigles, le texte sont tous considérés par le processus pour augmenter la qualité et la quantité des traductions extraites automatiquement.

2.3 La gestion des ordres de grandeur de documents

Comment analyser de manière raisonnée des documents de taille aussi variée que des dépêches d'agences, des articles scientifiques ou des romans ? Par analyse raisonnée, j'entends une analyse reposant sur un modèle visant la structuration de textes ou de documents de toute taille, en un nombre contrôlé et réduit d'unités. Résoudre cette question, c'est gérer les ordres de grandeur.

En traitement des langues, le modèle propositionnel est dominant. Il est construit autour de la relation sujet-verbe. C'est un modèle simple : peu de positions, peu de fonctions. C'est ce qui fait la force de ce modèle. Son cadre d'étude classique, la phrase, est en pratique le seul étudié. Dans une approche orientée corpus, où il s'agit d'analyser des textes ou des documents, l'utilisation du modèle propositionnel dans le cadre de la phrase aboutit à la structuration d'un roman, d'une documentation technique ou d'un texte réglementaire, en une liste considérable et non structurée de propositions. Or, s'il est une représentation visée qui ne fait pas sens lorsque l'on s'intéresse à l'analyse d'un tout, qu'il s'agisse d'un article, d'un livre, ou d'une collection d'ouvrages, c'est bien une suite de quelques milliers de phrases ou de propositions analysées. Une analyse raisonnée consisterait davantage en une structuration de documents en quelques grandes parties fonctionnelles, et ce quelle que soit la taille des documents.

Apprendre à gérer les ordres de grandeur, c'est apprendre à gérer l'élasticité du discours. C'est une problématique difficile qui ne trouve pas de réponse dans l'état de l'art de l'analyse automatique du discours. Depuis la fin de ma thèse, mes recherches en matière de gestion d'échelle consistent à identifier et lever les verrous scientifiques qui entravent la confrontation automatique de modèles linguistiques à différents ordres de grandeur, et donc à des textes de toute taille. J'ai obtenu de réelles avancées sur la question en posant les fondations d'un cadre méthodologique.

Mon cadre méthodologique impose que la résolution soit *guidée par le modèle,* et non par les données, de manière à garantir la cohérence du résultat. Le modèle définit les fonctions en jeu, les positions absolues ou relatives des zones fonctionnelles, leurs rapports de forme. Ce cadre impose *l'abandon d'une unité minimale d'analyse de taille fixe*, au profit d'une unité minimale d'analyse compatible avec l'ordre de grandeur du texte, par exemple le mot pour l'analyse de phrases, la phrase pour l'analyse de dépêches d'agences, le message pour l'analyse de forums de discussion. Bien évidemment, l'on s'attend à ce que l'unité minimale de courtes dépêches soit différente de l'unité minimale de dépêches plus longues, la phrase versus le paragraphe. L'on s'attend également à ce que l'unité minimale de forums courts soit différente de l'unité minimale de forums longs, le message versus un groupe de messages formant un échange.

Dans ce cadre méthodologique, la résolution s'appuie sur une méthode différentielle, dans la lignée des travaux de Rastier (2002) et Coursil (2000) poursuivis par Beust (Roy et Beust, 2007) dans l'équipe DLU du GREYC, et qui constituent un des piliers de notre culture d'équipe. Elle est fondée sur la recherche *endogène* de contrastes formels permettant l'attribution des fonctions définies dans le modèle, sous contraintes des propriétés du modèle, les zones fonctionnelles et les frontières entre zones fonctionnelles devant être envisagées de taille variable et compatibles avec l'ordre de grandeur traité (Lucas, 2009).

Plusieurs applications ont été développées pour valider ces principes. Une application de structuration thématique a notamment été développée sur des articles journalistiques, sur des articles scientifiques, et sur des forums de discussion en toute langue. Cette application a montré sa capacité à produire des représentations textuelles compactes et toujours lisibles de ces documents, et ce quelle que soit leur taille. Une application plus récente concerne l'identification de la structure de livres entiers. Cette application consiste en la fabrication automatique de tables de matières interactives pour les documents numériques de toute langue. Elle permet au lecteur l'accès direct à tel ou tel chapitre, telle ou telle section, via une représentation compacte du document, à savoir sa table des matières.

3 Synthèse des travaux sur l'alignement automatique multilingue

L'alignement automatique est à la base des systèmes de traduction statistique (Brown *et al.*, 1988 ; Och et Ney, 2002 et 2004 ; Koehn *et al.*, 2003) et des systèmes d'aide à la traduction basés sur les mémoires de traduction (Isabelle, 1992 ; Foster *et al.*, 1997 ; Foster *et al.*, 2002). L'utilisation de telles mémoires repose sur le constat que certains types de documents, comme les documentations techniques ou les textes contractuels, comportent de nombreuses répétitions d'une version à l'autre, d'un client à l'autre ou d'un produit à l'autre. Le recours à l'historique des traductions déjà réalisées permettrait alors de réduire le coût des nouvelles traductions tout en favorisant la cohérence lexicale de l'ensemble des traductions produites.

Dans ce contexte, l'alignement consiste à repérer automatiquement les relations de traduction existant entre les versions traduites d'un document ou d'un groupe de documents, et à fabriquer des mémoires de traduction constituées de segments de textes en différentes langues. Ces relations sont à chercher au sein des traductions à différents niveaux, supra-phrastiques et sous-phrastiques (Isabelle, 1992), par exemple entre phrases (cette phrase est la traduction de celle-ci), entre paragraphes (ce paragraphe est la traduction de celui-ci), entre termes, entre mots. Les relations peuvent en outre être envisagées entre niveaux de granularité (Kraif, 2001), un mot pouvant être traduit par une proposition, par exemple.

3.1 État de l'art

Si l'alignement au sens large consiste à établir des relations de traduction entre segments de texte, à différents niveaux de granularité, force est de constater que l'alignement de phrases et l'alignement de mots ont constitué deux des grands axes du domaine.

L'alignement automatique de phrases a rapidement été considéré comme résolu. Les articles fondateurs de Gale et Church (1991 et 1993) et Brown *et al.* (1991) ont montré que des tests statistiques extrêmement simples, basés notamment sur la longueur des phrases, donnaient des résultats très acceptables lorsque les textes de départ ne présentent pas d'inversions importantes de phrases, ni d'éclatements trop importants de phrases sources en plusieurs phrases cibles, et réciproquement. C'est le cas pour des langues de même famille. On peut aussi s'appuyer sur un ancrage lexical grâce aux invariants graphiques appelés cognats (les nombres, les sigles, et autres mots similaires entre 2 langues) ou à des dictionnaires bilingues. Ce courant est illustré par les travaux de Simard (1992) et Chen (1993). Les travaux de Moore (2002) allient les deux approches précédentes.

Le succès rencontré par l'alignement de phrases repose sur une hypothèse forte : le parallélisme des traductions (Langé et Gaussier, 1995). Cette hypothèse suppose que la chronologie du texte est préservée, qu'à une unité source correspond le plus souvent une unité cible de même niveau (une phrase est traduite par une phrase), et que les proportions sont préservées (un passage court est traduit par un passage court).

L'alignement automatique de mots est généralement associé à l'outil GIZA++ (Och et Ney, 2003). Plus que le dépassement des performances de cet outil, les travaux portent aujourd'hui sur la

simplification des modèles à mettre en œuvre (Moore *et al*., 2006 ; Lardilleux et Lepage, 2008) et l'obtention de scores comparables sur des tâches ou des tests répertoriés. Si l'alignement de mots permet également la production d'alignement de suites de mots contigus, la question de la production d'alignement de suites de mots discontinus a également été abordée (Goutte *et al*., 2004). La production d'alignement de suites de mots améliore la traduction statistique, mais un des reproches adressés à cette technique tient au fait que les suites de mots alignées ne correspondent pas à une définition intuitive ou linguistique et ne font pas sens pour l'humain, surtout lorsque celui-ci est traducteur.

L'alignement sous-phrastique regroupe les travaux portant sur l'alignement d'unités intermédiaires à la phrase et au mot. L'objectif de ces travaux est de produire des alignements qui font sens, c'est-à-dire facilement interprétables. Le *chunk*, apparu dans la communauté scientifique du traitement automatique des langues à la fin des années 1980 (Church, 1988; Abney, 1991) est un des concepts utilisés à cette fin. Kupiec (1993) fait ainsi appel à deux segmenteurs en chunks, l'un du français, l'autre de l'anglais, pour réaliser ce type d'alignement, Gaussier (1998) effectue une extraction terminologique en partant d'une segmentation en chunks de l'anglais et en recherchant le meilleur alignement de mots français, sous contrainte de contiguïté, de similairité de longueur des segments, de probabilités de traduction entre mots. D'autres techniques d'alignement sous-phrastiques ont également été explorées : Nakamura-Delloye (2007) s'est par exemple intéressée à l'alignement de propositions, Wu (1997) à l'alignement de structures de phrases.

La nature des unités alignées, c'est-à-dire le résultat de la procédure d'alignement automatique, est l'angle que j'ai choisi pour structurer cet état de l'art. Un regard sur l'entrée de la procédure permet cependant de relever quelques invariants méthodologiques. L'alignement automatique porte classiquement sur une paire de langues, impliquant généralement l'anglais. Dès 1999, une expérience de Simard portant sur l'alignement phrastique (1999) laissait cependant envisager de meilleures performances en alignant trois langues simultanément. L'alignement automatique sous-phrastique suppose classiquement une segmentation en mots. En 2006, Crosmières (2006 ; 2010) présente un alignement au grain caractère. Enfin, si l'entrée attendue d'une procédure d'alignement correspond soit à un ensemble de phrases comme le BTEC (Takezawa et al., 2002), soit à une collection de textes comme l'Acquis Communautaire (Erjavec, 2005), on comprend que l'alignement est davantage envisagé comme un traitement *littéral* que comme un traitement de document ayant une structure textuelle et une mise en forme matérielle.

3.2 Émergence de la thématique, encadrements et collaborations

C'est au travers de l'alignement sous-phrastique grec-anglais que j'aborde l'alignement automatique et c'est par ces travaux que la thématique de recherche prend naissance au GREYC en 2005 (Giguet et Apidianaki, 2005). L'étude porte alors sur un corpus de résumés d'articles scientifiques médicaux en anglais et en grec, proposé par Mariana Apidianaki du laboratoire Lattice (Apidianaki, 2008). Mon objectif est de montrer qu'il est possible de mettre en évidence des équivalences sous-phrastiques entre des textes rédigés dans des langues ayant des morphologies très différentes, en n'utilisant aucune ressource linguistique, mais des similarités de distribution.

Le choix de réaliser cette étude n'a rien de fortuit. Il s'agit au départ d'une application des traitements indépendants des langues auxquels je m'intéresse. Pouvant être appliqués à n'importe quelle langue, les prototypes que j'avais conçus au Lattice dans la lignée des travaux de Hervé Déjean (1998) étaient à même de croiser ceux de la traduction automatique. C'est par l'alignement automatique que l'occasion s'est présentée. Au laboratoire GREYC, une collaboration avec Pierre-Sylvain Luquet permet d'appliquer la méthode d'alignement sous-phrastique que j'ai mise au point, sur n'importe quelle couple de langues de l'Acquis Communautaire, un corpus de documents

multilingues en 20 langues (Giguet, 2005a, 2005b ; Giguet et Luquet 2005).

Mes recherches sur l'alignement prennent un véritable essor au travers des actions que j'ai entreprises pour initier une thématique de recherche dans le laboratoire GREYC, créer des collaborations internes à l'Université avec les Sciences de la traduction, et initier des collaborations extérieures avec notamment les sociétés Pertimm, concepteur de moteurs de recherche, et Lingua et Machina, éditeur de l'environnement logiciel d'aide à la traduction Similis.

Avec Jacques Vergne, nous établissons une collaboration pluri-disciplinaire avec les Sciences de la traduction au sein de l'Université de Caen Basse-Normandie. Par l'intermédiaire de Christine Durieux, professeur en Sciences de la traduction, nous proposons des sujets de stage de Master recherche pour le Master conjoint franco-hellénique mention Sciences du langage, spécialité Sciences de la Traduction : Traductologie et Sciences cognitives. Nous encadrons ainsi avec Nadine Lucas les étudiantes Anne Lemoine et Calliopi Sachtouri en 2006 (Lemoine, 2006 ; Sachtouri, 2006) puis Charlotte Lecluze et Marina Trichaki en 2007 (Lecluze, 2007 ; Trichaki, 2007). Les sujets portent sur l'alignement et la recherche d'invariants dans une collection de traductions massivement multilingues. Parallèlement, nous encadrons le stage de Master recherche en informatique de Romain Brixtel portant sur l'alignement endogène de structures de documents, contraint par la mise en forme (Brixtel, 2007).

Notre groupe de travail s'inscrit dans l'organigramme du laboratoire avec la création d'un axe thématique « Alignement et traduction » au sein de l'équipe ISLanD. Il rassemble alors quatre permanents : notre groupe composé à l'origine de trois permanents se trouve renforcé par l'arrivée d'Yves Lepage. Il apporte son expérience internationale en traitements multilingues et en traduction automatique, notamment pour les langues asiatiques et slaves, et pour la participation à des campagnes d'évaluation internationale (Lepage *et al.*, 2009 ; Gosme *et al.*, 2010).

De 2007 à 2011, j'ai co-encadré la thèse de Romain Brixtel, intitulée « Alignement endogène de documents, une approche multilingue et multi-échelle » a été soutenue en janvier 2011 (Brixtel, 2011). Le sujet de thèse a été classé prioritaire par le laboratoire et nous avons établi une co-direction pluridisciplinaire composée de Jacques Vergne, Christine Durieux et moi-même.

Depuis 2008, je co-encadre la thèse CIFRE de Charlotte Lecluze, avec Jacques Vergne, et, côté entreprise, Patrick Constant et Loïs Rigouste. J'ai effectué la négociation du contrat d'encadrement de Charlotte auprès de Pertimm que nous connaissions de longue date, participé activement au montage du dossier auprès de l'ANRT et j'en suis responsable scientifique officiel.

3.3 Principales contributions

Ma contribution consiste en la réalisation d'une modélisation permettant de se dégager de la primauté des formes dans le traitement informatique. J'étudie la répartition de formes dans un document, à l'aide d'une fenêtre d'observation, selon la méthode distributionnelle. En cela, mes préoccupations rejoignent celles de Habert et Zweigenbaum (2003). Cependant, la position prend le pas sur la forme, suivant les travaux de Nadine Lucas (2009), un positionnement que l'alignement sous-phrastique me permet de promouvoir.

3.3.1 Une approche distributionnelle de l'alignement

La méthode d'alignement sous-phrastique que j'ai proposée repose sur deux hypothèses principales :

- la première suppose que des correspondances sémantiques peuvent être établies entre des segments sous-phrastiques ayant des répartitions similaires dans des versions traduites d'un

même document. Il s'agit ici de capter la composante récurrente de la traduction : les segments qui sont traduits le plus souvent de la même manière.

- la seconde hypothèse suppose qu'un segment sous-phrastique apparaissant dans plusieurs documents de même langue n'a pas nécessairement un unique correspondant sémantique dans la collection. Il s'agit d'accroître la confiance en un alignement lorsqu'il résulte de l'analyse de plusieurs documents. Il s'agit également de laisser place à la découverte de correspondants multiples au sein de la collection, pour une séquence sous-phrastique donnée.

La **similarité des distributions** est à la base de ma méthode d'alignement sous-phrastique. Au travers des deux hypothèses précédentes, l'idée est en effet de privilégier des contraintes relationnelles fortes, basées sur le critère de position des segments à relier, pour relâcher les contraintes de forme portant classiquement sur les segments reliés.

L'approche distributionnelle permet de générer des correspondances très fiables, à différents niveaux de granularité sous-phrastique : alignements de mots, de termes, d'expressions. L'exécution du programme sur des volumes de documents importants met en évidence que la méthode ne génère pas de résultat lorsque les textes ne sont pas traductions l'un de l'autre. La qualité des résultats ne se dégrade pas lorsqu'une version n'est qu'une traduction partielle de l'autre, ou lorsque l'alignement phrastique préalable produit un résultat erroné, par exemple lorsque les textes ne sont pas parallèles. La méthode compare cependant les distributions de segments répétés et ne permet donc pas de mettre en relation des segments « hapax ». Cette question a été abordée plus tard dans l'équipe par Lardilleux (2010).

Les correspondances établies se distinguent de l'état de l'art par le fait que les tailles des segments mis en relation ne sont pas contraintes par la méthode. Les alignements sous-phrastiques peuvent ainsi être produits entre segments de taille très différentes tout en restant fiables. La méthode n'impose pas de contraintes de catégorie, de forme, ou de probabilité de traductions sur les mots des segments mis en relation. Elle ne nécessite donc ni base lexicale monolingue, ni base lexicale bilingue. Cela ne gêne en rien la production de correspondances interprétables, notamment entre termes de spécialité. La méthode impose cependant une contrainte de forme sur les segments reliés, à savoir la contiguïté des mots, elle ne peut donc pas directement mettre en relation des segments discontinus.

Exemple d'alignements sous-phrastiques anglais-français produits par WimsAlign

Anglais	**Français**
and	et
council	conseil
the commission	la commission
having regard to the	vu
member states	états membres
or	ou
directive	directive
regulation	règlement
of	de
this directive	la présente directive
article	article
whereas	considérant
community	la communauté
this regulation	présent règlement

in particular	notamment
economic	économique
of the european communities	des communautés européennes
the council	conseil
no	nº
the treaty	traité
provisions	dispositions
regulation (eec) no	règlement (cee)
council directive	du conseil
having regard to the opinion of the	vu l'avis

Cet exemple présente les alignements sous-phrastiques les plus fréquents produits à partir d'un sous-corpus de 250 documents anglais-français de l'Acquis Communautaire. Les alignements de chiffres, de ponctuations, ou de combinaisons de chiffres et ponctuations ont été filtrés, bien que pertinents.

Exemple d'alignements sous-phrastiques anglais-néerlandais produits par WimsAlign

Anglais	**Néerlandais**
article	artikel
the	de
and	en
of	van
the commission	de commissie
member states	lid-staten
the european	europese
directive	richtlijn
whereas	overwegende dat
regulation	verordening
community	gemeenschap
having regard to the opinion of the	gezien het advies van het
council	de raad
article 1	artikel 1
annex	bijlage
; whereas	; dat
having regard to the	gezien het
the council	de raad
article 2	artikel 2
member	lid-staten
member state	lid-staat
the treaty	het verdrag

Cet exemple présente les alignements sous-phrastiques les plus fréquents produits à partir d'un sous-corpus de 250 documents anglais-néerlandais de l'Acquis Communautaire. Les alignements de chiffres, de ponctuations, ou de combinaisons de chiffres et ponctuations ont été filtrés, bien que pertinents.

3.3.2 Une approche multilingue de l'alignement

La méthode d'alignement sous-phrastique proposée au début de mes travaux a l'intérêt d'être immédiatement applicable à n'importe quel corpus de documents traduits, quel que soit le couple de langues. Sa mise en œuvre sur un corpus de textes réglementaires en 20 langues européennes –

l'acquis communautaire – montre sa propension à fournir des alignements corrects sur n'importe quel couple de langues et à moindre frais, sans requérir la disponibilité de tel ou tel lexique monolingue ou bilingue, de tel ou tel analyseur linguistique (étiqueteur, segmenteur en chunks, lemmatiseur, ...). Ce positionnement répond à la préoccupation de prendre en compte les langues peu dotées, comme le préconise l'Union Européenne.

Dans sa première phase de conception, la méthode concernait uniquement les langues où le mot est facilement accessible, c'est-à-dire délimité par des espaces ou des ponctuations. Depuis 2006, le grain d'analyse utilisé pour former les segments à aligner est paramétrable et il est donc possible de choisir une segmentation au caractère, dans la lignée de Crosmières (2006) pour traiter les langues où le mot n'est pas graphiquement marqué.

La **différence de richesse morphologique** entre des langues comme l'anglais et le grec engendre un certain silence lorsque l'on cherche à aligner des textes non lemmatisés (Déjean *et al.*, 2003). À une marque casuelle près, deux séquences de mots sont graphiquement différentes et considérées comme telles. Elles ne peuvent par conséquent pas être alignées avec une séquence invariante d'une autre langue. Cela viole la condition de similarité de distribution de séquences sous-phrastiques *identiques*. Cette contrainte n'est pas satisfaisante car trop stricte pour l'objectif visé qui est de produire un maximum d'équivalences. Nous avons donc choisi de ne pas considérer significatives les altérations flexionnelles en procédant à une lemmatisation.

Pour normaliser les séquences à apparier, j'utilise mon algorithme de lemmatisation endogène WimsMorph basé sur la recherche d'affixes. Les affixes sont calculés à partir des vocables du corpus par un algorithme de recherche de frontières entre noyaux et affixes ou entre affixes, selon la méthode de Déjean (1998). Dans le cas de l'alignement entre le grec et l'anglais, deux langues qui n'ont pas la même richesse morphologique, ce processus de normalisation permet d'augmenter les mises en correspondance sans dégrader la fiabilité des résultats.

Exemple de suffixes produits par analyse morphologique endogène avec WimsMorph

français : ~t ~ait ~ant ~ent ~aient ~e ~elle ~s ~ais ~es ~ons ~ion ~on ~ux ~eux ~u ~er ~our ~ir ~é

grec : ~α ~ου ~ς ~ικές ~ίας ~ικής ~ους ~ούς ~ο ~νται ~εται ~ει ~ων ~ών ~ουν ~η ~κού ~ική ~εί

anglais : ~cial ~al ~sion ~ation ~reas ~eas ~ments ~ions ~ing ~nder ~ment ~ent

Dans cet exemple, l'analyse morphologique endogène est calculée sur la liste des vocables du corpus à analyser.

Exemple d'alignement sous-phrastique de séquences de mots après lemmatisation endogène par WimsAlign

Anglais	**Grec**
and	και
patients	ασθενείς
insulin	ινσουλίνη~
between	μεταξύ
with	με
diabetes	διαβήτη
glucose	γλυκόζη~
levels	επίπεδα
women	γυναικες
but	αλλά

or	ή
studies	μελέτες
during	κατά τη διάρκεια
this	αυτή~
vs	έναντι
significantly	σημαντικά
for	για
respectivly	αντίστοιχα
REVIEW	ΑΝΑΣΚΟΠΗΣΗ
results	Αποτελέσματα
Key words	Λέξεις ευρετηρίου

Cet exemple présente quelques alignements sous-phrastiques produits après lemmatisation endogène – marquée par le ~ – sur un corpus de résumés bilingues anglais-grec, alignés au grain phrase, d'articles scientifiques. La lemmatisation endogène permet de compenser les différences de richesses morphologiques qui aurait empêché l'alignement de séquences comme insulin *et* glucose.

3.3.3 Une approche de l'alignement à grain d'analyse paramétrable

Lorsque l'alignement est mené sur une langue agglutinante, comme l'estonien ou le finnois, la quantité d'alignements générée par la méthode diminue sensiblement. On pourrait imputer ce phénomène à l'analyseur morphologique endogène WimsMorph qui n'est pas adapté aux traitements des langues agglutinantes. Il est vrai qu'une segmentation endogène en morphèmes, à la manière de (Déjean 1998), serait plus appropriée et préserverait les atouts de l'approche endogène en terme de généricité. J'ai cependant préféré envisager ce problème sous l'angle de la **compatibilité des grains d'analyse**.

Une des hypothèses implicites du travail précédent réside dans le fait que l'alignement peut être réalisé par la mise en correspondance de suites de mots, des mots « graphiques », délimités par des espaces ou des ponctuations, repérables sans lexique par l'ordinateur. Or, le mot graphique correspond à des réalités très diverses selon les langues (Lecluze, 2011). Chercher des alignements en observant des similarités de distributions de séquences de mots graphiques n'est donc pas forcément productif. C'est pour cela que la lemmatisation endogène a été introduite : pour considérer identiques des mots qui graphiquement ne le sont pas. La lemmatisation permet de porter un autre regard sur la graphie. Dans notre version endogène, la lemmatisation permet de *masquer* les affixes.

Pour calculer les alignements en présence de langues agglutinantes, j'ai choisi de segmenter automatiquement ces langues non plus en mots graphiques mais en caractères, dans la lignée des travaux de Crosmières (2006, 2010). La similarité de distribution n'est alors plus évaluée entre séquences d'unités de même nature, à savoir des mots graphiques, mais entre séquences d'unités pouvant être de nature différente, mots graphiques, mots lemmatisés, caractères, selon la ou les fonction(s) de segmentation choisies (une par langue). Le processus d'appariement des séquences reste lui inchangé, à savoir la similarité de distribution. Par rapport à l'état de l'art qui envisage l'alignement comme prenant en entrée le résultat d'une unique fonction de segmentation pour les deux langues, nous travaillons désormais avec deux fonctions de segmentation, une par langue, les deux fonctions pouvant bien entendu être identiques. C'est là une contribution originale au domaine.

Notons qu'en autorisant l'alignement de séquences de caractères, les langues comme le chinois, où le mot graphique n'est pas délimité par l'espace, sont désormais prises en compte, ce qui accroit sensiblement le champ d'application de notre méthode d'alignement.

Les premières expérimentations que j'ai conduites tendent à montrer qu'en segmentant en caractères les langues agglutinantes et en segmentant en mots les langues non agglutinantes, on augmente la quantité d'alignements produits sans perdre en qualité. Ces mêmes expérimentations tendent à montrer qu'en sélectionnant une segmentation en caractères pour les deux langues, alors les résultats s'en trouvent globalement dégradés. C'est le même constat qu'effectue Brixtel (2011) en segmentant systématiquement ces deux langues en chunks. Sous réserve d'expérimentations plus abouties, il est tentant de faire un parallèle entre la dégradation des résultats lorsque le grain est systématiquement le caractère, lorsque le grain est systématiquement le mot, ou lorsque le grain est systématiquement le chunk, c'est-à-dire lorsqu'une unique fonction de segmentation est utilisée pour les deux membres du couple, indépendamment des familles auxquelles elles appartiennent.

Exemple d'alignements sous-phrastiques anglais-estonien à grains différenciés produits par WimsAlign, ici mot pour l'anglais et caractère pour l'estonien.

Anglais	**Estonien**
having regard to the	[võttes arvesse]
article	[artik]
the european	[euroopa]
economic	[majandus]
or	[või]
and	[ja]
whereas	[;]
member	[liikmesrii]
having regard to the opinion of the	[arvamust]
community	[ühenduse]
council	[nõukogu]
the commission	[komisjon]
having regard to	[võttes arvesse]
having	[võttes arvesse]
the council	[nõukogu]
this directive	[käesolev]
this	[käesolev]
this regulation	[käesolev]
the treaty	[asutamislepingu]
directive	[direktiiv]
committee	[komitee]
regulation	[määrus]
of the european communities	[euroopa ühenduste]
. member states shall	[. liikmesriigid]
the	[e]
member states	[liikmesrii]
not	[ei]
proposal from the commission	komisjoni ettepaneku]

Dans cet exemple, les alignements ont été produits à grains différenciés : segmentation en mots pour l'anglais, en caractères pour l'estonien. Les séquences de caractères ont été entourées de crochets pour rendre compte des espaces débutant ou clôturant les séquences.

Au grain caractère, la qualité des séquences est souvent discutée car les segments de caractères ne correspondent pas à des segments directement interprétables. Il y a régulièrement des caractères manquants ou supplémentaires aux frontières, par rapport à un attendu qui fait sens. Ce biais d'interprétation est traité spécifiquement dans le chapitre *Regard épistémologique*.

3.3.4 Une approche structurelle et typo-dispositionnelle de l'alignement

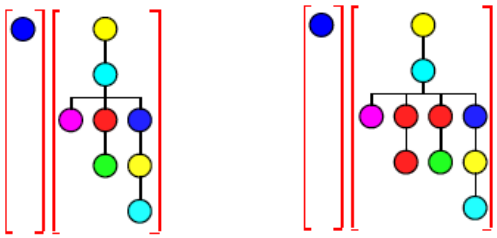
Le succès rencontré par l'alignement de phrases repose sur une hypothèse forte : le parallélisme des traductions. Cette hypothèse suppose que la chronologie du texte est préservée, qu'à une unité source correspond le plus souvent une unité cible de même niveau (une phrase est traduite par une phrase), et que les proportions sont préservées (un passage court est traduit par un passage court).

Pourtant, dans la pratique, l'hypothèse de parallélisme est régulièrement mise à mal. La qualité de l'alignement produit par les méthodes d'alignement phrastique peut se dégrader au fil du texte, et ce, malgré les techniques de recalage basées sur les « ancres fortes » (e.g., balises de titre ou de paragraphe). Ces méthodes d'alignement sont par ailleurs totalement inefficaces en cas de non parallélisme des structures. Véronis (2000) rappelle à ce titre que l'intérêt porté à l'alignement de structures de documents non parallèles est fondamental. Quoique l'alignement phrastique et supra-phrastique n'ait plus fait vraiment l'objet de travaux, c'est sous cet angle que le sujet a été reposé au sein de notre équipe et étudié au travers de deux thèses.

La thèse de Romain Brixtel (2011) que j'ai co-dirigée avec Jacques Vergne et Christine Durieux a renouvelé la problématique de l'alignement supra-phrastique. Elle porte sur l'alignement de structures de documents en utilisant la mise en forme matérielle. L'alignement de structures contraint par la mise en forme exploite la notion de contraste et généralise l'alignement phrastique à d'autres unités graphiquement identifiables comme les paragraphes, les listes, ou les sections.

La notion de contraste de mise en forme repose sur le constat que la mise en valeur du texte est préservée lors de la traduction, mais que l'on ne peut pas présager de la forme particulière que revêt cette mise en valeur. Ainsi un texte en gras dans une version pourra correspondre à un texte en italique ou en couleur dans une autre version. Romain Brixtel calcule les structures de document à partir de la mise en forme matérielle et réalise l'alignement des structures, mettant ainsi en évidence la suppression d'annexes, ou l'ajout d'un paragraphe dans telle ou telle section.

Exemple de représentation schématique de la structure de deux documents à aligner, par Romain Brixtel (2011)



Dans cet exemple, le calcul automatique des structures des deux documents à aligner, à partir de leur mise en forme relativisée, permet d'observer une différence de structuration interne dont la procédure d'alignement va rendre compte.

3.4 Méthode d'alignement multilingue proposée

3.4.1 Spécificités

La méthode d'alignement sous-phrastique que j'ai mise au point se distingue de l'état de l'art par les faits suivants :

- elle repose sur une approche distributionnelle, où la similarité des distributions des segments reliés prime sur la forme des segments reliés. La similarité des distributions est évaluée à l'intérieur du document et consolidée à l'intérieur de la collection, permettant de générer des correspondances très fiables, sans contrainte forte sur les éléments reliés,
- elle est multilingue et endogène : elle est applicable d'emblée à n'importe quel couple de langue et ne fait appel à aucune ressource externe, que ce soit en terme de lexiques bilingues ou d'analyseurs linguistiques propres à une langue (étiqueteur morpho-syntaxique, lemmatiseur, segmenteur en chunks,...),

Dans l'état de l'art, la dimension textuelle du document est souvent écartée ou ignorée et les alignements sous-phrastiques sont calculés à partir de corpus de phrases isolées, décontextualisées. L'alignement sous-phrastique fait traditionnellement appel à des lexiques bilingues et à des analyseurs linguistiques propres à chaque langue. La taille des alignements produits est contrôlée.

La méthode partage avec l'état de l'art les caractéristiques suivantes :

- elle extrait des alignements par couple de langue, et s'avère par conséquent combinatoire pour établir des alignements entre toutes les paires de langue d'un corpus massivement multilingue,
- elle suppose un alignement préalable au grain phrase, et dépend par conséquent des avancées dans ce domaine.

La méthode a séduit une partie de la communauté internationale, que ce soit par sa légèreté et sa simplicité de mise en œuvre sur des langues variées, que par la qualité des alignements produits. Les sources d'amélioration de la méthode restent cependant nombreuses. Cela tient principalement aux hypothèses réductrices qui ont été formulées. C'est au travers du co-encadrement de stages de master recherche, de la co-direction de deux thèses, dont l'une en entreprise, et de la constitution d'un groupe de travail interne au laboratoire que mes travaux sur l'alignement automatique se sont poursuivis.

3.4.2 Éléments d'implémentation

Le prototype que j'ai réalisé a été implémenté en PHP et intégré à ma plateforme d'expérimentations Wims. La récurrence des séquences sous-phrastiques à l'intérieur d'une version traduite est calculée à partir du tableau de suffixes associé (Crochemore *et al.*, 2001). La récurrence des séquences est calculée par un algorithme de recherche de chaînes répétées de taille maximale. La similarité de distribution des séquences récurrentes entre deux versions traduites est estimée à l'aide de la mesure du cosinus. L'espace orthonormé permettant le calcul du cosinus est construit par l'alignement « phrastique » des versions traduites, un alignement correspondant à une dimension de l'espace. La qualité d'un alignement est estimée par le cosinus des séquences, dans le document, et par sa fréquence, dans la collection. Dans nos expérimentations, l'alignement « phrastique » a été délégué au logiciel HunAlign (Varga *et al.,* 2005) pour l'analyse de documents « à la volée », ou bien téléchargé sur le site du laboratoire JRC (Joint Research Center) pour l'Acquis Communautaire (Steinberger *et al.*, 2006).

3.5 Prospective

3.5.1 Vers l'alignement de structures de documents non parallèles

La thèse de Charlotte Lecluze (2011), en cours, porte sur l'alignement de structures de documents, sans présupposition de l'hypothèse de parallélisme à quelque niveau que ce soit. En cela, la thèse de Charlotte Lecluze vise à opérer une synthèse : il s'agit de généraliser la méthode d'alignement endogène multilingue en remettant en cause l'hypothèse de parallélisme des méthodes d'alignement phrastiques ou supra-phrastiques. L'hypothèse de parallélisme présuppose la conservation de l'ordre des constituants phrastiques ou supra-phrastiques lors de la traduction. Ce n'est qu'au niveau de l'alignement sous-phrastique que cette hypothèse est levée, et que l'ordre des mots ou des propositions est considéré plus libre. Dans ce travail de thèse, il s'agit de rassembler ces regards contradictoires au sein d'une méthode unique et cohérente. On constate en effet que des versions d'un même document peuvent avoir des macro-structures véritablement différentes et qu'à l'inverse des versions d'une phrase peuvent avoir un ordre des mots ou des propositions très stables. La thèse prend en charge ce constat et propose une méthode d'alignement où l'ordre et le désordre ne sont plus présupposés à tel ou tel niveau, mais constatés et gérés comme tels. Cette problématique rejoint celle de Bourdaillet et Ganascia (2007).

3.5.2 Vers la sélection automatique du grain d'analyse et de la fenêtre d'observation

Les méthodes d'alignement de l'état de l'art utilisent sur une fonction de segmentation en mots pour produire des alignements de séquences de mots. Dans la lignée de Crosmières (2006), nous avons montré qu'il était pertinent de réaliser l'alignement en manipulant, non plus une unique fonction de segmentation, mais des fonctions de segmentation différenciées, tenant compte des caractéristiques des familles de langues en présence.

Il conviendrait aujourd'hui d'approfondir ces travaux et d'étudier la question de la sélection automatique de la fonction de segmentation, c'est-à-dire la sélection automatique du grain d'analyse, en s'appuyant sur le corpus à analyser. Les fonctions de segmentation de chacune des langues ne seraient alors plus choisies par l'utilisateur en fonctions des langues du corpus à analyser, mais seraient choisies automatiquement par la méthode d'alignement.

En utilisant les propriétés de notre approche distributionnelle qui ne produit pas ou peu de résultat en cas d'incompatibilité des grains d'analyse, une première approche de la sélection automatique du grain d'analyse consisterait à tester différentes combinaisons de fonctions de segmentation sur un échantillon du corpus, et à paramétrer automatiquement l'alignement par les fonctions ayant eu le meilleur rendement en terme de nombre d'alignements sur l'échantillon.

Dans cette même perspective, il conviendrait également de remettre en cause la fenêtre d'observation qui est toujours fixée, généralement la phrase, parfois le paragraphe, parfois l'alinéa. Cette fenêtre d'observation qui est utilisée pour calculer la distribution des séquences a un impact majeur sur la possibilité d'observer tel ou tel phénomène. Étudier des distributions d'objets avec une fenêtre d'observation calée par exemple sur le paragraphe ne permet pas de faire les mêmes observations qu'avec une fenêtre d'observation calée par exemple sur la phrase.

En utilisant les alignements de structures de document produits par Romain Brixtel, une première approche du paramétrage de la fenêtre d'observation consisterait à considérer chacun des niveaux de la hiérarchie d'alignements supra-phrastiques comme fenêtre d'observation.

La perspective de la sélection automatique du grain d'analyse et de la fenêtre d'observation correspond en quelque sorte à la « mise au point automatique », ou à l'*auto-focus,* de l'outil

d'observation, l'aligneur, dans une perspective d'expérimentations tenant compte des ordres de grandeur.

3.5.3 Vers un alignement massivement multilingue généralisé

C'est au travers des études d'Anne Lemoine (2006) et de Charlotte Lecluze (2007) que j'ai co-encadrées avec Jacques Vergne, qu'ont été montrés la faisabilité et l'intérêt d'un alignement massivement multilingue. Auparavant, les algorithmes étaient majoritairement bilingues. et quelques travaux isolés comme ceux de Simard (1999) évoquaient la pertinence d'aligner en considérant non plus deux mais trois versions traduites. Dans ces travaux, il s'agit d'effectuer l'alignement non plus par couple ou par triplet, mais en utilisant simultanément toutes les versions traduites disponibles. Lorsque l'alignement est réalisé entre paire de textes, cela aboutit à une forte combinatoire lorsque l'on travaille sur des corpus comme ceux de l'Union Européenne. Contrairement au sens commun, l'alignement sur un grand nombre de langues en même temps est plus facile que par paires. La démarche manuelle des stagiaires, sur 20 langues à la fois, est très laborieuse, mais la faisabilité est montrée. En revanche, la tâche convient à l'ordinateur.

Dans notre équipe, une première expérience d'alignement automatique massivement multilingue a été réalisée par Adrien Lardilleux, sous la direction d'Yves Lepage (Lardilleux et Lepage, 2008 et 2009 ; Lardilleux, 2010). Cette expérience portait sur des corpus de phrases. Il conviendrait aujourd'hui de généraliser ces travaux en tenant compte de la structure textuelle.

3.5.4 Vers un alignement automatique des chaînes de coréférence

Parmi les perspectives, l'alignement automatique des chaînes de coréférence est pour moi une des plus enthousiasmantes. J'ai encadré à ce titre une étude sur la préservation des chaînes anaphoriques dans l'activité de traduction en 2006. L'étude de Calliopi Sachtouri (2006) menée simultanément sur 20 langues confirme que la structure des chaînes anaphoriques varie légèrement d'une langue à l'autre. Elle montre cependant qu'au-delà des différences locales dans la structure de telle ou telle chaîne, l'organisation globale du discours est préservée et se manifeste par la stabilité du maillage textuel que définissent ensemble les chaînes anaphoriques globales et locales.

Il s'agit là d'une piste très prometteuse en terme d'avancée de l'alignement automatique sous-phrastique. Pour atteindre cet objectif, il conviendrait de développer la modélisation informatique de l'approche distributionnelle que nous promouvons dans ces travaux. Il s'agirait d'affiner le modèle relationnel, basé sur le critère de position des segments à relier, en utilisant la structure textuelle, pour relâcher davantage les contraintes de formes portant sur les segments reliés, en l'occurrence celles portant sur l'identité des segments à comparer.

> Exemple d'alignement manuel multilingue des chaines anaphoriques par Calliopi Sachtouri (2006) sur le communiqué IP/05/1451 (europa.eu)
>
> **Version française :**
>
> *Le multilinguisme dans l'Union européenne : la Commission européenne appelle à œuvrer pour la promotion des langues et lance un nouveau portail web*
>
> Sous le mot d'ordre « Plus tu connais de langues, plus tu es humain », la Commission européenne réaffirme son propre engagement en faveur du multilinguisme en adoptant aujourd'hui la première communication de son histoire sur ce sujet. Le document explore les diverses facettes des politiques de la Commission en la matière et

présente une nouvelle stratégie-cadre pour le multilinguisme, assortie de propositions *d'actions spécifiques*. *Celles-ci* portent sur trois domaines distincts dans lesquels les langues occupent une place importante dans la vie quotidienne des Européens : la société, l'économie et les relations de la Commission elle-même avec les citoyens de l'Union. La Commission incite les États membres à jouer leur rôle dans *la promotion de l'enseignement, de l'apprentissage et de l'utilisation des langues*. Pour marquer l'occasion, un nouveau portail web consacré aux langues et consultable dans les 20 langues officielles a été lancé sur EUROPA, le site web de l'ensemble des institutions de l'Union.

Version suédoise :

Flerspråkighet i EU: Europeiska kommissionen uppmanar till en insats för språk och lanserar en ny webbportal

Under mottot ”Varje gång du lär dig ett nytt språk blir du en ny människa” bekräftar Europeiska kommissionen åter igen sin vilja att främja flerspråkighet och antar i dag sitt första meddelande någonsin om ämnet. I meddelandet undersöks de olika aspekterna av kommissionens politik på området, och det formuleras en ny ramstrategi för flerspråkighet med förslag *till särskilda insatser*. *De* omfattar tre olika områden där språk har betydelse för EU-medborgarnas vardag: samhället, näringslivet och kommissionens egen kontakt med EU-medborgarna. Kommissionen uppmanar medlemsstaterna att medverka *till att främja språkundervisning, språkinlärning och språkanvändning*. Med anledning av meddelandet lanserar man på EU:s interinstitutionella webbplats Europa en ny webbportal för språk, som är tillgänglig på alla de 20 officiella språken.

Version finnoise :

Monikielisyys EU:ssa – Euroopan komissio kehottaa toimiin kielten hyväksi ja käynnistää uuden verkkoportaa

Mottonaan ”Jokaisen oppimasi kielen myötä kasvat ihmisenä” Euroopan komissio on tänään jälleen osoittanut olevansa sitoutunut monikielisyyteen, kun se hyväksyi lajissaan ensimmäisen monikielisyyttä käsittelevän tiedonannon. Siinä tarkastellaan eri kulmista aiheeseen liittyvää komission politiikkaa ja esitetään uusi monikielisyysstrategia *toimintaehdotuksineen*. *Toimet* kohdistuvat kolmeen eri osa-alueeseen, joilla kielillä on erityistä merkitystä EU:ssa asuvien arkielämässä: yhteiskuntaelämään, talouteen sekä komission omiin kansalaissuhteisiin. Komissio kehottaa jäsenvaltioita *edistämään omalta osaltaan kielten opetusta, opiskelua ja käyttöä*. Tapahtuman kunniaksi käynnistettiin EU:n toimielinten Europa-sivustolla uusi kieliaiheinen portaali unionin kaikilla 20:llä virallisella kielellä.

Dans cet exemple d'alignement manuel multilingue de chaînes anaphoriques, présenté ici en trois langues, on constate la stabilité du maillage textuel que définissent ensemble les chaînes anaphoriques, au-delà des petites variations concernant le nombre de maillons de chacune des chaînes ou la forme des maillons. Cette stabilité du maillage laisse envisager une détection et un alignement automatique des chaînes anaphoriques, basés sur cette contrainte relationnelle.

3.6 Conclusion

Au travers de ces recherches sur l'alignement automatique, j'ai souhaité développer et promouvoir une *approche distributionnelle* de l'alignement permettant d'obtenir des alignements sous-phrastiques fiables de nature très variée. J'ai proposé une *approche multilingue et endogène* où toutes les langues sont placées sur un plan d'égalité, alignables sans condition de ressources lexicales ou de composants linguistiques. La *prise en compte de la structure du document* a été placée au centre de la problématique de l'alignement et j'ai tenté de montrer qu'elle offrait des perspectives multiples. Une réflexion sur la prise en compte des ordres de grandeurs, sur l'influence du grain d'analyse, et sur le choix de la fenêtre d'observation, a été menée et permet de porter un autre regard sur l'état de l'art.

Enfin, j'ai proposé des nouveaux axes de recherche permettant de faire une synthèse de plusieurs techniques, certains ambitieux, mais pour lesquels j'ai souhaité montrer l'accessibilité en esquissant le chemin à suivre. Certaines études ont d'ores et déjà débuté, *l'alignement de structures de documents non parallèles,* d'autres font l'objet d'un travail de fond, *l'alignement automatique des chaînes de coréférence,* d'autres sont à mener, *la sélection automatique du grain d'analyse et de la fenêtre d'observation, l'alignement massivement multilingue généralisé*.

3.7 Publications liées

En publications liées, mes propres publications ainsi que celle des étudiants encadrés

GIGUET, Emmanuel et APIDIANAKI, Marianna. 2005. Alignement d'unités textuelles de taille variable. *Journée Internationale de la Linguistique de Corpus*. Lorient. Septembre.

GIGUET, Emmanuel. 2005a. Multi-grained alignment of parallel texts with endogenous resources. *Proceedings of the Recent Advances in Natural Language Processing (RANLP) International Workshop "New Trends in Machine Translations"*. pages 12-17. Borovets, Bulgaria. 24 septembre.

GIGUET, Emmanuel. 2005b. Linguistic-poor, multi-grained alignment of parallel text sequences. *Workshop "EU Enlargement and Integration Workshop: Exploiting parallel corpora in up to 20 languages"*. 26-27 septembre. Arona, Italy.

GIGUET, Emmanuel et LUQUET, Pierre-Sylvain. 2005. Multilingual Lexical Database Generation from parallel texts with endogenous resources. *PAPILLON-2005 Workshop on Multilingual Lexical Databases*. Chiang Rai, Thaïland. December 12-14.

GIGUET, Emmanuel et LUQUET, Pierre-Sylvain. 2006. Multilingual Lexical Database Generation from parallel texts in 20 European languages with endogenous resources. *Poster Proceedings of the ACL-COLING-2006 International Conference*. July 16-22. Sydney, Australia.

MANGEOT, Mathieu et GIGUET, Emmanuel. 2005. *Multilingual aligned corpora from movie subtitles*. Technical report, Condillac-LISTIC.

Publications des doctorants

BRIXTEL, Romain. 2007. Extraction endogène de structure pour un alignement multilingue. *TALN-RECITAL'2007* : pp. 367-376. Toulouse, France. June 2007.

BRIXTEL, Romain. 2009. Extraction d'une structure endogène de document pour l'alignement. *Congrès de l'ACFAS 2009*. Ottawa, Canada. 2009.

BRIXTEL, Romain. 2011. *Alignement endogène de documents, une approche multilingue et multi-échelle*. Thèse de doctorat. Université de Caen Basse-Normandie, Caen, France. Janvier.

LECLUZE, Charlotte. 2011. Recherche d'une granularité optimale pour l'alignement multilingue : N-grammes de caractères ou N-grammes de mots ? *Jetou*. Toulouse, France.

Publications des stagiaires de master recherche en informatique

BRIXTEL, Romain. 2007. *Alignement automatique de corpus multilingues*. Mémoire de Master Recherche. Université de Caen Basse-Normandie, Caen, France.

Publications des stagiaires de master recherche en sciences de la traduction

LECLUZE, Charlotte. 2007. *Méthode d'alignement sémantique multilingue appliquée à une collection de multidocuments : Un apport aux systèmes d'aide à la traduction*, Mémoire de Master conjoint franco-hellénique mention Sciences du langage, spécialité Sciences de la Traduction : Traductologie et Sciences cognitives, Université de Caen Basse-Normandie, Caen, France, et Université Ionienne, Corfou, Grèce.

LEMOINE, Anne. 2006. *Alignement sémantique de corpus multilingue : une perspective traductologique*, Mémoire de Master conjoint franco-hellénique mention Sciences du langage, spécialité Sciences de la Traduction : Traductologie et Sciences cognitives, Université de Caen Basse-Normandie, Caen, France, et Université Ionienne, Corfou, Grèce.

SACHTOURI, Calliopi. 2006. *Étude comparative des chaînes anaphoriques dans vingt langues européennes*, Mémoire de Master conjoint franco-hellénique mention Sciences du langage, spécialité Sciences de la Traduction : Traductologie et Sciences cognitives, Université de Caen Basse-Normandie, Caen, France, et Université Ionienne, Corfou, Grèce.

TRICHAKI, Marina. 2007. *Étude sur l'apport des cognats à l'alignement des textes*, Mémoire de Master conjoint franco-hellénique mention Sciences du langage, spécialité Sciences de la Traduction : Traductologie et Sciences cognitives, Université de Caen Basse-Normandie, Caen, France, et Université Ionienne, Corfou, Grèce.

4 Synthèse des travaux sur la structuration automatique des documents

Contrairement à notre intuition de lecteur qui perçoit naturellement la structure des documents au travers de leur mise en forme, l'ordinateur n'a généralement pas la capacité à manipuler la structure d'un document qu'il présente ou dont il extrait de l'information. Pour la machine, tout n'est qu'octets, caractères, images ou formes graphiques, disposés dans un espace particulier, avec une mise en forme particulière.

Dans les documents numérisés ou au format PDF par exemple, des concepts aussi primaires que le paragraphe, le chapitre, le titre, l'entête ou la figure, ne sont pas modélisés. Ils ne peuvent par conséquent pas être exploités directement par l'ordinateur. La structure du document, pourtant si évidente, presque tangible, est le résultat d'un processus cognitif complexe qui échappe le plus souvent aux traitements informatiques d'aujourd'hui

Dans la perspective de traitements informatiques pouvant prendre systématiquement appui sur cette structure dont comprend aisément l'utilité, la structuration automatique des documents consiste à mettre en évidence les différentes parties et sous-parties d'un document et à établir les relations qu'elles entretiennent.

4.1 Enjeux de la structuration automatique des documents

4.1.1 Enjeux applicatifs

Tout étudiant, tout chercheur, en informatique, en linguistique, en traductologie, qui s'intéresse à l'étude des langues se trouve tôt ou tard confronté à la constitution d'un corpus. Dans un projet étudiant de master recherche en informatique qui implique des fonds textuels, la constitution d'un corpus, allant du téléchargement des documents à la standardisation du seul contenu jugé pertinent, correspond à un effort pouvant atteindre la moitié de la durée totale du projet. Quand il s'agit d'étudiants, d'enseignants ou de chercheurs non informaticiens, ce temps passé peut être plus élevé en l'absence de compétences techniques pour automatiser la tâche. En thèse, la préparation d'un corpus est également critique et peut également demander un effort conséquent difficilement valorisable.

Au regard du coût que nécessite la préparation d'un corpus, certains projets de recherche privilégient la réutilisation de corpus existants plutôt que de se lancer dans la création d'un nouveau corpus, quitte à transiger sur sa représentativité. Il y a quelques années, en marge de la conférence TALN, n'ironisait-on pas sur l'opportunité de rebaptiser les recherches sur l'analyse de l'écrit « recherches sur l'analyse du journal *Le Monde* », tant les travaux sur ce corpus si facile d'accès occupaient une place prépondérante dans la communauté.

Pour les entreprises des secteurs de la recherche d'information ou de la veille stratégique sur Internet, la décision d'intégrer ou de ne pas intégrer un service de *Web scraping* pour extraire le contenu pertinent d'une page internet est stratégique. En 1999, lors du projet de transfert de technologies entre le laboratoire GREYC et la société d'intelligence économique Datops, aujourd'hui Lexis-Nexis Intelligence, dont j'assurais la responsabilité technique, une tâche et un

livrable de type W*eb wrapper* avaient dû être ajoutés en cours de contrat pour assurer ce service, tant la question tout d'abord négligée était devenue critique pour l'analyse du contenu des dépêches (Giguet *et al.*, 2000). En 2001, alors directeur de la recherche et du développement de la société d'intelligence économique Startem, je constatais qu'un de nos partenaires affectait deux ingénieurs en informatique à temps plein pour garantir le détourage de qualité des quelques milliers de sites d'information surveillés en continu.

L'enjeu de l'identification de la structure logique va bien entendu au-delà de la constitution de corpus. Cette opération permet en effet de relativiser et de hiérarchiser l'information dans les traitements informatiques. Il est admis que les informations contenues dans les différentes parties d'un document ne peuvent être considérées de manière uniforme, par une même application de traitement des langues. Ainsi, dans un système de veille épidémiologique, (Lejeune *et al.*, 2010) montrent qu'il est pertinent d'accorder plus d'importance à une information située en chapeau d'article et reprise en conclusion, qu'à une information située uniquement en corps d'article. De plus, l'importance à donner à telle ou telle partie du document dépend de l'application visée : s'il peut être nécessaire de détecter les menus de navigation d'une page internet pour les ignorer dans une perspective d'indexation automatique, il peut être nécessaire de détecter ces menus, non plus pour les ignorer, mais pour les reformuler dans un dispositif d'aide à la navigation pour déficients visuels.

4.1.2 Enjeux scientifiques

La structuration des documents a longtemps été perçue comme simplement technique, comme secondaire, face au véritable enjeu que constitue « le calcul du sens ». Pour l'étudiant en master ou en thèse, l'effort accompli, en marge de son sujet, reste rarement reconnu au regard du coût investi. Pour le chercheur, l'attrait pour le calcul du sens domine bien souvent l'attrait pour une opération de prétraitement assimilée à un « nettoyage » de corpus. La structuration des documents est à ce titre longtemps restée synonyme de suppression d'informations : suppression d'informations paratextuelles pour éliminer encarts publicitaires ou menus de navigation, suppression d'informations figuratives pour écarter illustrations ou schémas sortant du cadre de l'étude, suppression d'informations structurelles pour éliminer des attributs de mise en forme pouvant perturber une analyse automatique strictement littérale.

Tout d'abord envisagée sans réelles perspectives méthodologiques, la structuration de corpus d'étude en traitement des langues est alors une opération « assistée par ordinateur » : elle est techniquement réalisée à l'aide d'une cascade d'opérations de suppression et de réécriture propres à la source, à l'application visée et à la démarche suivie. Bien entendu, cette approche de la structuration par normalisation du contenu nuit à la confrontation avec des approches qui exploiteraient des indices effacés ou altérés (Lejeune, Giguet et al., 2011). Cela nuit également à l'évolution d'une démarche originale vers la prise en compte de ces mêmes indices. Aujourd'hui, l'importance des contenus multimédias diffusés en ligne, l'essor des nouvelles formes de communication écrite, la numérisation massive de fonds documentaires contribuent à l'émergence de ce véritable domaine de recherche. De par sa jeunesse, le domaine de la structuration des documents souffre cependant de concepts mal établis et la théorisation du domaine est balbutiante (Déjean, 2010).

Si le traitement du document ne peut se résumer à l'analyse du seul contenu textuel, c'est-à-dire à un traitement littéral, l'observation des nouvelles formes de communication écrite montre que le traitement ne peut non plus se réduire à l'étude de la modalité texte-image : il convient de pouvoir analyser un forum de discussion en tenant compte des smileys ou des avatars, de pouvoir s'appuyer sur les informations de mise en forme pour distinguer un site institutionnel d'un site activiste, de pouvoir exploiter la richesse de la mise en forme matérielle et des informations dispositionnelles

pour permettre l'analyse de revues ou de magazines (Valette et Rastier, 2008 ; Lucas, 2009). Les nouvelles formes de communication écrite témoignent de la mutation du document numérique et entraînent nécessairement le traitement du document sur le terrain de la sémiotique. Il convient désormais de s'appuyer sur l'ensemble des modalités de communication présentes, mais également sur les modalités absentes, dans une perspective d'analyse différentielle ou contrastive. La structuration automatique du document ne peut que s'inscrire dans cette dynamique, en préservant le contenu extra textuel, mais également en enrichissant ses modèles pour intégrer la dimension sémiotique des documents numériques actuels.

La structuration des documents soulève des questions complexes d'interprétation, de relativisation, et de hiérarchisation de l'information. Alors que cette opération a longtemps été perçue comme un nettoyage ou un passage obligatoire pour tenter d'accéder le plus directement au sens, alors qu'elle est aujourd'hui abordée comme un véritable enjeu technologique faisant l'objet d'une action d'évaluation internationale, il conviendrait, dans une perspective cognitive, de la considérer comme une composante à part entière de la structuration du discours.

4.2 État de l'art de la structuration automatique des documents

La structuration automatique des documents est un sujet abordé depuis longtemps dans la communauté scientifique (Wong *et al.*, 1982 ; Virbel, 1989 ; Salton *et al.* 1994 ; Mao *et al.,* 2003).

La question mobilise aussi bien des équipes industrielles telles que celle du Xerox Research Centre Europe, XRCE (Chanod *et al*., 2005 ; Chidlovskii & Fuselier, 2005 ; Déjean & Meunier, 2010) ou le Microsoft Development Center Serbia, MCDS (Dresevic *et al.*, 2009), que des équipes universitaires (Sakamoto *et al.,* 2002 ; Shafait *et al.*, 2008).

Les conférences internationales *Int'l Conference for Document Analysis and Recognition* (ICDAR), *Document Engineering,* et *Document Analysis Systems (*DAS) y sont entièrement consacrées. De nombreuses conférences généralistes de la recherche d'information, par exemple SIGIR, ou de l'analyse d'images, comme *Computer Vision Theory and Applications,* ont une section dédiée à ces questions.

Le domaine de la structuration automatique des documents se structure autour de deux grandes thématiques : la structuration physique du document et la structuration logique du document.

La structuration physique des documents concerne l'analyse de la mise en page du document. Il s'agit d'organiser le contenu des pages en régions homogènes en terme de densité d'information inscrite sur la page. Il faut par exemple regrouper les caractères en mot, les mots en ligne, les lignes en pavé de texte, les pavés de texte en pavage, et il faut également étiqueter ces régions à l'aide d'attributs comme le type de région (texte, image, tracé, graphique, tableau, formule mathématiques, séparateur), ou comme la mise en forme (motif de fond ou *background*, couleur, bordure, police de caractères...). L'interprétation des régions, en terme de paragraphe ou de chapitre par exemple, ne fait pas partie de la structuration physique mais de la structuration logique. La séparation des deux champs n'est cependant pas si claire (Antonacopoulos, 2009).

Les méthodes utilisées pour l'analyse de la structure physique des documents sont basées sur la reconnaissance de la géométrie de la page. Ces méthodes sont issues des travaux sur la reconnaissance des formes, sur l'analyse d'images et sur la reconnaissance optique des caractères. Elles ont pour trait commun d'être guidées par les données. Shafait *et al.* (2008) en font un excellent comparatif.

Les méthodes d'intersection « XY-cut » et « whitespace analysis » cherchent à obtenir une segmentation de la page par la détection de blancs, XY-cut cherchant de manière récursive des axes

horizontaux et verticaux vides, whitespace analysis cherchant des rectangles blancs de taille maximale dans la page. Il s'agit de méthodes dites descendantes ou « top-down ».

La méthode « constrained text-line finding » raffine « whitespace analysis » en sélectionnant les lignes de texte dans les blocs situés entre les rectangles blancs.Les méthodes « smearing » et « diagramme de Voronoi » sont des méthodes génériques pour la segmentation d'images, que ces images contiennent du texte ou non. La méthode « Docstrum » traite d'image de pages avec un modèle de structuration typographique sous-jacent. Ces méthodes sont dites ascendantes ou « bottom-up ».

La structuration physique du document fait l'objet d'une compétition internationale *ICDAR Page Segmentation Competition* (Antonacopoulos, 2009).

La structuration logique des documents désigne la tâche de de structuration des documents numériques ou numérisés, au format HTML ou PDF par exemple, vers un format dit « structuré », convenable pour les manipulations informatiques, ordinairement XML pour sa capacité à représenter des structures arborescentes. Classiquement, la structuration logique organise le document en une hiérarchie de constituants interprétables : parties, chapitres, sections, paragraphes, figures ou encadrés. Elle repose sur le résultat de la structuration physique.

Mao *et al.* (2003) témoignent des nombreux travaux de l'état de l'art. Ces travaux consistent à construire l'arbre des constituants logiques à partir de la structure physique, à base de règles de réécriture ou de grammaires formelles.

À l'écart de ces approches, la méthode d'analyse de (Déjean et Meunier, 2007) s'intéresse à la structuration globale du document. Il s'agit d'une méthode descendante. Le processus de structuration est conçu de manière modulaire : détection de la table des matières, détection de l'entête et du pied de page, détection de la numérotation, détection des titres et sous-titres. Cette approche modulaire se retrouve également dans les travaux de (Besagni et Bélaïd, 2004) pour l'analyse des citations dans les articles scientifiques et (Belaïd et Toussaint, 2000) pour la reconnaissance de tables des matières.

La structuration logique du document fait l'objet d'une compétition internationale annuelle INEX devenue *ICDAR Structure Extraction Competition* à laquelle nous participons depuis 3 ans. Cette compétition porte sur l'évaluation et la comparaison des techniques automatiques d'extraction de la structure des livres et des ouvrages numérisés (Doucet *et al.*, 2009, 2011).

4.3 Émergence et développement de la thématique

Mes recherches sur l'identification de la structure logique des articles scientifiques et journalistiques se sont véritablement structurées en 2007, dans le cadre d'un projet régional que j'ai porté et pour lequel j'ai obtenu un financement d'un an sur fonds FNADT, avec pour partenaire industriel la société Memodata. J'en ai assuré la responsabilité scientifique. Ces travaux ont porté sur l'identification automatique de la structure logique d'articles scientifiques bio-médicaux en anglais, et d'articles journalistiques en diverses langues téléchargés en ligne sur des sites de presse.

Ces travaux ont trouvé un prolongement naturel dans le cadre des campagnes d'évaluation internationales *ICDAR Structure Extraction competition* qui portent sur l'extraction de la structure de livres numérisés et ocrisés. L'évaluation porte sur la capacité à produire automatiquement au format XML les tables des matières d'un corpus de livres ocrisés au format PDF. J'ai coordonné la participation du laboratoire à trois de ces campagnes d'évaluation internationale, en 2009, 2010 et 2011. La participation à ces campagnes d'évaluation internationale a apporté de la visibilité au laboratoire, en le présentant comme un acteur du domaine. Cela a surtout permis d'améliorer les

algorithmes et les modèles d'analyse sous-jacents en terme de gestion des « écarts au modèle », ce qui est crucial pour le traitement de documents ocrisés.

Depuis 2006, dans le cadre de mon activité sur l'analyse des forums de discussions au travers de l'Erté Calico, et dans le cadre de mon activité d'expertise judiciaire portant régulièrement sur l'analyse de conversations instantanées sur internet, je me confronte régulièrement à l'analyse des nouvelles formes de communication écrite. Leur étude illustre l'inadéquation des attentes normatives d'un traitement de l'écrit figé, pas assez adaptable à des marqueurs de structures aussi inhabituels qu'imprévisibles, propres à chaque communauté d'utilisateurs ou à chaque utilisateur en quête de reconnaissance. Ces études, débutées dans une perspective plutôt textuelle, m'ont amené à remettre en cause profondément ma vision de la structuration automatique, et à aller vers une perspective résolument sémiotique, avec la prise en compte d'informations davantage stylistiques.

4.4 Positionnement

Mes recherches sur la structuration des documents ont l'ambition de contribuer à la modélisation de la construction du sens par le lecteur, tout en offrant aux utilisateurs de systèmes d'information la perspective d'un accès amélioré au contenu des documents numériques. Il s'agit d'une activité de recherche fondamentale qui tire sa substance d'applications en lien avec les enjeux sociétaux et économiques liés à l'accès à l'information. Mes activités applicatives bénéficient quant à elles de la consolidation des connaissances relatives à la modélisation de la construction du sens.

Au centre de ce thème de recherche se situe la mise au point de processus de structuration automatique permettant la confrontation de modèles linguistiques à un corpus. Mon ambition est ici est de systématiser l'utilisation de l'approche distributionnelle dans le traitement informatique du document. C'est également pour moi une manière de travailler sur des modèles portant des hypothèses de contraintes cognitives liées à la production et à l'interprétation du contenu du document, dans la lignée de Jacques Vergne (1999).

Sur le versant fondamental de cette recherche, l'objectif est la construction d'un processus de *macrostructuration du discours.* Les structures linguistiques que je cherche à calculer correspondent à différents niveaux d'organisation du discours dans les documents textuels formatés, niveaux d'organisation qu'il ne me semble pas possible d'appréhender séparément. C'est donc avec une *approche systémique* de la structuration du discours dans le document que j'effectue mes recherches.

La méthode de macrostructuration du discours que je mets au point prend appui sur les propriétés du document textuel formaté qui permettent à l'auteur de transmettre son discours, que ce soit en terme de restitution d'objets textuels, figuratifs ou tabulaires, qu'en terme d'organisation et de hiérarchisation de ces objets, au travers de leur mise en forme, de leur mise en page, ou de leur référence. C'est à ce titre que mes travaux sur la structuration des documents convergent vers une *analyse sémiotique automatique à vocation multilingue*.

Sur le versant applicatif de cette recherche, mes travaux participent à améliorer l'accès au contenu des documents textuels formatés en permettant la mise en œuvre de techniques d'analyse automatique basée sur la structure du discours, comme le filtrage de documents pour les systèmes de veille, l'alignement automatique contextualisé pour les systèmes de traduction automatique ou d'aide à la traduction, ou la construction automatique de tables des matières ou de représentations compactes pour l'aide à la navigation au sein de collections de documents numériques.

4.5 Les structures du document considérées

L'idée de *structure de document* est inhérente au concept même de document si bien qu'il ne saurait y avoir document sans structure de document. Le document, support de communication et objet porteur de sens, comporte différents niveaux d'organisation qui se traduisent par autant de structures contribuant à la transmission de l'information et à son interprétation. Aussi, Je ne puis que m'interroger sur le bienfondé des approches envisageant les textes comme « *non structurés* ». La structure physique et la structure logique du document ne sont que le reflet de deux regards portés sur le document, des regards aussi complémentaires qu'indissociables, des regards qui devront cependant en croisés d'autres, notamment ceux portant sur l'organisation du discours.

4.5.1 La structure physique du document

La *structure physique* du document est très certainement la structure première dans le sens où elle résulte du premier contact du lecteur avec le document, avant même qu'il en commence la lecture. La structure physique reflète l'organisation de l'information sur son support. Le modèle de structure physique envisage le document comme un volume, ou un ensemble de volumes, constitué d'une suite de pages. Cette structure décrit la mise en page des différentes pages du document, notamment les marges, le colonage, les entêtes et pieds de page. La structure physique est porteuse de deux contenus complémentaires, que j'appelle *enveloppe éditoriale* et *corps de document*. Ces deux sortes de contenus sont à mettre en relation avec les notions de *paratexte* et de *texte de* Genette (cité par Dupuis, 2009). Il va sans dire que dans cette perspective, *texte* n'est pas à interpréter au sens informatique puisqu'il est bien clair qu'ils ne se réduisent pas à une simple suite de caractères.

4.5.2 La structure logique du document

La *structure logique* du document reflète l'organisation du corps de texte en parties et en objets visuellement perceptibles, comme les figures ou les encadrés. La structure logique dépend d'un modèle logique de document qui a pour objet de refléter l'organisation hiérarchique des corps de document. Une première approche du modèle logique consiste à envisager le document comme organisé en parties consécutives, parties pouvant elles-mêmes être subdivisées en sous-parties en seconde approche. Au-delà de cette structure hiérarchique, le modèle logique reflète également des structures d'objets flottants, qui accompagnent la structure hiérarchique pour l'enrichir et permettre des parcours alternatifs dans une modalité propre à l'écrit (Derrida cité par Bachimont, 1998). Les objets flottants souvent composites, structurellement isomorphe, sont liés au genre textuel et prennent leur appellation dans ce contexte. On parle ainsi de tableaux et de figures dans les articles scientifiques, de cartes dans les manuels géographiques, d'encadrés dans les articles de vulgarisation...

4.6 Contribution à la structuration physique du document

Alors que l'état de l'art aborde la question de la structure physique de manière guidée par les données, je propose une **analyse de la mise en forme des pages guidée par le modèle**. Cette nouvelle approche de la structuration physique est une réelle avancée en terme d'interprétabilité de la structure produire.

Le modèle de mise en forme des pages peut être en quelque sorte envisagé comme un calque qui, placé sur une page donnée, va permettre de décider de son caractère interprétable. Soit le calque est complètement différent de la mise en page constaté et il n’est pas possible d’interpréter, soit le calque est cohérent et l’on peut se permettre d’interpréter. La force du modèle est de définir des positions indépendamment du contenu. Il permet par exemple de tester si du contenu apparaît ou

non dans la marge, définie comme une position, sans que la marge soit positivement matérialisée dans le document : elle est matérialisée comme espace vide.

Le calcul de la mise en forme des pages du document est basé sur l'utilisation conjointe de deux modèles abstraits : (1) le modèle périphérie–centre qui permet de mettre en relation un contenu périphérique et un contenu central, (2) et le modèle séquentiel qui permet de coordonner plusieurs contenus, c'est-à-dire de les placer sur un plan d'égalité, tout en proposant un mode de lecture, qui peut être du premier au dernier item ou bien du dernier item au premier item de la série.

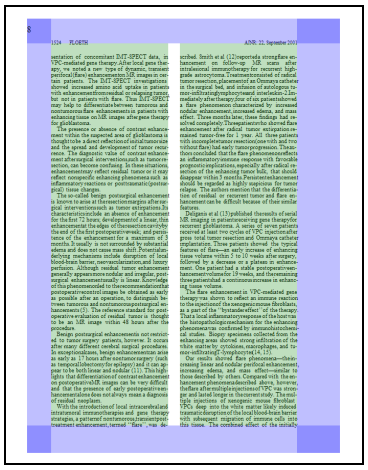


Figure : Illustration de l'instanciation du modèle périphérie-centre (bleu/blanc) et du modèle séquentiel (vert), après induction des traits communs de pages et superposition du résultat obtenu sur une des pages du document.

L'analyse de la mise en forme des pages du document est guidée par le modèle sous-jacent de mise en forme des pages. Ce modèle oppose les marges et le corps de page, selon le motif périphérie–centre. Il définit ces deux positions documentaires, la marge et le corps de document, permettant ainsi le classement des objets selon qu'ils se situent dans la marge, dans le corps de page, ou en chevauchement de la marge et du corps. Le corps de page est envisagé sous la forme d'un colonage, selon le motif séquentiel. Le colonage permet d'observer l'appartenance d'un objet à telle ou telle colonne et le chevauchement d'objets sur plusieurs colonnes.

Plusieurs modèles de mise en forme des pages peuvent cohabiter à l'intérieur d'un même document. La mise en forme des pages de préambule ou d'annexes peut par exemple être différente de la mise en forme des pages du corps de document. Si les pages du corps de document se prêtent aisément à l'induction d'un modèle de mise en forme associé, la première page d'un article, ou les premières et dernières pages d'un livre, ne sont pas toujours exploitables pour le calcul d'une instance de modèle. Il convient d'en tenir compte lors du traitement. Le processus d'induction d'un modèle de mise en forme des pages nécessite en effet plusieurs pages pour produire un résultat fiable – un minimum de trois pages de même mise en forme est souhaitable. Il ne peut être appliqué sur une ou seulement deux pages, à l'échelle d'un livre.

Le document étant considéré comme un tout structuré en parties consécutives, une hypothèse de séquentialité des pages de même mise en forme est prise en compte lors de l'induction des modèles de mise en page. Cette contrainte qui trouve sa justification dans la structuration même du document en parties consécutives doit cependant être gérée, parfois même relâchée. Le pré-supposé de séquentialité atteint notamment ses limites en frontière de parties, ou en présence de pages « flottantes », pouvant par exemple correspondre à des planches illustrées, ou à des pages de transition entre les parties d'un ouvrage. Ces pages flottantes de même mise en forme et réparties dans le document rompent la séquentialité du texte. Il convient d'en tenir compte tant pour calculer un modèle de mise en page sous l'hypothèse de séquentialité, que pour calculer le modèle de ces

pages particulières qui généralement ne sont pas contigües.

L'analyse du corps de chaque page est guidée par le modèle de mise en forme des pages associé à la page. Elle repose sur un modèle de corps de page. Ce modèle de corps de page est fondé sur l'opposition de deux classes d'objets : les objets dits normaux, internes à une colonne et les objets dits spéciaux, chevauchant plusieurs colonnes. Cette opposition va permettre l'identification des objets flottants (figures, tableaux, encarts) macroscopiques, c'est-à-dire qui ne sont pas internes à une colonne. Le calcul de la mise en forme du corps de page va produire un pavage composé de zones dites normales, et de zones dites spéciales. Les zones normales sont notées en vert, les zones spéciales sont mises en valeur en rouge.

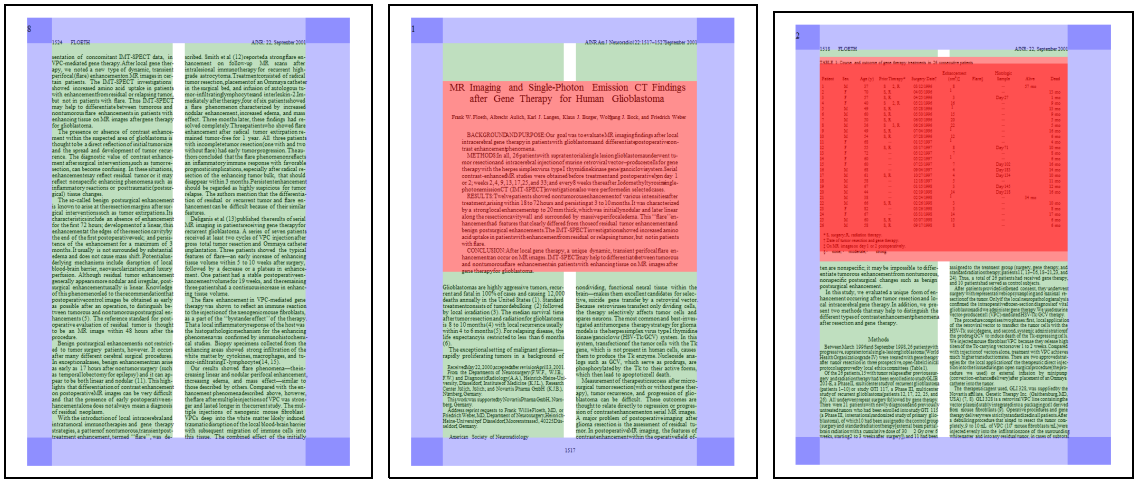
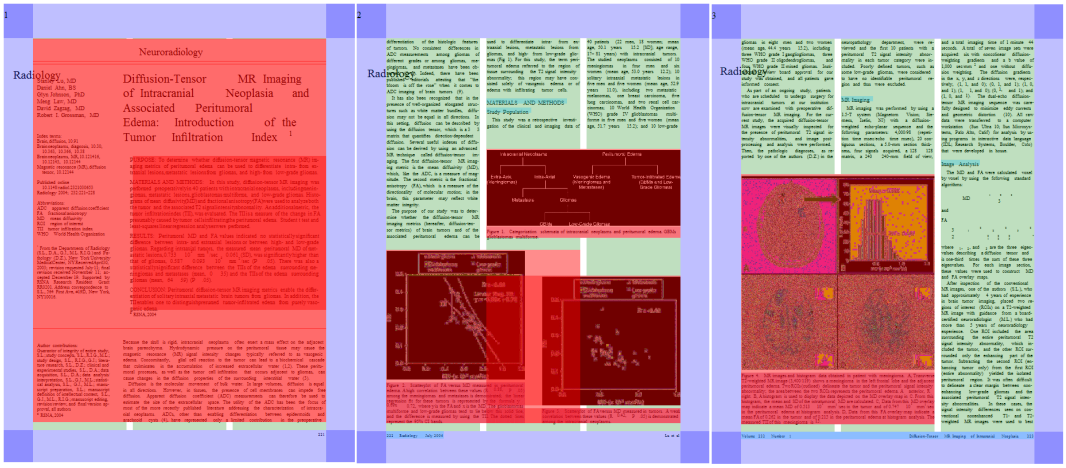


Figure : Le calcul de la structure du corps de page, guidé par le modèle de mise en forme des pages, met en évidence des zones « spéciales » (en rouge), cohabitant avec le colonage induit et attendu (en vert).

La compatibilité d'un modèle de mise en forme des pages avec la mise en page d'une page donnée est évaluée à ce stade. Intuitivement, on ne souhaite pas avoir beaucoup d'objets documentaires non compatibles avec la mise en page modèle. On calcule donc le pourcentage d'objets qui ne sont inclus dans aucune zone. En outre, intuitivement, on souhaite que les zones couvrent une bonne densité d'objets documentaires compatibles avec le modèle. On calcule donc le ratio de la surface des objets strictement inclus dans un type de zone par rapport à la surface totale des zones de même type. Il s'agit d'un diagnostic automatique de la qualité de la reconnaissance. En cas de qualité trop faible, le diagnostic permet de rechercher un modèle de mise en page adéquat.



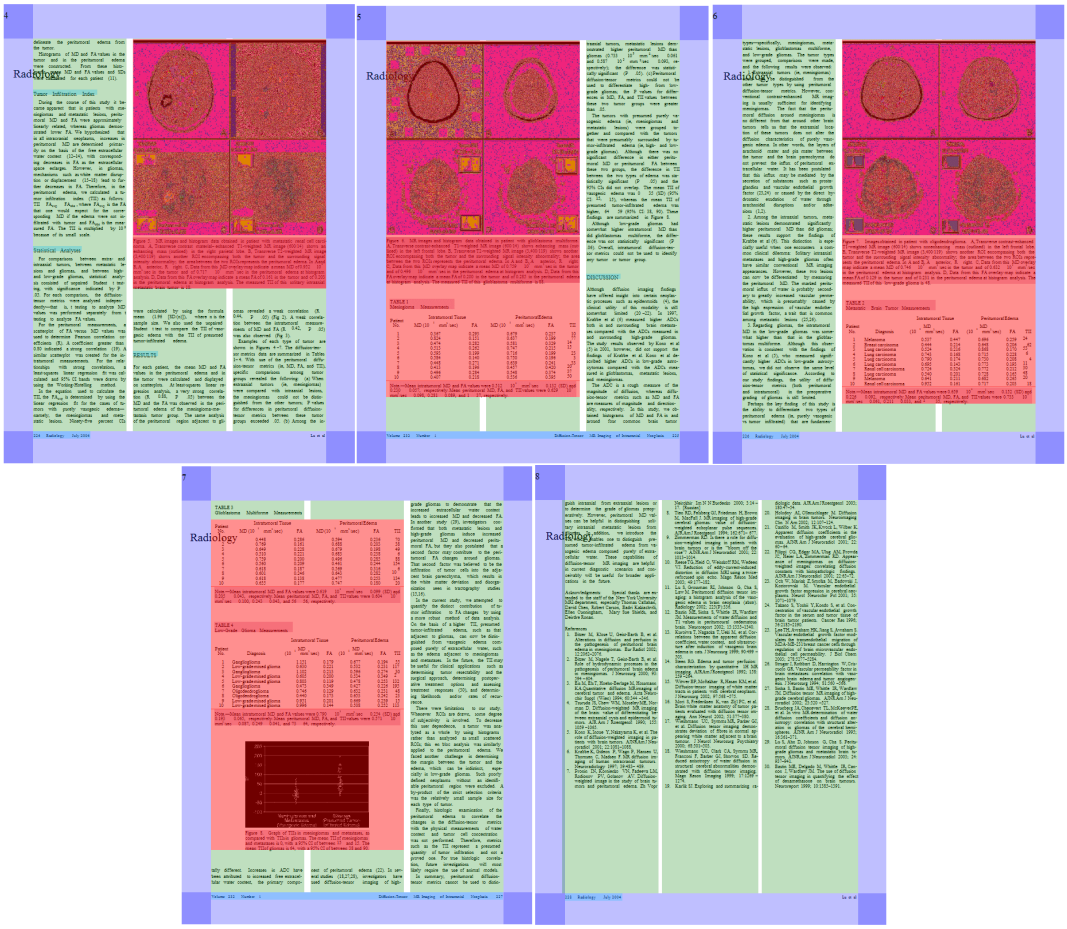


Figure : Le modèle de mise en forme des pages du corps de ce document est tri-colonne. Appliqué à l'ensemble des pages du document, il se révèle incompatible avec la page 1. L'analyse de la page 1 n'est donc pas interprétable avec ce modèle. Il faudra calculer un modèle de mise en forme spécifique.

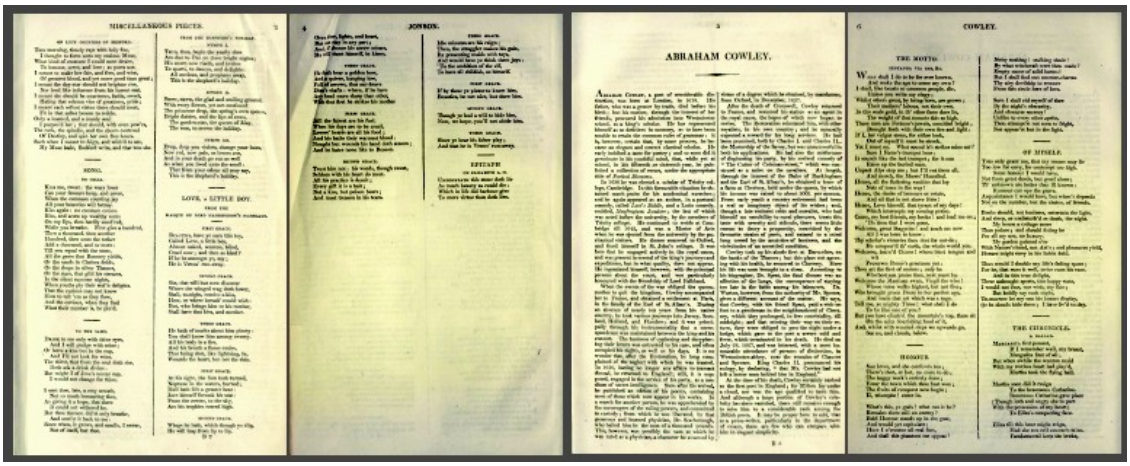
4.7 Contribution à la structuration logique automatique

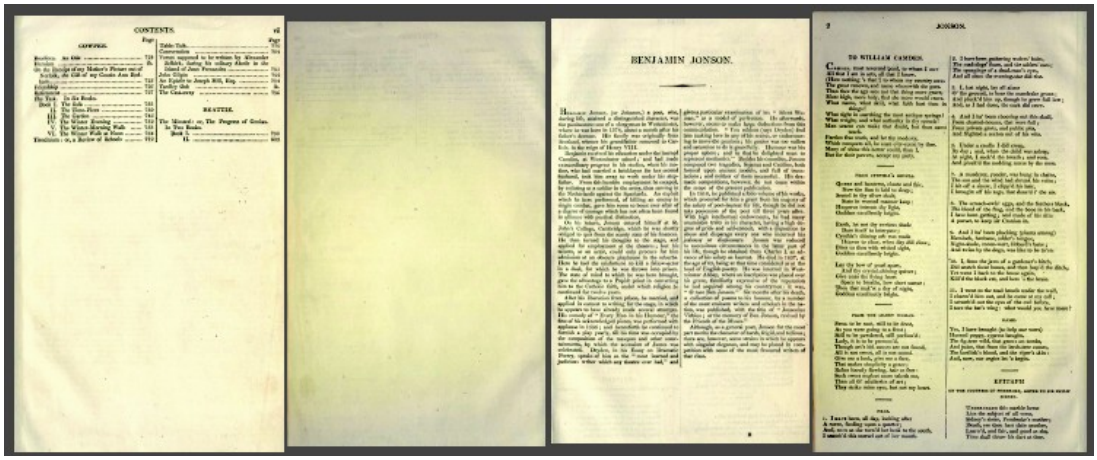
La méthode de structuration logique que nous proposons est à la fois efficace et légère. Elle systématise l'approche distributionnelle en recherchant des répétitions de mise en forme dans des fenêtres d'observations de différentes tailles, à l'intérieur de différentes positions du document. Nous avons testée cette méthode pour la détection de titres de partie et pour la détection de titres de chapitre. Elle est actuellement en cours de mise au point pour la détection des titres internes à la page (titres de section, légendes de figures et de tables)

Les titres de chapitre sont recherchés avec une fenêtre d'observation de quatre pages glissantes consécutives, glissant sur le corps de document. La mise en forme observée dans la fenêtre d'observation porte sur la répartition de la quantité d'information dans chacune des pages, la page du titre étant moins remplie en haut qu'en bas. La distribution attendue du motif observé doit être régulière sur l'ensemble des pages du corps de document.

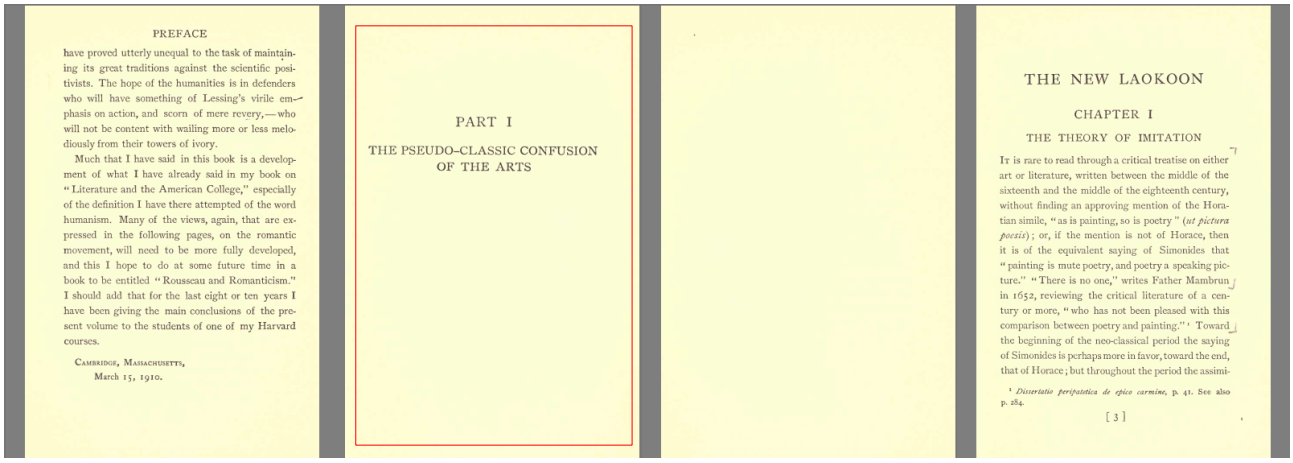
Les titres de parties sont recherchés avec une fenêtre d'observation d'une page, glissant sur l'ensemble du document. La mise en forme observée dans la fenêtre d'observation porte sur la répartition de la quantité d'information dans la page : un haut et d'un bas de page vide, et entre 1 et 3 lignes de texte globalement centrées. La distribution attendue du motif observé doit être régulière sur l'ensemble des pages du document.

Une contrainte relationnelle permet de vérifier la compatibilité de la distribution des parties et des chapitres, avec notamment des effectifs compatibles.





Exemple de deux fenêtres d'observation de 4 pages permettant la détection de titre de chapitre



Exemple de fenêtre d'observation d'1 page (en rouge) permettant la détection d'un titre de partie (1 pages de contexte à gauche et 2 à droite sont affichées mais non utilisées)

Cette méthode de structuration originale porte sur l'ensemble du document et non sur l'analyse d'une éventuelle table des matières, comme le proposent certains travaux de l'état de l'art. Elle peut donc gérer des documents n'en possédant pas, et peut également gérer des titres non présents dans les tables des matières (préface, table des matières) . Elle est dépendante de la mise en forme qui est très stable sur les corpus étudiés (des livres numérisés d'une part, des articles scientifiques d'autre part). Elle ne nécessite aucun lexique, elle est tolérante aux erreurs d'OCRs et est indépendante des langues, puisque ne s'appuyant pas sur le contenu textuel. Elle est par ailleurs simple et efficace. L'approche distributionnelle permettant l'identification des titres internes au corps de page (les titres de sections et les légendes) est en cours de mise au point.

4.8 Prospective

4.8.1 Vers un modèle cognitif de la structure logique du document

La mise en évidence d'une structure hiérarchique est au centre du processus de structuration logique. En informatique, la représentation et l'analyse de structures hiérarchiques, en l'occurrence, de parties consécutives pouvant elles-mêmes être subdivisées en sous-parties, fait immédiatement écho au concept de structure récursive. Cet écho résonne d'autant mieux que tout un arsenal de représentations, d'algorithmes, d'études de complexité est disponible.

La récursivité s'avère cependant inappropriée pour la modélisation des phénomènes linguistiques car beaucoup trop puissante. Tout d'abord, il convient de noter que la récursivité intègre un concept de développement potentiellement infini, concept qui certes contribue à la caractérisation des langages de programmation mais qui ne reflète en rien les caractéristiques du discours. Ainsi, on constate que la profondeur des structures logiques est corrélée au genre et à l'ordre de grandeur du document, et que cette profondeur est extrêmement limitée, peut-être devrais-je dire majorée, au sens mathématiques, dans le sens où il existe une borne supérieure. Ainsi dans un roman, on trouvera selon la taille de l'ouvrage tout au plus deux niveaux, la partie et le chapitre. Dans un article scientifique court, on trouvera communément deux niveaux de profondeur, et dans un article long on en trouvera plutôt trois, rarement cinq ou six.

La récursivité permet par ailleurs la modélisation d'un développement anarchique de structures hiérarchiques, ce qui va à l'encontre des caractéristiques du discours. Ainsi, il convient de noter que les parties situées en début et en fin de document sont généralement moins développées que les parties de corps de texte. Ainsi, dans un roman les liminaires sont généralement plats, c'est-à-dire non structurés en sous-parties, à l'inverse du corps qui lui l'est, en chapitres. Par ailleurs, l'observation montre que toutes les parties du corps de texte ne peuvent être structurées en sous-parties sur un même plan d'égalité. Ainsi, dans un article scientifique, on constate que des sections périphériques comme l'introduction, la discussion ou la conclusion, sont rarement structurées en sous-sections.

Enfin, il existe des rapports de forme dont ne rendent pas compte les structures récursives. Ainsi, dans un livre les parties liminaires représentent moins de pages que l'ensemble des parties du corps de texte. Nous reviendrons sur ces rapports de forme dans l'identification de la structure thématique des documents.

Bien entendu, cette inadéquation des structures récursives à modéliser des phénomènes linguistiques n'est pas une entrave à l'usage de techniques informatiques récursives dans un processus d'identification. Il convient cependant de tenir compte des propriétés du modèle linguistique pour produire des contraintes et éviter de produire des solutions non justifiées.

Le modèle de structure logique de document que nous cherchons à définir n'est donc pas complètement récursif. Si plusieurs critères de bonne formation de la structure logique ont été formulés et traduits sous forme de contraintes dans les implémentations, il reste encore un travail de recherche important pour consolider ce travail.

4.8.2 Réflexions sur les format de documents, la perception par le lecteur, et les conséquences pour le calcul de la structure

Les collections de documents auxquels je m'intéresse sont issues de sites internet ou de bibliothèques numériques accessibles en ligne.

La collection de livres numérisés comporte 2000 ouvrages, principalement en anglais. Cette collection a été mise à disposition dans le cadre des campagnes d'évaluation internationales *ICDAR Book structure evaluation* auxquelles je participe depuis trois ans. Il s'agit d'une collection de livres composés en imprimerie qui ont été intégralement numérisés et océrisés, couverture comprise, puis diffusés au format PDF. Chaque page du document numérique contient l'image de la page originale ainsi que le résultat de l'océrisation. L'océrisation est de qualité variable, fonction de la présence de taches, d'annotations manuelles, de caractères mal imprimés, de caractères de forme non reconnue, ou d'un mauvais positionnement de la page sur le scanner.

La collection d'articles scientifiques contient 300 articles principalement en anglais provenant de la base Medline. Cette collection a été constituée dans le cadre du projet ANR Bingo (Bases de données INductives et GénOmique), dont le laboratoire GREYC est partenaire. Elle contient exclusivement des articles scientifiques médicaux. Cette base à l'origine exclusivement en anglais a été enrichie de quelques articles en chinois, allemand, français et espagnol. Les articles diffusés au format PDF ont été rédigés à l'aide de traitements de texte, pour les plus récents, et sont diffusés au format PDF.

De la perception par le lecteur à l'analyse automatique des structures du document

Contrairement à notre intuition de lecteur qui perçoit naturellement la structure des documents au travers de leur mise en forme, l'ordinateur n'a généralement pas la capacité à manipuler la structure d'un document qu'il présente ou dont il extrait de l'information. Pour la machine, tout n'est qu'octets, caractères, images ou formes graphiques, disposés dans un espace particulier, avec une mise en forme particulière. La structure, pourtant si évidente, si tangible pour le lecteur qui ne peut prendre en main un document sans l'appréhender, est le résultat d'un processus cognitif complexe qui échappe le plus souvent aux traitements informatiques d'aujourd'hui.

Dans les documents électroniques au format PDF par exemple, le colonage, le paragraphage, la titraison, les entêtes, les pieds de page ou même les marges ne sont pas systématiquement représentés. Il en est de même des concepts de figure, de légende, de note de bas de page, ou d'encadré, qui ne sont nullement matérialisés comme tels. Ceci n'entrave en rien un rendu de qualité. En tant que format de visualisation et d'impression, PDF exploite les informations de mise en forme nécessaires au rendu fidèle du texte d'origine et n'intègre pas le meta-langage utilisé pour sa production (les nouvelles versions de PDF commencent à l'intégrer). Le parti pris de PDF est de se reposer sur les facultés de perception et d'interprétation du lecteur.

Perception et interprétation des documents en tant qu'images

Dans la philosophie de conception du format PDF, l'ordinateur n'a pas à manipuler les concepts de marge ou le concept de colonne de texte pour réaliser l'affichage ou l'impression d'un document. Il suffit en effet de disposer des mots écrits en noir sur un fond blanc de manière à former un rectangle pour que le lecteur perçoive une colonne de texte entourée de marges.

Seuls suffisent le concept de page et d'objets textuels, avec pour chaque page, ses dimensions et son orientation, les coordonnées des objets à afficher sur la page, et les attributs de mise en forme de ces objets ; la colonne de texte en tant qu'objet n'existe à aucun moment dans l'ordinateur. Elle n'est que construction intellectuelle résultant de la perception visuelle de l'interprétant au travers du jeu des régularités et des différences de mise en forme, des rapports de forme attendus, de modèles d'organisationnels connus. En l'occurrence, le concept de colonne de texte se forme dans l'esprit du lecteur par la différence de texture entre un intérieur rempli de mots écrits en noir et interprété comme le texte de la colonne, et un extérieur blanc interprété comme la marge ou l'entre-colonne, l'intérieur étant associé à une forme géométrique abstraite de type rectangle et qui se concrétise par des lignes de texte parallèles, et une orthogonalité de lignes virtuelles droite et gauche créées par l'alignement relatif à droite et à gauche de mots situés respectivement en début et en de fin de ligne.

Nul besoin en effet que l'ordinateur ait à disposition l'information sur le statut de titre ou de figure de tel ou tel objet pour l'afficher ou l'imprimer correctement. Nul besoin que l'ordinateur indique au lecteur que tel ou tel objet est un titre ou une figure pour que celui-ci le déduise en contexte. Les informations représentées dans le format PDF suffisent au rendu et n'entravent en rien l'interprétation du lecteur qui perçoit naturellement la structure du document. Cependant, dans la perspective d'un traitement informatique qui exploiterait la hiérarchie des titres ou bien la présence de figures à tel ou tel endroit du document, la structure du document qui n'est pas directement manipulable car non représentée doit être préalablement calculée par l'ordinateur.

Représentation de la structure dans les formats traitement de texte

Dans les documents au format Microsoft Word (DOC ou DOCX) ou au format OpenDocument Text (ODT), formats de traitement de texte et non de diffusion, la situation est différente et a beaucoup évolué ces dernières années. Les informations sur la structure du texte étaient tout d'abord absentes du document. C'est par l'intermédiaire des attributs de mise en forme appliqués directement sur le texte que l'utilisateur pouvait rendre compte de la structure de son document. Parmi les attributs de mise en forme à disposition, on trouve notamment la police de caractères utilisée, les effets appliqués sur les caractères (graisse, italique, couleur, surlignage), l'alignement du texte (à droite, à gauche, centré ou justifié), les retraits de début et de fin de ligne, les espacements avant et après. Ces informations de mise en forme permettant de rendre compte de la structure ont par la suite été partiellement détachées du contenu et rendues manipulables sous la forme de feuille de style. Un style rassemble les attributs de mises en forme applicable à une même catégorie d'objets textuels (les titres, les légendes, les notes de bas de page par exemple). Ces styles sont manipulables pour l'auteur qui peut les appliquer sur son texte via l'interface utilisateur et plus facilement garantir la cohérence de son manuscrit, ainsi que la cohérence de tous ses écrits de même genre. Ils sont également manipulables pour la machine qui peut ainsi créer un affichage du plan ou de la table des matières à partir de la hiérarchie des titres. Les séries d'objets documentaires ont également été intégrées au format (figures, tableaux, équations par exemple) et sont rendues manipulables tant par l'utilisateur que par l'ordinateur.

On constate cependant que c'est principalement dans les communautés organisées ou dans les situations contraintes, comme le monde scientifique ou l'édition numérique, que des contraintes de production sont effectivement respectées. En effet, de même qu'il peut sembler absurde pour un auteur d'avoir à indiquer la langue dans laquelle il écrit, il n'est pas non plus naturel d'avoir à expliciter le métalangage documentaire et le statut de « section » ou de « figure » de tel ou tel objet, de surcroît en cours d'écriture.

Alors que la technologie WYSIWYG avait été une innovation majeure en matière de transparence et d'appropriation par les utilisateurs, l'immaturité des technologies d'identification de structures équivaut à un frein technologique. Là où l'auteur ne devrait écrire que pour ses lecteurs, il

se voit contraint à écrire également pour la machine. En l'absence d'une technologie permettant l'identification de la structure d'un document et la production d'une représentation manipulable, il n'est pas possible de construire des programmes informatiques hiérarchisant, relativisant l'information. C'est l'objet de l'identification automatique de la « structure logique » du document que de construire une telle structure.

4.9 Conclusion

Alors que l'état de l'art aborde la question de la structure physique de manière guidée par les données, nous avons proposé une analyse de la mise en forme des pages guidée par le modèle. Cette nouvelle approche de la structuration physique est une réelle avancée en terme d'interprétabilité de la structure produire. Concernant la structuration logique, nous proposons une méthode à la fois efficace et légère qui systématise l'approche distributionnelle en recherchant des répétitions de mise en forme dans des fenêtres d'observations de différentes tailles, à l'intérieur de différentes positions du document. Enfin, nous avons proposé un principal axe de recherche, portant sur la conception d'un modèle cognitif de la structure logique du document, et qui nous paraît prometteur.

4.10 Publications liées

GIGUET, Emmanuel, LUCAS, Nadine, et COUSIN, Grégoire. 2000. Document structure identification as a means for relevant indexation. *International Conference on Intelligent text processing and Computational Linguistics (CICLING-2000).* Mexico, February.

GIGUET, Emmanuel. 2008. *Rapport scientifique du projet Résurgence*. Rapport interne, Groupe de Recherche en Informatique, Image, Instrumentation et Automatique de Caen. Octobre.

GIGUET, Emmanuel, LUCAS, Nadine, et Chircu, Catalina. 2008. Le projet Résurgence : Recouvrement de la structure logique des documents électroniques. *JEP-TALN-RECITAL'08 Session "Show & Tell"*, juin, France.

GIGUET, Emmanuel, BAUDRILLART, Alexandre, et LUCAS Nadine. 2009. "Resurgence for the Book Structure Extraction Competition." Paper presented at the INEX, Woodlands of Marburg, Ipswich, Queensland, Australia, December 6–10, pp. 136-142.

GIGUET, Emmanuel, et LUCAS, Nadine. 2010. The Book Structure Extraction Competition with the Resurgence Software for Part and Chapter Detection. *INEX 2010 pre-proceedings*, Amsterdam.

GIGUET, Emmanuel, et LUCAS, Nadine. 2010. "The Book Structure Extraction Competition with the Resurgence Software at Caen University." *Inex 2009*, edited by Shlomo Geva, Jaap Kamps and Andrew Trotman, pp. 170-178. Heidelberg: Springer.

GIGUET, Emmanuel, et LUCAS, Nadine. 2011. "The Book Structure Extraction Competition with the Resurgence Software for Part and Chapter Detection at Caen University." *Comparative Evaluation of Focused Retrieval: 8th International Workshop of the Initiative for the Evaluation of Xml Retrieval (Inex 2010)*, ed. by S. Geva, J. Kamps, R. Schenkel and A. Trotman. Berlin / Heidelberg: Springer.

LEJEUNE, Gaël, BRIXTEL, Romain, et GIGUET, Emmanuel. Deft 2011: Appariement de résumés et d'articles scientifiques fondé sur des distributions de chaînes de caractères. *DEFT 2011*, Montpellier. À paraître.

LUCAS, Nadine, et GIGUET, Emmanuel. 2008. Robust adaptive discourse parsing for e-learning fora. *The 8th IEEE International Conference on Advanced Learning Technologies*. July 1st- July 5th. Santander, Cantabria, Spain. Learning technologies in the Information society.

5 Synthèse des travaux sur l'analyse des forums de discussion

Les forums de discussion et autres nouvelles formes de communication écrite telles que les listes de discussion qui se développent avec les moyens électroniques posent plusieurs problèmes de manipulation et d'analyse interprétative. Sur le plan technique, ils posent le problème des documents volumineux. Sur le plan linguistique, ils posent le problème du registre et du style collectif (Beaudoin, 2002). On y observe en effet un type d'écrit à la fois informel et polyphonique (Marcoccia, 2010).

Le traitement des langues, qui s'est longtemps focalisé sur l'analyse d'écrits propres et lissés, rencontre là un véritable obstacle. Les outils d'analyse des langues, même robustes, peinent à analyser les productions dont la qualité est aussi éloignée de l'attendu. Les techniques de correction orthographique ou de ré-accentuation, parfois utilisées en prétraitement, sont inopérantes.

L'attente forte d'un style rédactionnel soutenu explique la dégradation des résultats des logiciels lorsqu'ils sont confrontés à ces nouvelles formes de communication. Face à tout un ensemble de techniques simultanément malmenées, c'est sur les hypothèses fondamentales du traitement des langues que nous nous interrogeons.

Dans la perspective de traitements informatiques pouvant prendre en charge des productions humaines protéiformes, nous présentons ici un des aspects de nos travaux réalisés dans le cadre d'un projet national, Calico, détaillé plus bas. Nous mettons l'accent sur l'analyse des discussions dans les forums, afin d'aborder deux questions : la prise en compte des ordres de grandeur, et la prise en compte du style.

5.1 Les forums de discussion

Revenons sur les bases. Sur Internet, un forum de discussion est un espace de discussion accessible publiquement ou par authentification. La possibilité de discuter dans un tel espace correspond à un service offert par un logiciel de gestion des discussions et de l'espace du forum.

Les discussions sont archivées par le système, ce qui permet une communication asynchrone. Les actions des utilisateurs, comme la lecture de message ou la réponse à un message, sont également en partie archivées. Ces informations sont horodatées et servent à informer le lecteur de l'activité de telle ou telle discussion, ainsi qu'à effectuer divers traitements statistiques. Le système de forum prévoit l'attribution de rôles aux participants, qui dépendent du contexte (modérateur, utilisateur authentifié, visiteur, enseignant, étudiant, …), et qui confèrent des droits d'utilisation différenciés (possibilité de lire uniquement, possibilité d'initier des discussions, …).

Anne Lavallard (2008), membre de notre équipe, a réalisé une étude approfondie de l'organisation des forums de discussion. Cette étude témoigne de la diversité des points de vue sur l'objet. Couramment, les forums sont organisés en thématiques ou sous-thématique de discussions, et les discussions sont appelés « fils de discussions ». Ils sont initiés par un message auquel répondent les autres participants intéressés. Le forum offre donc naturellement deux modes de parcours, l'un chronologique, par l'horodatage des message, l'autre hiérarchique, au travers de l'organisation thématique du forum et des fils de discussion.

La taille des forums de discussion est très variable, selon l'objectif qu'il vise et la popularité qu'il rencontre. Ainsi dans le domaine de la formation à distance, un forum lié à un groupe d'étudiants pourra totaliser quelques dizaines de messages, alors qu'un forum lié à une formation pourra en compter plusieurs centaines. Dans le jeu en ligne EVE ONLINE (http://www.eveonline.com/ingameboard.asp), le forum comptait plus d'un million quatre cents mille messages début juin 2011. On comprend alors que l'analyse des forums de discussion soulève tant la question de la gestion du style d'écriture que la question de la gestion des ordres de grandeur.

5.2 État de l'art

L'analyse de l'activité sur les forums est une première contribution à l'analyse des forums de discussion. Elle est basée sur les statistiques portant sur le nombre et le type des messages lus, postés, répondus, sur des périodes de temps à définir. Bratitsis et Dimitracopoulou (2007) intègrent de nombreux outils de la sorte dans leur plateforme DIAS. Cette analyse peut être comparative : entre participants, entre discussions, entre forums, et à différents moments.

Une deuxième contribution à l'analyse des forums de discussion concerne l'analyse des interactions entre participants. Ces recherches ont pour objet d'étude le graphe des interactions entre individus. Ces recherches visent à identifier des schémas d'organisation de groupe de participants à partir de la forme du graphe, ainsi qu'à modéliser l'évolution de l'organisation du groupe en fonction de l'évolution de la forme du graphe (de Laat, 2007). Ces travaux bénéficient des recherches portant sur la théorie des réseaux sociaux. L'analyse peut tenir compte ou ignorer le rôle des individus, ainsi que le type des interactions entre individus (envoyer un message à, lire un message de, …).

Les forums de discussion ne sont pas traités en tant que tels par les méthodes de traitement des langues. Les analyses classiques sont en effet faites au niveau de la phrase, et présupposent un style soutenu, au moins que les mots soient correctement orthographiés. Or dans le registre des forums, les limites de phrase ne sont pas fiables, et l'orthographe est peu respectée (spécialement en français qui a une orthographe difficile). Les analyses statistiques lexicales, le plus souvent appuyées par des systèmes d'apprentissage automatique, sont les plus répandues (Cress, 2008 ; Romero et Ventura, 2010).

5.3 Émergence de la thématique

Mes travaux sur l'analyse des forums de discussion s'inscrivent dans le cadre d'une ERTé, Équipe de Recherche en Technologie de l'Éducation, qui a coopéré sur une période totale de quatre ans, de 2006 à 2009. L'ERTé CALICO – Communauté d'apprentissage en ligne instrumentation, collaboration – a pour objet la recherche sur les formations à distance ou partiellement à distance, c'est-à-dire des formations comportant des phases à distance et des phases en présence, et qui intègrent des modalités de travail collaboratif. Son but est d'améliorer la compréhension des conditions de bon fonctionnement de l'apprentissage collaboratif (Bruillard, 2010). Le projet, dirigé par Éric Bruillard, professeur à l'UMR STEF, a réuni 5 laboratoires de recherche (STEF, LIUM, SaSo, CREAD, GREYC) et 6 IUFMs (Caen, Créteil, Nantes, Rennes, La Réunion, Rouen).

5.4 Positionnement

Avec Nadine Lucas, nous avons choisi d'aborder la problématique de l'analyse des forums par la lisibilité des visualisations produites. C'est d'une part une préoccupation majeure pour les utilisateurs d'outils d'analyse qui n'ont pas le temps de lire l'intégralité des discussions (Romero et Ventura, 2010). C'est d'autre part une confrontation avec la réalité de l'objet forum, qui peut contenir aussi bien quelques dizaines de messages que plusieurs milliers. Notre objectif est de parvenir à

construire des représentations globales et synthétiques, qui restent lisibles sur un écran d'ordinateur, quelle que soit la taille du forum. Il s'agit en fait d'approfondir la question de la prise en compte des ordres de grandeur de document et la question de la sélection de la fenêtre d'observation.

Parallèlement, nous souhaitons également ouvrir des perspectives en terme d'analyse de contenu. Notre hypothèse est qu'une structuration automatique du discours est envisageable sur les forums de discussion, quel que soit le style ou la langue, pour peu que l'on remette en cause l'hypothèse fondamentale de primauté des formes dans le traitement.

5.5 Principales contributions

5.5.1 Contribution à la prise en compte des ordres de grandeur

La communauté de l'analyse automatique de forum ne s'est que très peu intéressée à la question de la gestion des ordres de grandeur pour produire des résultats lisibles. La raison réside très certainement dans le fait que les communautés d'apprentissage en ligne ou les communautés du travail collaboratif ont affaire à des forums de discussions restant dans un ordre de grandeur raisonnable.

Anne Lavallard (2008) s'est confrontée à l'analyse de forums de taille variée, en étudiant les forums de jeu en ligne et les forums des communautés d'apprentissage en ligne. La question de la taille des forums est soulevée. Comme dans beaucoup d'autres approches, le problème est reporté sur l'utilisateur qui doit choisir les bons paramètres d'affichage pour obtenir un résultat lisible, en filtrant les messages par différents critères comme l'intervalle de dates.

Il faut se tourner vers les réseaux sociaux et vers les travaux récents de Seifi *et al.* (2010) pour trouver une proposition de « visualisation interactive multi-échelle des grands graphes », appliquée à un réseau de blogs. L'approche est guidée par les données et la structuration est ascendante, pour conserver la topologie du réseau.

Si notre objectif est similaire à celui de Seifi *et al.* (2010), nous l'abordons par une autre voie. La méthode d'analyse que nous proposons est guidée par un modèle d'interprétation et la méthode de résolution est descendante. Le premier intérêt de cette approche réside dans le fait que le modèle est garant de la lisibilité du résultat, puisqu'il définit un nombre fixe et raisonné de constituants en relation au niveau macro (couvrant le document). Le second intérêt réside dans le fait que le modèle construit pas les linguistes permet de juger immédiatement de l'interprétabilité ou de la non-interprétabilité du résultat.

Le modèle linguistique choisi est un modèle de structuration thématique, basé sur une opposition thème/rhème. Ce modèle théorique est proposé par le linguiste japonais Yamada, cité par Nadine Lucas (2005). La première mise en œuvre de ce modèle dans un traitement informatique portait sur l'analyse de textes courts en français. Le logiciel Thema a été réalisé par Pascalie Pinatel (2003), encadrée au sein de notre équipe par Nadine Lucas.

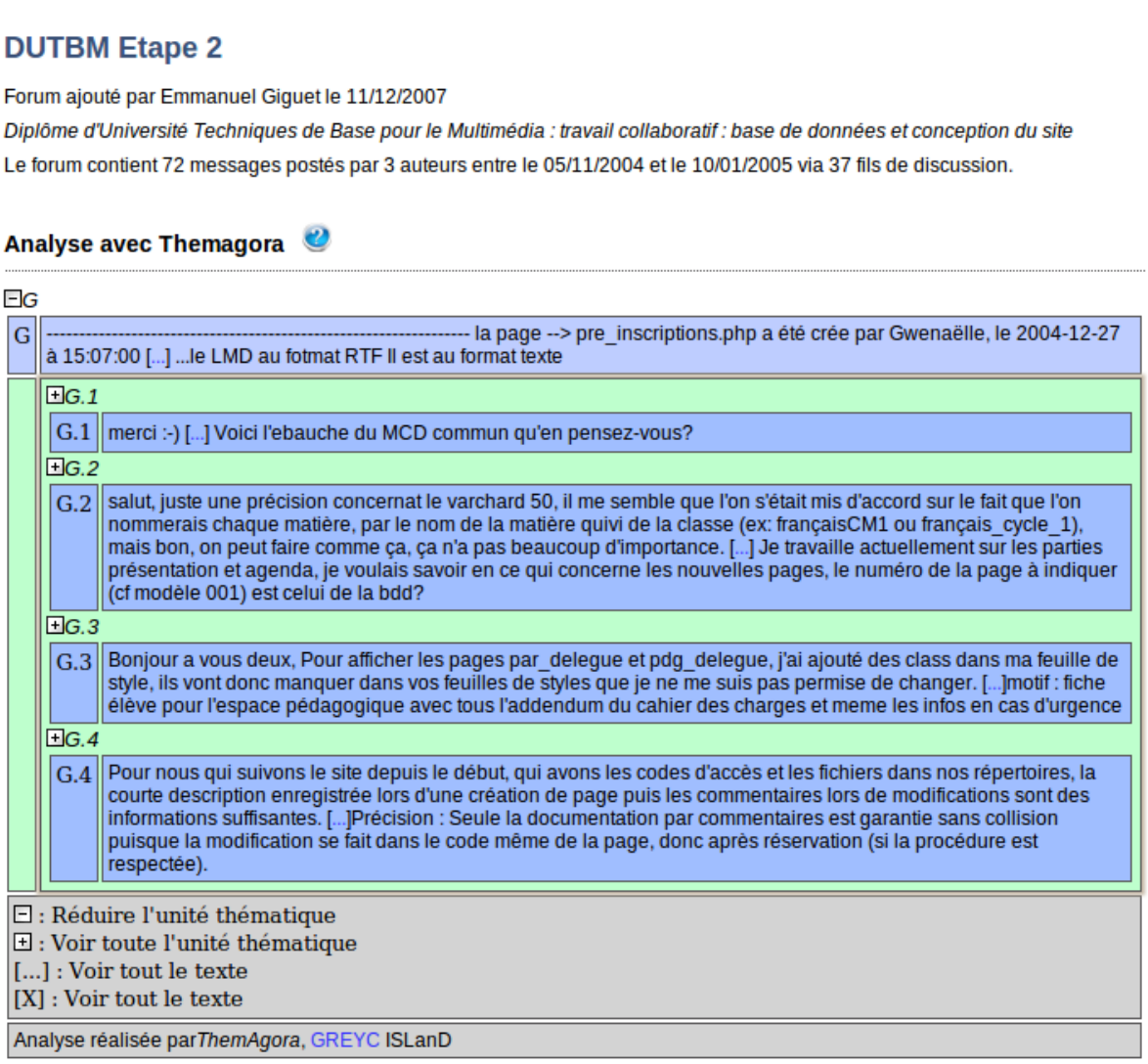
Nadine Lucas a adapté le modèle et nous avons réalisé la mise en œuvre multi-échelle, baptisée Themagora, et permettant d'analyser des forums de discussion de différente taille en intégrant une prise en compte automatique de l'ordre de grandeur (Lucas et Giguet, 2008 et 2010). Cette version multi-échelle a été réalisée à partir de la version multilingue, baptisée UniThem, que Nadine Lucas et moi-même (2005) avons mis en œuvre pour l'analyse d'un corpus de presse internationale.

Le **principe de la résolution multi-échelle guidée par le modèle** repose sur l'utilisation de l'approche distributionnelle et sur la sélection automatique de la taille de la fenêtre d'observation. La fenêtre est sélectionnée automatiquement de sorte que sa taille soit compatible avec l'ordre de

grandeur du document. Dans le cas des forums de discussion, la fenêtre d'observation peut être le fil de discussion, le message, ou une partie de message, selon la taille du forum.

Le modèle impose une segmentation inégale du document en deux parties successives, le thème puis le rhème. L'algorithme consiste alors à chercher dans le document et à l'aide de la fenêtre d'observation sélectionnée automatiquement, une distribution de marques qui met en évidence la frontière entre le thème et le rhème. Une approche similaire est ensuite appliquée pour séparer le rhème en un nombre raisonné de sous-thèmes.

Exemple de petit forum de discussion (72 messages) analysé par Themagora :



Dans cet exemple, le forum de discussion DUTBM, composé de 72 messages et 35 fils de discussion est analysé par Themagora. La vue synthétique tient sur un écran. La macro-structuration thématique thème/rhème est représentée par un bloc bleu, le thème, au-dessus d'un bloc vert, le rhème. Themagora a structuré le rhème en 4 parties appelées sous-thèmes (en bleu). La vue est interactive et permet à la demande d'observer la structure thématique de ces parties.

Exemple de forum moyen (244 messages) analysé par Themagora :

OS Concepts

Forum ajouté par Emmanuel Giguet le 05/12/2007

Forum téléchargé sur la plateforme phpBB du Department of Mathematical and Computer Sciences at the Colorado School of Mines.

Le forum contient 244 messages postés par 22 auteurs entre le 30/07/2006 et le 28/11/2007 via 77 fils de discussion.

Analyse avec Themagora

⊟G

G	When one thread opens a file, do other threads see the file? Same as with ULTs, all the threads can see the files opened by one thread. Does the OS choose a thread, a process, or both to execute? The OS chooses a process, then chooses a thread within the process to run. [...]If a writer arrives before V(wrt) is called it will be forced to wait since P(wrt) is called at the beginning of Writer (making wrt=-1). However, readers can continue to arrive which prevents V(wrt) from being called and causes writers to starve. In this situation we will have deadlock when a philosopher picks up the chopstick to his/her left when the philosopher to his right does the same. If this continues around the circle, each philosopher will have one chopstick and no one will be able to eat.

⊞G.1

G.1	The first example with semaphores: [...]There is also a problem with starvation where its possible that someone will never get more than one chopstick and keep giving any chopsticks they get back.

⊞G.2

G.2	Critique of paging: [...] Yeah, I also added the export line to my .bashrc (and ran it in the terminal) before I tried make and I got the same results as Brandon. _____ -Sara McF

⊞G.3

G.3	Teresa Davies wrote: I've found that when I ssh into alamode, I have to set LD_LIBRARY_PATH every time. [...]The remainders should be the offsets.

⊞G.4

G.4	Consider a memory system with following parameters? [...]One example of temporal locality in the code is when one value of i is used multiple times in the second loop.

⊞G.5

G.5	I had an alamode account like 2 years ago that I rarely used so as a result, I don't remember the password or if it is even still active. [...] Instead, it should be a major emphasis to use the DMA module, and only use the CPU for direct access to the disk when it requires it - implying the DMA controller should have much higher priority to disk access than the CPU. _____ Ryan Carpenter ADS Student Consultant/Mines Help Center rcarpent@mines.edu 719/930-3412

⊞G.6

G.6	For equation 1.1... The 2-level equation is Code: T_s = H*T_1 + (1-H)*(T_1+T_2) [...]And the same thing applies to why we have 200 GB hard drives instead of 200 GB of main memory (this assumes that the memory is made so that it doesn't lose its data on shutdown, like flash memory).

⊞G.7

G.7	SeanB wrote: I had an alamode account like 2 years ago that I rarely used so as a result, I don't remember the password or if it is even still active. [...]Wasn't there something on that test about how little endian and big endian only deals with memory storage and doesn't change how values are stored in registers?

⊞G.8

G.8	What is the purpose of System Calls and how do system calls relate to the operating system and to the concept of dual-mode (kernel mode and user mode) operations? System calls allow user-level processes to request services from the operating system that the process itself is not allowed to do. These system calls are classified into broad six categories - Filesystem, Process, Scheduling, Interprocess communication, Networking, and Miscellaneous. [...] I would like to add that this is functional abstraction at work and that a consistent interface is presented to userland applications (and also those that implement userland applications) making communicating with the kernel and subsequent hardware much easier. _____ -Kenneth Melby III "Progress isn't made by early risers. It's made by lazy men trying to find easier ways to do something." - Robert Heinlein

Dans cet exemple, la macro-structuration thématique porte sur 244 messages et tient sur l'écran.

5.5.2 Contribution à la prise en compte du style

Le traitement des langues, qui s'est longtemps focalisé sur l'analyse d'écrits propres et lissés, rencontre avec l'analyse des forums de discussion un véritable obstacle. Les outils d'analyse des langues, même robustes, peinent à analyser les productions dont la qualité est si éloignée de l'attendu. Selon nous, l'attente trop forte d'un style rédactionnel soutenu explique la dégradation des résultats.

C'est en remettant en cause l'hypothèse fondamentale du traitement des langues qui consiste en la primauté de la forme sur la relation que nous proposons une solution. Notre méthode repose sur l'utilisation de la méthode distributionnelle. La mise en relation des distributions et la préservation des proportions sont à la base de la méthode de structuration du discours que nous proposons.

L'idée est de privilégier des contraintes relationnelles fortes, basées sur la position des segments à relier, pour relâcher les contraintes de forme portant classiquement sur les segments reliés. Dans l'approche de la structuration thématique guidée par le modèle, le modèle linguistique est basé sur l'opposition thème/rhème. Le modèle envisage une succession du thème et du rhème, ainsi qu'un rapport de forme attendu entre le thème et le rhème : le thème doit avoir une taille inférieure et compatible avec celle du rhème. Le modèle n'impose pas de contraintes sur la forme des marques.

Pour traiter les forums de discussion, nous effectuons la recherche de marques, à l'aide de critères non lexicaux, à savoir ponctuations fortes (points d'interrogation, d'exclamation, de suspension, sauts de paragraphe), smileys, les nombres de constituants graphiques facilement identifiables : nombre de caractères, de mots s'ils sont délimités, de parties de messages, de messages, par exemple les citations « emblématiques » (Lucas, 2011) ou de messages particuliers (par exemple les relances par les tuteurs). Sans contrainte sur le lexique, la méthode est alors naturellement tolérante aux fautes d'orthographe, qui ne seraient visibles qu'à un grain très fin, et sensible au style qui se manifeste en partie dans les forums via les attributs que nous utilisons.

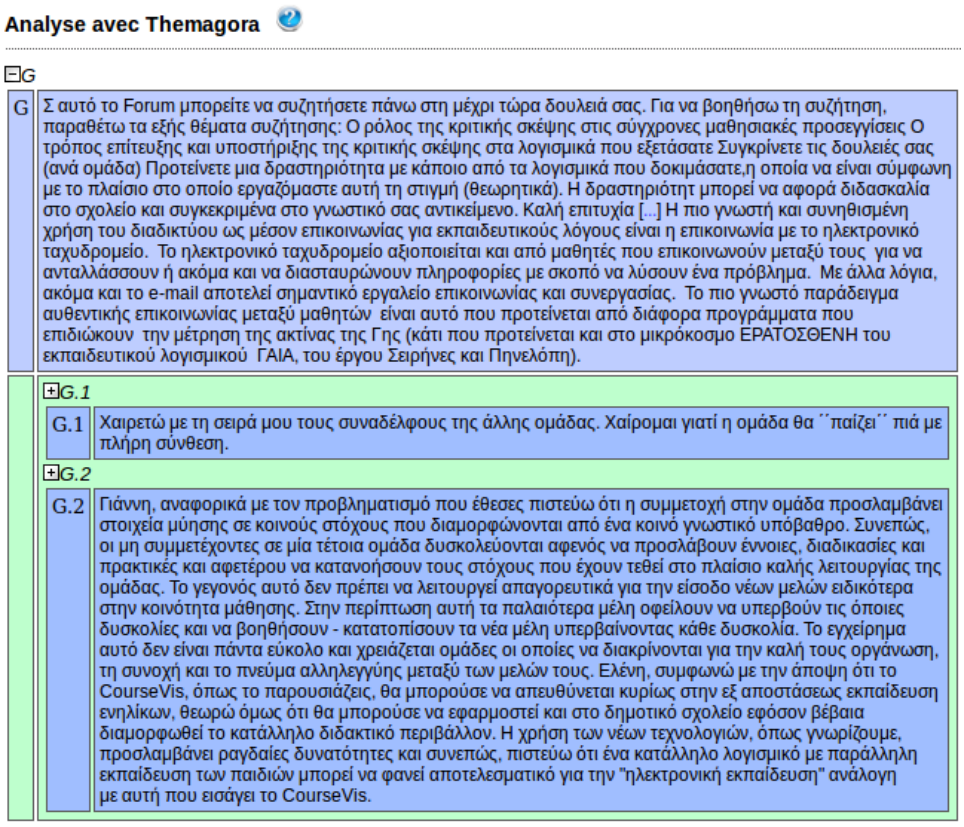
5.5.3 Contribution à la prise en compte de la variation en langues

Pour des raisons similaires à celle de la prise en compte du style, l'approche distributionnelle que nous utilisons privilégie l'utilisation de contraintes relationnelles fortes, pour relâcher les contraintes portant sur la forme. Elle est ici mise en œuvre en combinaison avec une recherche de marques non lexicales. Cette stratégie aboutit naturellement à une méthode de structuration automatique du discours guidée par le modèle, non dépendante de l'alphabet et du lexique.

5.5.4 Réflexions sur la prise en compte des ordres de grandeur

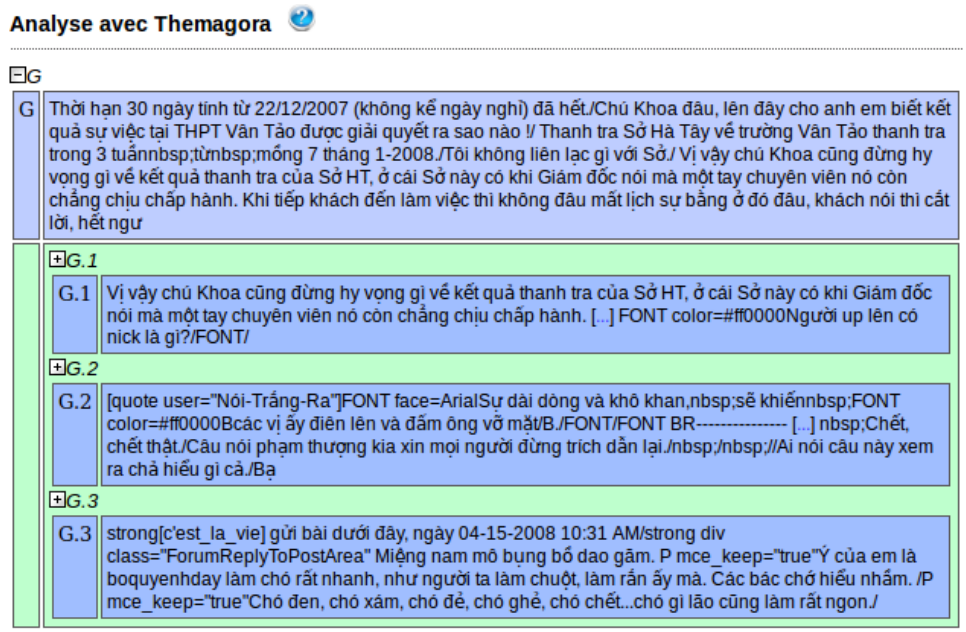
Au travers des logiciels Themagora, et Anagora (Giguet et Lucas 2009) conçus dans le cadre de l'ERTé Calico, nous avons cherché à concevoir une méthode de résolution qui prenne en compte les ordre de grandeur de document, c'est-à-dire leur taille. Nous établissons un parallèle avec les principes de construction de cartes, en géographie (Hubert, 2003). Les principes de construction des cartes géographiques nous semble particulièrement informatifs pour notre étude. Le format des cartes est constant quelle que soit la taille du phénomène représenté : les cartes de France, de départements, de randonnées, ont sensiblement la même taille une fois dépliées. Dans le cas des forums de discussion, nous souhaitons que les représentations générées tiennent sur un écran d'ordinateur quelle que soit la taille du forum. De nombreux concepts sont alors transférables d'un domaine à l'autre. Des cartes de départements de surface légèrement différente sont comparables car toujours représentées à même échelle. Certaines informations qui ne devraient pas être visibles à une échelle donnée sont parfois représentées (une rivière ou une route, à l'échelle de la France). Certaines informations sont au contraire absentes à une certaine échelle car non pertinentes.

Exemple de forum en grec analysé par Themagora



Dans cet exemple, la macro-structuration du discours porte un forum en grec.

Exemple de forum en vietnamien analysé par Themagora



Dans cet exemple, la macro-structuration du discours porte un forum en vietnamien.

5.6 Prospective

5.6.1 Vers une analyse sémiotique des forums de discussions

L'approche distributionnelle que nous utilisons pour effectuer une macro-structuration du discours nous a permis de relâcher les contraintes de forme. Les marques actuellement recherchées pour effectuer la structuration sont les ponctuations fortes, les smileys à base de caractères, ainsi que le nombre d'occurrences de constituants textuels. La mise en forme matérielle (la graisse, l'italique, la couleur, l'alignement, la couleur de fond) n'est pas encore exploitée. La présence d'images dans le corps de message pourrait également enrichir la construction du faisceau de marques : présence de smileys au format image, présence d'autres images. La présence d'un avatar associé à l'auteur, d'une signature de messages personnalisée sont autant de marques qu'il conviendrait d'exploiter, dans la lignée des travaux de Valette et Rastier (2008). Il y a ici tout un effort de recherche à effectuer pour apprendre à utiliser ces marques, à les relativiser. L'analyse purement textuelle, ou « littérale », confrontée au forum de discussion, atteint ses limites et c'est il me semble vers une analyse davantage sémiotique qu'il faut aujourd'hui se tourner. Il y a ici matière à une thèse.

5.6.2 Réflexions sur la pertinence des représentations synthétiques textuelles

Dans ces travaux portant sur l'analyse des forums de discussion, un de nos objectifs était de parvenir à construire des représentations globales et synthétiques, qui restent lisibles sur un écran d'ordinateur, quelle que soit la taille du forum. Il s'agissait d'approfondir la question de la prise en compte des ordres de grandeur de document et la question sous-jacente de la sélection de la fenêtre d'observation.

Si cet objectif nous semble aujourd'hui atteint, l'appropriation par l'utilisateur des représentations synthétiques que nous construisons pose question. Nous constatons en effet que les vues synthétiques *graphiques* produites par l'outil Anagora (Giguet et Lucas, 2009), qui restent également lisibles sur un écran d'ordinateur quelle que soit la taille du forum, sont plus appréciées que les vues synthétiques *textuelles* produites par l'outil Themagora.

Nous faisons l'hypothèse que la difficulté d'appropriation de l'outil Themagora ne tient pas tant à la capacité de l'outil à produire des vues interprétables, que dans la rapidité de l'utilisateur à identifier si la vue fait sens ou pas. Sur une représentation synthétique schématique, nous pensons que l'interprétant identifie très rapidement si la vue fait sens et que si ce n'est pas le cas, il choisit une autre représentation. Sur une représentation synthétique textuelle, l'interprétant se croit obligé de s'arrêter pour lire. Si la vue ne fait pas sens, il considèrera avoir perdu son temps. La validation de cette hypothèse pourrait faire l'objet d'une collaboration interdisciplinaire.

5.7 Conclusion

Les forums de discussion et autres nouvelles formes de communication écrite telles que les listes de discussion qui se développent avec les moyens électroniques posent plusieurs problèmes de manipulation, de visualisation et d'analyse interprétative. Au travers de l'ERTé Calico, notre objectif a été de parvenir à construire des représentations globales et synthétiques, qui restent lisibles sur un écran d'ordinateur, quelle que soit la taille du forum. Parallèlement, nous souhaitions ouvrir des perspectives en terme d'analyse de contenu. Une structuration automatique du discours nous semblait envisageable malgré la particularité du style très libre.

Nous pensons avons fait un réel progrès dans la prise en compte des ordres de grandeur de document en la liant la question à celle de la sélection automatique de la fenêtre d'observation. La

méthode d'analyse de discours guidée par le modèle que nous avons proposée, alliée à la méthode distributionnelle permet de garantir une macro-structuration motivée, quelle que soit la taille du forum de discussion. La méthode distributionnelle a en outre permis de relativiser l'importance des formes dans le traitement informatique des forums, verrou qui nous semblait important à lever pour parvenir à une véritable analyse de contenu. L'analyse distributionnelle, à bonne résolution, sur des formes non lexicales, lève en partie ce verrou. Des travaux supplémentaires sont nécessaires sur ce plan.

La plateforme Calico pour le partage et l'analyse de forum de discussions qui intègrent entre autres les outils Themagora et Anagora témoigne de la maturité des concepts. Cette plateforme est présentée dans le chapitre suivant, en section 6.2.

Dans la prospective, nous avons cependant noté que les marques actuellement utilisées n'étaient pas suffisantes pour tirer pleinement partie des spécificités du document et nous plaidons pour une analyse sémiotique des forums de discussion.

5.8 Publications liées

BLONDEL, François-Marie et GIGUET, Emmanuel. 2009. CALICO, une plate-forme pour visualiser et analyser des discussions. Animation d'un atelier dans le cadre du Colloque international EPAL "Echanger pour apprendre en ligne", Grenoble, France, 4 juin.

GIGUET, Emmanuel, LUCAS, Nadine, BLONDEL, François-Marie, BRUILLARD, Éric. 2009. Share and explore discussion forum objects on the Calico website. Dans "*8th International Conference on Computer Supported Collaborative Learning (CSCL2009)*", June 8-13, 2009, University of Aegean, Rhodes, Greece. Best Technology Design Nomination

GIGUET, Emmanuel, LUCAS, Nadine, BLONDEL, François-Marie, BRUILLARD, Éric. 2009. The Calico Platform: Multilingual Monitoring of Online Discussions. Dans *8th International Conference on Computer Supported Collaborative Learning (CSCL2009) Workshop: "Interaction Analysis and Visualization for Asynchronous Communication: Analysis Methods, Tools, and Research Questions*", June 8-13, 2009, University of Aegean, Rhodes, Greece.

GIGUET, Emmanuel, et LUCAS, Nadine. 2009. Creating Discussion threads graphs with Anagora. Dans "*8th International Conference on Computer Supported Collaborative Learning (CSCL2009)*", June 8-13, 2009, University of Aegean, Rhodes, Greece.

LUCAS, Nadine, et GIGUET, Emmanuel. 2005. UniTHEM, un exemple de traitement linguistique à couverture multilingue. *Conférence Internationale sur le Document Electronique (CIDE 8*). Beyrouth, Liban, pp. 115-132

LUCAS, Nadine, SIDIR, Mohamed, et GIGUET, Emmanuel. 2006. Analyse de forums dans la formation à distance. *Conférence Internationale sur le Document Electronique (CIDE 9*). Fribourg, Suisse. Europia. pp. 169-180.

LUCAS, Nadine, et GIGUET, Emmanuel. 2008. Robust adaptive discourse parsing for e-learning fora. *The 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008)*, Santander, Cantabria, Spain July 1st- July 5th, 2008, P. Diazet al., IEEE. pp. 730-732.

LUCAS, Nadine, et GIGUET, Emmanuel. 2010. L'analyse de forums par ThemAgora. JOCAIR. Amiens.

SIDIR, Mohamed, LUCAS, Nadine, et GIGUET, Emmanuel. 2007. De l'analyse des discours à l'analyse structurale des réseaux sociaux : une étude diachronique d'un forum éducatif. *Revue Sticef*, vol. 13. numéro spécial forum. ISSN : 1764-7223, mis en ligne le 20/03/2007.

6 Les cadres d'expérimentation et de diffusion scientifique

6.1 Le cadre d'expérimentation scientifique

Dans le cadre d'une démarche orientée corpus, l'expérimentation tient une place prépondérante. Sans expérimentation, sans cadre applicatif, il n'y a pas de validation possible des concepts. Il s'agit donc d'un passage incontournable qui nécessite une instrumentation spécifique. La plateforme d'expérimentation est la solution que j'ai retenue.

La plateforme Wims est une plateforme d'expérimentation linguistique que j'ai conçue, réalisée, et que je maintiens depuis mon entrée au CNRS. C'est au travers de cette plateforme que je mets à l'épreuve la plupart des prototypes réalisés par moi-même ou par l'équipe. Il m'est ainsi possible, en toute transparence, de valider ou d'invalider les concepts sous-jacents sur diverses collections. Cette plateforme est principalement réservée à un usage interne mais sert aussi parfois de vitrine dans le cadre de démonstrations organisées.

La plateforme Wims se caractérise par sa simplicité d'utilisation, pour réaliser des traitements sur des collections de documents multilingues, et pour explorer les résultats. Une fonctionnalité complémentaire de la plateforme Wims est d'assurer la disponibilité des prototypes et la reproductibilité des expériences au fil des ans. C'est un point stratégique car bien des prototypes disparaissent faute d'avoir été entretenus ou d'avoir conservé leur documentation.

D'un point de vue épistémologique, la plateforme d'expérimentation Wims reflète une approche de la résolution de problème orthogonale à celle reflétée par des plateformes d'analyse comme Pandore (Clouard *et al.*, 1997), pour l'analyse d'image, et Linguastream (Bilhaut, 2006a), pour l'analyse de texte. Dans ces deux plateformes, l'interface est conçue pour construire de manière interactive une chaîne de traitement par assemblage de composants d'analyse. C'est donc la chaîne de traitement qui motive l'interface. Dans Wims, la chaîne de traitement, également modulaire, n'est pas présentée. L'interactivité porte sur l'exploration du corpus et des résultats d'analyse. C'est donc le corpus et les résultats d'analyse qui motivent l'interface.

6.2 Le cadre de diffusion scientifique

À l'heure où la publication scientifique est un des principaux critères d'évaluation de la recherche scientifique, peu d'équipes de recherche peuvent aujourd'hui se prévaloir d'une plateforme accessible librement par Internet et qui permet la mise à l'épreuve et la transmission des concepts, en toute transparence. C'est ce défi que j'ai souhaité relever en concevant et en assurant la disponibilité de la plateforme Calico.

La plateforme Calico est spécialisée dans l'analyse de forums de discussion en toute langue. Elle est très ouverte et bénéficie d'une visibilité internationale au sein de la communauté de la recherche pédagogique. Elle a notamment permis à des chercheurs de France, de Belgique, de Grèce, des États-Unis, du Canada, d'Israël, du Vietnam, de Corée et de Hong-Kong de mettre à l'épreuve nos concepts et d'expérimenter nos technologies sur des forums de discussion de différente taille et écrits dans leur propre langue, à savoir en français, en anglais, en grec, en hébreu et en vietnamien.

Soumettre ses concepts à l'épreuve du réel est particulièrement délicat. La principale difficulté d'une mise à disposition très ouverte réside en fait dans la capacité à maîtriser l'intégralité de la chaîne de traitement. Il faut garantir la prise en charge de textes ou de documents « bruts» ayant des formats variés. Il faut ensuite réaliser leur analyse en un temps raisonnable, et produire une représentation toujours lisible sur écran, quelles que soient la taille et la langue des données. Il n'est plus seulement question d'ingénierie mais d'une mise en œuvre des concepts de gestion d'échelle, de variation en langue et de prise en compte d'un genre.

Cette initiative de diffusion scientifique a permis à la plateforme Calico d'être sélectionnée dans la catégorie « Best Technology Design » à la conférence internationale reconnue CSCL. Cette distinction accroit la visibilité des concepts sous-jacents qui ont tous fait l'objet de publications internationales.

6.3 Publications liées

BLONDEL, François-Marie, GIGUET, Emmanuel. 2009. CALICO, une plate-forme pour visualiser et analyser des discussions. Animation d'un atelier dans le cadre du Colloque international EPAL "Echanger pour apprendre en ligne", Grenoble, France, 4 juin.

GIGUET, Emmanuel et LUCAS, Nadine. 2002. Intégration d'Unicode - conception d'un agent de recherche d'information sur Internet. Document numérique, n° spécial "Unicode, écriture du monde?" sous la direction de A. Jacques et H. Hudrisier. Vol. 6 n° 3-4. pp. 225-236.

GIGUET, Emmanuel. 2005. Modélisation de l'activité expérimentale du chercheur en traitement des langues sur corpus multilingues. Journée "Articuler les traitements sur corpus", du 12 février, organisée par Benoît Habert, Serge Heiden et André Salem.

GIGUET, Emmanuel, LUCAS, Nadine, BLONDEL François-Marie, et BRUILLARD, Éric. 2009. The Calico Platform: Multilingual Monitoring of Online Discussions. Dans 8th International Conference on Computer Supported Collaborative Learning (CSCL2009) Workshop: "Interaction Analysis and Visualization for Asynchronous Communication: Analysis Methods, Tools, and Research Questions", 8-13 juin 2009, University of Aegean, Rhodes, Greece.

GIGUET, Emmanuel, LUCAS, Nadine, BLONDEL François-Marie, et BRUILLARD, Éric. 2009. Share and explore discussion forum objects on the Calico website. Dans "8th International Conference on Computer Supported Collaborative Learning (CSCL2009)", June 8-13, 2009, University of Aegean, Rhodes, Greece. Best Technology Design Nomination

7 Regard épistémologique sur mes recherches

7.1 Le modèle linguistique au centre de la gestion des ordres de grandeur

Pour analyser un texte ou un document, il faut un modèle. Un modèle linguistique lorsqu'il s'agit d'un texte, un modèle sémiotique lorsqu'il s'agit d'un document composé de texte, d'images, de schémas... Le modèle guide l'analyse des données et permet leur interprétation. C'est dans la confrontation du modèle aux données que s'effectue la construction du sens. Pour l'interprétant, « *ça fait sens* » ou « *ça ne fait pas sens* ». Dans cette perspective, le modèle ne peut être que central.

Dans mon activité expérimentale, le modèle est choisi par le linguiste, en fonction du contexte applicatif. Un modèle de structuration thématique a par exemple été sélectionné et mis en œuvre sur des articles journalistiques, scientifiques, et sur des forums de discussions en toute langue, pour produire une forme condensée et structurée du contenu. Un modèle inspiré de Jakobson a, quant à lui, été utilisé en analyse de la presse internationale pour mettre en évidence les différents points de vue sur un événement.

Le modèle est aujourd'hui central dans mes travaux. Il ne l'a pas toujours été. J'ai bien longtemps soutenu une approche guidée par les données, avec pour principale motivation la robustesse des traitements, c'est-à-dire une moins grande sensibilité à la diversité des phénomènes de surface constatés. Je travaillais alors en analyse syntaxique automatique et les approches à base de grammaire formelle avaient à mon sens montré leur incapacité à analyser des productions réelles. Une analyse guidée par le modèle n'est cependant pas moins robuste : elle est plus contraignante. Elle impose la mise en place d'un contrôle global du schéma relationnel en cours d'instanciation.

Si le modèle est devenu central, c'est grâce à l'introduction de contextes applicatifs où la nécessité de comprendre domine : est-ce que l'analyse « *fait sens* » ou « *ne fait pas sens* » pour l'interprétant, pour l'utilisateur du système ? Faire de l'analyse syntaxique automatique sans contexte applicatif ne m'a en fait questionné que très tardivement. Pourtant aucun utilisateur ne s'intéresse en fait à des milliers de phrases analysées automatiquement une à une. Et ce n'est pas en ajoutant le calcul de relations anaphoriques sortant du cadre de la phrase que le point de vue de l'utilisateur change. Ce travail souffrait en fait de l'absence d'un fil directeur, tel un édifice construit sans plan architectural, sans information sur sa destination.

En traitement des langues, les travaux portant sur le modèle propositionnel restent majoritaires. Au dessus de la proposition et de la phrase, lorsqu'il s'agit de structurer le texte ou le document, on parle d'analyse du discours. Dans la communauté francophone du traitement des langues, les principaux modèles côtoyés en analyse automatique du discours sont la SDRT (Lascarides et Asher, 1993), le modèle d'encadrement du discours (Charolles, 1997) et le Centrage de Grosz, Joshi et Weinstein (Pery-Woodley, 2000). Ces modèles n'ont pas de relation avec le modèle d'architecture textuelle (Pascual et Virbel, 1996). La SDRT a pour vocation de modéliser l'organisation du discours, en s'intéressant aux liens entre propositions. Centré sur la proposition, ce modèle est utilisé pour la modélisation de petits ordres de grandeur, pas vraiment pour l'analyse de textes entiers. Le modèle d'encadrement du discours et le modèle d'architecture textuelle permettent la structuration d'ordres de grandeur un peu plus variés, le modèle d'architecture textuelle ayant

l'ambition de représenter la structure de textes entiers, alors que l'encadrement du discours propose une modélisation de quelques paragraphes. Cependant, ces deux modèles sont davantage descriptifs que fonctionnels et nourrissent peu de logiciels, notamment (Bilhaut, 2006) pour l'encadrement du discours.

Si le modèle propositionnel n'est pas vu comme un modèle d'analyse de discours malgré les tentatives de Kintsch et Van Dijk (1978), rien ne prédispose pourtant ce modèle à n'être confronté qu'à la phrase. Il se trouve que ce modèle est systématiquement associé à cet unique ordre de grandeur. Il faut se tourner vers des modèles d'interprétation comme celui de Van Dijk (1985, 1988) pour constater que les choses ne sont pas si figées. C'est dans cette perspective de confrontation de modèles linguistiques à des données, dans leur ensemble, quelle que soit leur taille, que je me suis intéressé à la mise au point des techniques informatiques appropriées, et c'est au travers du modèle thème-rhème proposé par Nadine Lucas que mes travaux ont pris une nouvelle dimension, avec des applications sur des articles de presse, des articles scientifiques, et des forums de discussion.

7.2 Le statut incertain de constituants non linguistiques

S'il me semble aujourd'hui incontournable de placer le modèle linguistique au cœur du contrôle, il ne m'a pas été aisé de comprendre sa place dans la construction d'une application de traitement des langues. Il s'agit en effet d'un problème difficile. Il faut en effet garder à l'esprit que l'évaluation de plusieurs analyseurs syntaxiques guidés par le modèle et réalisée par Anne Abeillé (1991) témoignait de l'incapacité de ces approches à analyser de véritables écrits. Il faut également avoir à l'esprit l'engouement suscité par les méthodes d'analyse robuste, dirigées par les données, notamment le « *part-of-speech tagging* », le « *phrase chunking* » ou le « *clause bracketing* ». C'est autour des concepts de modèle opératoire et d'échelle de constituants que se sont cristallisées mes recherches sur des principes fondamentaux de résolution.

Un modèle linguistique ne peut être projeté directement sur un texte par une technique informatique. Il n'existe par exemple pas de programmes qui découpent parfaitement et systématiquement un texte en mots, en phrases, en paragraphes, ou qui le structurent en syntagmes, en propositions, en cadres de discours, au sens linguistique. Certes, il existe des implémentations produisant des approximations globalement satisfaisantes pour certains objets linguistiques. Pour le cas de la segmentation en mots, c'est par le recours à des listes de mots et des listes de locutions *à ne pas manquer* que la communauté parvient à un compromis acceptable entre linguistes et informaticiens. Mais on finit cependant toujours par relever un mot *mal découpé*. Il en est de même pour tous les autres objets : « ça ne tombe jamais systématiquement juste ». Il y a toujours une phrase *mal découpée*, un paragraphe *mal délimité*, un syntagme *incomplet*, une proposition *manquée,* un cadre de discours *mal borné*. S'il est d'usage de se satisfaire de tels résultats, l'image du traitement des langues ne cesse d'être attaquée pour ces « approximations ».

Pour la défense des concepteurs de technologies, on pourrait arguer qu'il s'agit d'ergotage, que la majorité des objets ne sont pas si mal calculés, évaluation à l'appui. On pourrait aussi arguer qu'il ne s'agit que d'une question de vocabulaire mal choisi, que le mot ou la phrase du grammairien ne sont pas les mêmes que le mot ou la phrase de l'informaticien et que l'usage de termes identiques pour deux objets différents nuit à la reconnaissance de la qualité des analyses automatiques proposées. Il n'en reste pas moins que l'on ne peut s'empêcher d'observer le mot calculé sous l'éclairage du mot linguistique, de chercher au travers du mot calculé un mot immédiatement interprétable, un mot qui fasse sens. Et cette comparaison est effectuée quel que soit l'objet calculé. Lorsqu'apparut par exemple dans les années 1990 le terme « *chunk* », introduit et promu par les partisans d'une approche robuste du traitement des langues, on ne put s'empêcher de confronter ce nouvel objet avec des objets linguistiques classiques, interprétables : on le rebaptisa alors, à la guise des

circonstances, syntagme minimal, syntagme non récursif ou segment prosodique minimal. Lors d'initiatives d'évaluation, on chercha à le redéfinir de sorte qu'il s'apparente mieux à un objet déjà connu, donc interprétable.

Le mot calculé comme le chunk ne sont manifestement pas des objets linguistiques. Ils n'en sont cependant pas si éloignés puisque c'est en cherchant à les calculer qu'ils ont été produits, ou imaginés. Lorsque l'on souhaite les redéfinir de sorte qu'ils coïncident davantage avec des objets linguistiques, ils échappent aux traitements informatiques et ne sont alors plus calculables. Ils n'ont pas de statut mais l'on s'accorde à dire qu'ils sont tout de même bien pratiques. Ce sont des objets que l'on ne peut véritablement interpréter mais ce sont pourtant les seuls que l'on sache de fait calculer et manipuler. Ils font partie de la « cuisine » du traitement des langues, d'un certain empirisme du traitement des langues.

Ne pouvant me satisfaire d'un traitement des langues où les méthodes de résolution ne seraient que d'*inspiration* linguistique, j'ai souhaité mener une réflexion sur ces objets qui ne sont manifestement pas le fruit du hasard, et cela, afin de leur conférer un véritable statut. C'est dans la définition d'un modèle opératoire, basé sur une échelle de constituants opératoires que nous proposons une modélisation plus élégante, et une méthode de résolution servant de cadre dans divers contextes applicatifs.

7.3 La délimitation et l'interprétation des constituants opératoires

Contrairement à ce que laisse penser la variété d'objets énumérés ci-dessus – mot, phrase, proposition, cadre de discours – le problème de la délimitation des unités et de leur interprétation ne se cantonne pas à la seule structuration textuelle. Il s'agit d'une problématique générale de reconnaissance de formes. En traitement des langues, on la rencontre ainsi dans des contextes applicatifs aussi variés que l'extraction d'information, l'alignement automatique, ou la structuration automatique de documents.

En extraction d'information, la question se manifeste sous la forme de la pertinence du mot comme unité d'indexation : la suite de caractères constituerait-elle une meilleure alternative ? Elle laisse en effet espérer un traitement uniforme de toutes les familles de langue ainsi que la mise en valeur de facteurs communs non accessibles au traitement par mot (McNamee et al., 2008). En alignement automatique, c'est la différence de statut du mot dans les langues à aligner qui conduit à sa remise en cause en faveur d'un traitement au caractère (Lecluze, 2011). Les motivations en alignement automatique sont semblables à celles en indexation automatique : l'absence de délimiteur de mot, ainsi que les phénomènes d'agglutination et de flexion font du mot calculé une unité qui n'est pas d'emblée compatible avec l'objectif d'extraire des unités qui font sens. En structuration automatique de documents, la problématique est rencontrée à différentes étapes du traitement. Par exemple, lors de la détection des lignes de corps de page, il faudra décider du statut de la suite de caractères située physiquement dans la marge, et qui peut correspondre à une note de marge ou bien à la fin d'une ligne de corps de texte ayant fortuitement débordé dans la marge suite à un problème de mise en page.

Ces questions, dans leur apparente diversité, soulèvent toutes l'unique question du statut des objets calculés et de leur interprétation dans le cadre d'un modèle linguistique ou sémiotique. Bien évidemment, cette question du statut des objets calculés est incontournable puisque l'objectif d'un traitement des langues est précisément de construire des objets qui font sens. Mais il convient de déterminer à quel stade de la résolution il est raisonnable de confronter l'objet calculé à l'objet attendu, à l'objet qui fait sens. Je fais l'hypothèse que l'analyse est souvent considérée achevée trop tôt et que la confrontation entre l'objet calculé et l'objet qui fait sens intervient par conséquent sur des critères partiels. La question de l'interprétation se posant sur des objets aussi variés que le mot

calculé, le chunk calculé, ou le cadre de discours calculé (Bilhaut, 2006a), je suppose que le problème n'est pas tant lié à telle ou telle méthode de structuration qu'au paradigme de résolution sous-jacent.

Nous sommes en effet dans le cadre d'une démarche de structuration sur corpus et à chaque unité calculée peut correspondre un processus d'analyse robuste : pour le mot, la segmentation en mots (ou *tokenization, word segmentation*), pour le chunk, la segmentation en chunks ou en constituants minimaux (ou *chunking)*, pour les propositions, la délimitation en proto-propositions (ou *clause-bracketing*), pour les segments thématiques, la délimitation des segments thématiques (ou *text tiling*). C'est dans une perspective de structuration ascendante que sont conçus ces processus et c'est en ce sens que malgré la diversité des constituants ciblés, je les considère comme appartenant tous au même paradigme.

Pour l'ensemble de ces processus, la résolution consiste en une structuration ascendante conduisant à la délimitation de constituants textuels sur critère formel. A mon sens, les propriétés des constituants textuels engendrés dans un tel paradigme sont encore trop éloignées des propriétés des constituants *fonctionnels* linguistiques. Ces constituants textuels doivent être considérés comme des objets intermédiaires de calcul, comme des constituants opératoires. Tendre vers une meilleure adéquation entre ces constituants opératoires et des constituants linguistiques est, je pense, illusoire à l'intérieur de ce paradigme, et ne peut engendrer que des améliorations *marginales*. C'est sous ces hypothèses et dans la recherche d'un paradigme de résolution plus puissant que je mène mes recherches.

7.4 Les constituants opératoires au centre du modèle opératoire

Si la linguistique a su définir des constituants fonctionnels pour de nombreux modèles, le traitement des langues montre que l'identification de tels constituants par un processus informatique n'est pas immédiate : « cela ne tombe jamais juste ». J'attribue l'inadéquation des deux représentations au fait que deux modèles se confrontent : le modèle linguistique, d'une part, et un modèle que je baptise modèle opératoire, d'autre part. Les deux modèles ne peuvent que cohabiter et c'est dans la cohabitation d'un modèle linguistique et d'un modèle opératoire que je définis une application de traitement des langues.

L'introduction du concept de modèle opératoire engendre la dissociation systématique des constituants relevant du modèle linguistique, de ceux relevant du modèle opératoire. Il contribue à délimiter le champ de l'activité de modélisation en traitement des langues, à éviter notamment la confusion fréquente qui consiste à assimiler des constituants opératoires comme le mot calculé ou le chunk, relevant de la modélisation opératoire, à des constituants linguistiques comme le syntagme ou la proposition, relevant de modélisation linguistique. Le modèle opératoire contribue à la définition d'une hiérarchie de constituants opératoires permettant le calcul des constituants du modèle linguistique par l'intermédiaire d'un processus de structuration.

Dans cette perspective, le constituant opératoire correspond à la meilleure approximation d'un constituant à valeur linguistique, calculable par un processus de structuration, souvent ascendante en pratique. Tout du moins le concepteur a-t-il cette légitime intention lors de la construction du modèle opératoire. Ainsi le mot calculé se trouve-t-il être la meilleure approximation du mot linguistique, c'est-à-dire du mot qui fait sens, le chunk la meilleure approximation du syntagme minimal, la proto-proposition la meilleure approximation de la proposition, et le cadre de discours calculé la meilleure approximation du cadre de discours (Charolles, 1997), et ainsi de suite.

Sous cet angle, le segment thématique de la méthode Texttiling (Hearst, 1997) revêt un caractère particulier puisqu'à ma connaissance, il n'aurait de correspondant direct dans aucun modèle

linguistique d'analyse de discours. Se pose alors la question du statut de cet objet et, par extension, de cette méthode de structuration ascendante vis à vis du traitement des langues. La conception du modèle opératoire précède semble-t-il celle du modèle linguistique. La pertinence des segments calculés me semble alors difficile à établir puisqu'aucun parallèle avec un objet interprétable, un objet qui fait sens, de la structure du discours, n'est encore établi.

Le constituant opératoire est intimement lié à la stratégie de résolution, il en fait pleinement partie. Et s'il est légitime que le constituant opératoire soit la meilleure approximation d'un constituant linguistique, des approximations diverses sont envisageables en fonction de la stratégie de résolution retenue. Ainsi, d'un analyseur à l'autre, la définition des constituants opératoires diffère, ce qui entraîne par exemple la variabilité du mot calculé, du chunk, et plus généralement de tout constituant opératoire, comme on l'a vu dans les évaluations d'analyseurs du français comme GRACE ou EASY. Bien entendu, s'il est pertinent de s'interroger sur la meilleure approximation, il n'en reste pas moins que la comparaison de deux segmentations en constituants opératoires n'a de sens que dans la perspective d'une même application, et dans l'analyse de leur propension à permettre le calcul de la représentation finale.

Considérons plusieurs illustrations tirées de notre expérience de l'analyse syntaxique robuste du français. Alors que la dissociation des pronoms clitiques inversés est d'usage lors de la segmentation en mots, par exemple *vais-je* dissocié en *vais* et *-je,* il est tout à fait envisageable de ne pas procéder à cette dissociation et de traiter localement ces inversions comme des suffixes. Cela n'a pas d'impact sur la qualité finale de l'analyse. De même, si l'on s'intéresse au calcul de la relation sujet-verbe, on constate que la délimitation des chunks prépositionnels n'est pas d'un intérêt fondamental puisque les prépositions ne jouent pas un rôle principal dans le calcul de cette relation. Enfin, si la perspective est une structuration en propositions, la délimitation des proto-propositions peut s'appuyer sur les introducteurs de propositions, comme les pronoms relatifs et les conjonctions, mais peut également reposer sur la virgule qui, comme le point, permet de caler le début ou la fin d'un segment prosodique avec le début ou la fin d'un segment propositionnel.

7.5 La sélection des marques pour la structuration en constituants opératoires

Le modèle opératoire doit refléter la hiérarchisation des marques captables par un processus de structuration automatique. Ces marques constituent les points d'accroche du processus sur le matériau linguistique. Elles sont bien entendu dépendantes de la tâche de structuration et intimement liées à la stratégie de calcul de la structure. Si le choix des marques appartient au linguiste ou au sémioticien, leur sélection doit être guidée avec, à l'esprit, les contraintes d'une résolution informatique et non strictement linguistique, ce qui à mon sens constitue la principale motivation du travail interdisciplinaire.

7.5.1 De la dissymétrie du marquage des frontières gauche et droite

Les constituants à valeur linguistique ont la particularité de n'être que rarement marqués à la fois en début et en fin. C'est une des caractéristiques qui différencie les langues des langages de programmation pour lesquels les blocs fonctionnels sont systématiquement « parenthésés » afin de faciliter leur identification. Ainsi, en français, le début des propositions subordonnées est systématiquement marqué alors que la fin ne l'est pas. Notons qu'il n'y a pas de marque en tête des propositions principales : comme nous l'avons vu précédemment, ce n'est pas tant l'introducteur de la proposition qui est constitutif de la proposition que la relation sujet-verbe.

La dissymétrie du marquage des frontières gauche et droite rend impossible l'identification des

constituants à valeur linguistique à partir de critères uniquement morphologiques. Une méthode de résolution appropriée ne peut donc se contenter d'être descriptive des structures pour les caractériser. Elle peut certes exploiter les marques de début ou de fin, utiliser des critères de cohérence interne exprimée sous forme de relation, mais cela ne suffit pas. Elle doit également modéliser l'environnement dans lequel ces constituants apparaissent, c'est-à-dire leur contexte. La modélisation du contexte passe par la sélection de marques stables, caractéristiques des constituants connexes et des constituants d'ordre supérieur.

> Exemple de segmentation en chunks réalisée par WimsParser :
>
> [**A** mesure] qu' [**elle** chantait] , [**l'**ombre] [descendait] [**des** grands arbres] , et [[**le** clair] [**de** lune naissant]] [tombait] [**sur** elle seule] , [isolée] [**de** notre cercle attentif] . - [**Elle** se tut] , et [personne] [**n'**osa] [rompre] [**le** silence] . [**La** pelouse] [**était** couverte] [**de** faibles vapeurs condensées] , qui [déroulaient] [**leurs** blancs flocons] [**sur** les pointes] [**des** herbes] . [**Nous** pensions être] [**en** paradis] . - [**Je** me levai enfin] , [courant] [**au** parterre] [**du** château] , où [**se** trouvaient] [**des** lauriers] , [plantés] [**dans** de grands vases] [**de** faïence peints] [**en** camaïeu] . [**Je** rapportai] [**deux** branches] , qui [furent tressées] [**en** couronne] et [nouées] [**d'un** ruban] . [**Je** posai] [**sur** la tête] [**d'**Adrienne] [**cet** ornement] , dont [**les** feuilles lustrées] [éclataient] [**sur** ses cheveux blonds] [**aux** rayons pâles] [**de** la lune] . [**Elle** ressemblait] [**à** la Béatrice] [**de** Dante] qui [sourit] [**au** poète errant] [**sur** la lisière] [**des** saintes demeures] .
>
> *La segmentation exploite le marquage régulier des têtes (en gras) mais cela ne suffit pas. Les marques caractéristiques du contexte permettent d'identifier le début et la fin des chunks non marqués : la marque introductrice d'un chunk connexe permet de clore le précédent, les marques caractéristiques de constituants d'ordre supérieur (en rouge) permettent d'ouvrir ou de clore les chunks non marqués en début ou en fin.*

> Exemple de segmentation en proto-propositions réalisée par WimsParser.
>
> **[** A mesure **[** qu'elle chantait **] ,** l'ombre descendait des grands arbres, **] [** et le clair de lune naissant tombait sur elle seule, isolée de notre cercle attentif **] . [** - Elle se tut, **] [** et personne n'osa rompre le silence **] . [** La pelouse était couverte de faibles vapeurs condensées, **] [** qui déroulaient leurs blancs flocons sur les pointes des herbes **] . [** Nous pensions être en paradis **] . [** - Je me levai enfin, **] [** courant au parterre du château, **] [** où se trouvaient des lauriers, plantés dans de grands vases de faïence peints en camaïeu **] . [** Je rapportai deux branches, **] [** qui furent tressées en couronne et nouées d'un ruban **] . [** Je posai sur la tête d'Adrienne cet ornement, **] [** dont les feuilles lustrées éclataient sur ses cheveux blonds aux rayons pâles de la lune **] . [** Elle ressemblait à la Béatrice de Dante **] [** qui sourit au poète errant sur la lisière des saintes demeures **]** .
>
> *La segmentation exploite le marquage régulier des têtes (en rouge et en rose). Les marques caractéristiques du contexte permettent d'identifier le début et la fin des proto-propositions non marquées : la marque introductrice d'une proto-proposition connexe permet de clore la précédente, les marques caractéristiques de constituants d'ordre supérieur (la ponctuation en vert) permettent d'ouvrir ou de clore les proto-propositions non marquées en début ou en fin.*

La modélisation de la structure des constituants à valeur linguistique ne suffit pas pour réaliser leur repérage automatique dans les textes. La modélisation du contexte d'apparition doit également être effectuée au moyen de marques stables, caractéristiques des constituants connexes et des constituants d'ordre supérieur.

7.5.2 De la fiabilité des marques dans la structuration en constituants opératoires

Si les constituants à valeur linguistique sont régulièrement marqués en début ou en fin, il convient de tenir compte du fait que toutes les marques ne se valent pas en terme de fiabilité. Ainsi, dans une perspective d'analyse propositionnelle, on constate que des marques introductrices comme *à laquelle*, *quand* ou *pourquoi* sont plus fiables que des marques telles que *que*, *dont* ou *comme* qui peuvent certes débuter une proposition mais peuvent également entrer dans d'autres constructions intra-propositionnelles comme la négation ou la comparaison. Ces marques non discriminantes sont à considérer comme secondaires ou comme indices, pas comme des marques directement exploitables. De même, dans une perspective d'analyse de la relation sujet-verbe, il est préférable de considérer la marque *de* comme secondaire, puisqu'elle peut en tant que partitif entrer dans l'identification du sujet, mais a cependant beaucoup plus souvent le statut de préposition. Elle n'est donc pas discriminante dans ce cas.

> Exemple de segmentation en proto-propositions, réalisée par WimsParser, tenant compte des marques discriminantes (en rouge) et des marques secondaires (en rose) :
>
> **[** A mesure **[** qu'elle chantait **]** , l'ombre descendait des grands arbres, **]** **[** et le clair de lune naissant tombait sur elle seule, isolée de notre cercle attentif **]** . **[** - Elle se tut, **]** **[** et personne n'osa rompre le silence **]** . **[** La pelouse était couverte de faibles vapeurs condensées, **]** **[** qui déroulaient leurs blancs flocons sur les pointes des herbes **]** . **[** Nous pensions être en paradis **]** . **[** - Je me levai enfin, **]** **[** courant au parterre du château, **]** **[** où se trouvaient des lauriers, plantés dans de grands vases de faïence peints en camaïeu **]** . **[** Je rapportai deux branches, **]** **[** qui furent tressées en couronne et nouées d'un ruban **]** . **[** Je posai sur la tête d'Adrienne cet ornement, **]** **[** dont les feuilles lustrées éclataient sur ses cheveux blonds aux rayons pâles de la lune **]** . **[** Elle ressemblait à la Béatrice de Dante **]** **[** qui sourit au poète errant sur la lisière des saintes demeures **]** .
>
> *Alors que les marques discriminantes engendrent l'ouverture systématique d'une nouvelle proto-proposition, c'est l'apparition d'un second verbe, violant la cohérence de la structure interne du modèle de proposition mono-verbale, qui déclenche la recherche en amont d'une marque secondaire de début de proposition.*

> Exemple de sous-segmentation des proto-propositions par insuffisance de marques discriminantes, réalisée par WimsParser :
>
> **[** sortant du théâtre avec l'amère tristesse que laisse un songe évanoui **]**
>
> *La marque secondaire* que *(en rose) n'est pas utilisée pour délimiter une nouvelle proto-proposition car l'absence de marque sur le verbe n'a pas permis d'établir la relation caractéristique sujet-verbe. Une modélisation plus fine des relations internes à la proposition permettrait de résoudre ce cas.*

La prise en compte de la fiabilité des marques, à différents niveaux de structuration, entraîne la prise de décisions différentiée, correspondant à différents moments de la résolution. La prise de décision doit en effet intervenir une fois construit le faisceau de marques fiable, seul garant de la fiabilité de la décision. Alors qu'une méthode ne distinguant pas la fiabilité des marques est amené à prendre des décisions trop tôt, quitte à les remettre en cause plus tard, ou bien à retarder inutilement des décisions utiles, la prise en compte de la fiabilité des marques dans le processus de structuration permet une résolution plus efficace.

7.5.3 De la dissociation des marques frontières et des marques constitutives

L'existence de marques discriminantes de frontières de constituants à valeur linguistique et

l'existence de marques constitutives discriminantes de ces mêmes constituants laissent envisager la possibilité d'une utilisation découplée des marques frontières et des marques constitutives. L'identification de marques frontières de constituants à valeur linguistique ne nécessite la capacité à identifier les marques constitutives que lorsqu'elles sont secondaires. Et l'identification de marques constitutives de constituant à valeur linguistique ne nécessite la délimitation de constituant que lorsqu'elles sont secondaires. La relative inter-dépendance des marques frontières et des marques constitutives des constituants à valeur linguistique permet de dissocier partiellement le calcul des frontières des constituants du calcul de leur structure interne.

Sur un plan opératoire, la dissociation de la délimitation des constituants à valeur linguistique de l'identification de leur structure interne conduit à une résolution en deux temps : dans un premier temps une segmentation en constituants *opératoires*, ayant une structure interne non motivée linguistiquement, et dans un second temps une structuration interne guidée par le modèle, et donc interprétable. Dans cette résolution en deux temps, le constituant opératoire constitue un objet intermédiaire du calcul qui n'est pas immédiatement interprétable. C'est un espace qui permet de caler le modèle d'interprétation, à partir des marques constitutives qu'il contient, mais également à partir de deux nouvelles positions exploitables par la résolution guidée par le modèle : le début et la fin du constituant opératoire.

Il n'est cependant pas toujours possible de caler un modèle d'interprétation dans l'espace du constituant opératoire. La recherche de marques frontières, dissociée de la recherche de marques constitutives, ne présage pas de l'existence des marques constitutives suffisantes pour caler le modèle d'interprétation dans l'espace délimité. En l'absence de marques constitutives suffisantes au calage du modèle d'interprétation dans l'espace du constituant opératoire, la structuration guidée par le modèle ne peut aboutir, et le constituant à valeur linguistique ne peut être construit ou validé. Pour être fiable, la résolution doit être reportée à un niveau de structuration supérieur, sans garantie de succès, même si d'un point de vue interprétatif le constituant opératoire non validé peut effectivement correspondre à un constituant à valeur linguistique.

> Exemple de détection de la relation sujet-verbe, réalisée par WimsParser, basée sur la dissociation de la délimitation des proto-propositions et de la relation constitutive sujet-verbe :
>
> [[A mesure [qu'elle chantait] , l'ombre descendait des grands arbres,]] [[et le clair de lune naissant tombait sur elle seule, isolée de notre cercle attentif] .] [[- Elle se tut,] [et personne n'osa rompre le silence] .] [[La pelouse était couverte de faibles vapeurs condensées,] [qui déroulaient leurs blancs flocons sur les pointes des herbes] .] [[Nous pensions être en paradis] .] [[- Je me levai enfin,] [courant au parterre du château,] [où se trouvaient des lauriers, plantés dans de grands vases de faïence peints en camaïeu] .] [[Je rapportai deux branches,] [qui furent tressées en couronne et nouées d'un ruban] .] [[Je posai sur la tête d'Adrienne cet ornement,] [dont les feuilles lustrées éclataient sur ses cheveux blonds aux rayons pâles de la lune] .] [[Elle ressemblait à la Béatrice de Dante] [qui sourit au poète errant sur la lisière des saintes demeures] .]
>
> *Le calcul des relations sujet-verbe est guidé par le modèle : une fois les proto-propositions délimitées, les sujets sont identifiés en suivant le modèle canonique SV (en surligné bleu), et à défaut le modèle de sujet inversé VS (en surligné rose).*

> Exemple de détection de la relation sujet-verbe guidée par le modèle, réalisée par WimsParser :
>
> [- La ressemblance d'une figure oubliée depuis des années se dessinait désormais avec une netteté singulière ;]
>
> [La couronne donnée par mes mains à la belle chanteuse était le sujet de ses larmes]

[Les longs anneaux roulés de ses cheveux d'or effleuraient mes joues]

[la vibration de sa voix si douce et cependant fortement timbrée me faisait tressaillir de joie et d'amour] .

[si les Huns, les Turcomans ou les Cosaques n'arrivaient pas enfin pour couper court à ces arguments de rhéteurs et de sophistes]

[Quelques-uns d'entre nous néanmoins prisaient peu ces paradoxes platoniques,]

Alors que la visualisation laisse penser que le système calcule la structure de sujets complexes, il n'en est rien. La dissociation de la délimitation des proto-propositions et de la relation constitutive sujet-verbe permet la localisation des sujets, sans que leur structure ait été préalablement calculée. Le sujet est simplement caractérisé par une marque introductrice dite « de définition », sans vérification d'accord interne, sans vérification d'accord avec le verbe, sans résolution de la coordination ou de l'énumération, sans résolution des rattachements prépositionnels. Suivant le modèle canonique SV, le sujet a une position contrainte, entre la marque de début de proposition et le verbe : c'est cette position qui est surlignée. Les premiers exemples illustrent la couverture de cette approche, au vu de sa légèreté. Le dernier exemple témoigne des limites du modèle actuel qui ne tient pas compte des adverbes de proposition.

Exemple de calcul de la relation sujet-verbe guidée par le modèle propositionnel, en cas d'enchâssement de proposition :

[[l'époque [où j'avais rencontré Adrienne devant le château] n'était plus déjà qu'un souvenir d'enfance]]

[[L'immense bouquet de la fête, enlevé du char [qui le portait] , avait été placé sur une grande barque ;]

Dans ces deux exemples, la proto-proposition centrale (en rouge) est tout d'abord délimitée et interprétée comme proposition car elle contient les marques permettant d'établir la relation constitutive sujet-verbe. Aucune des proto-propositions connexes ne contient les marques permettant de construire une relation sujet-verbe constitutive. Le modèle d'interprétation ne peut être calé et les proto-propositions connexes ne sont pas retenues comme proposition. C'est au niveau de structuration supérieur que le constituant opératoire (en vert) est utilisé pour établir la relation constitutive sujet-verbe de la principale, en recherchant les marques constitutives de la relation à l'intérieur des deux proto-propositions sans statut. Il s'agit encore ici d'un cas de structuration interne guidée par le modèle.

Exemple de structuration sujet-verbe-objet guidée par le modèle, dans les textes journalistiques, réalisée par le logiciel Quotes que j'ai écrit pour la société Startem :

...

"Il faut passer un nouveau contrat avec les agriculteurs européens. Je pense que ce nouveau contrat doit être de produire mieux", a estimé le ministre de l'Agriculture à l'Assemblée nationale.

Mettant l'accent sur le respect de l'environnement, la sécurité sanitaire et la qualité des produits, Jean Glavany a estimé que la PAC devait "être refondée en profondeur".

"Le contrat qu'elle avait passé avec les agriculteurs européens, il y a quarante ans, qui était de produire plus pour répondre à des problèmes d'autosuffisance alimentaire après la Seconde Guerre mondiale, est un contrat qui a été bien rempli par les agriculteurs, mais qui est totalement dépassé", a relevé le ministre de l'Agriculture.

…

L'analyse guidée par un modèle de discours rapporté intra-propositionnel exploite des invariants du genre journalistique comme la présence de guillemets pour calculer la structure sujet-verbe-objet.

La dissociation de la délimitation des constituants à valeur linguistique de l'identification de leur structure interne conduit à une méthode de structuration légère et agile, exploitant la relative inter-dépendance des marques frontières et des marques constitutives des constituants à valeur linguistique, là où des approches plus classiques, basées notamment sur les grammaires, lient intimement les deux. La résolution en deux temps, segmentation en constituants opératoires, puis structuration interne guidée par le modèle, permet de produire avec fiabilité des résultats interprétables, à valeur linguistique.

7.5.4 De la détermination du local par le global

À un niveau de structuration donné, la capacité à identifier des constituants à valeur linguistique dépend tant de la capacité à identifier les marques caractéristiques des constituants de ce niveau qu'à identifier les marques de structuration supérieure. C'est ce que nous avons vu au travers de la dissymétrie du marquage des frontières de constituants à valeur linguistique, qui dépend en partie de la capacité à identifier des marques d'ordre supérieur, mais c'est également ce que nous venons de voir dans la résolution guidée par le modèle d'interprétation, qui s'effectue dans le cadre d'un constituant d'ordre supérieur délimité.

La délimitation des constituants, préalable à l'analyse de leur structure interne, permet de réaliser une structuration d'ordre supérieur sans que la structure interne des constituants de base ait été totalement construite. Par exemple, alors qu'une structuration classique en propositions présuppose une structuration en syntagmes, l'identification de marques discriminantes de début ou de fin de proposition permet d'entamer une « méso-structuration », supérieure à la proposition, sans que la structuration en syntagmes soit finalisée.

> Exemple de calcul d'une chaîne de co-références dans les textes journalistiques, réalisée par le logiciel Quotes :
>
> 31Janv2001 FRANCE: **Glavany plaide pour une nouvelle PAC où l'on produirait "mieux".**
>
> PARIS, 31 janvier (Reuters) - Préconisant une rupture avec le modèle productiviste de l'après-guerre, Jean Glavany a plaidé pour une politique agricole européenne s'appuyant sur la qualité et le respect de l'environnement.
>
> "Il faut passer un nouveau contrat avec les agriculteurs européens. Je pense que ce nouveau contrat doit être de produire mieux", a estimé *[@Jean Glavany@]*le ministre de l'Agriculture à l'Assemblée nationale.
>
> Mettant l'accent sur le respect de l'environnement, la sécurité sanitaire et la qualité des produits, Jean Glavany a estimé que la PAC devait "être refondée en profondeur".
>
> "Le contrat qu'elle avait passé avec les agriculteurs européens, il y a quarante ans, qui était de produire plus pour répondre à des problèmes d'autosuffisance alimentaire après la Seconde Guerre mondiale, est un contrat qui a été bien rempli par les agriculteurs, mais qui est totalement dépassé", a relevé *[@Jean Glavany@]*le ministre de l'Agriculture.
>
> *[#Jean Glavany#]*Il a prévenu que la France avait l'intention d'utiliser la crise de la vache folle et ses conséquences pour convaincre ses partenaires européens et la Commission de Bruxelles de la nécessité d'un changement de cap dans les pratiques agricoles.
>
> "Maintenant devant le débat public européen qui est posé, nous devons essayer d'aller plus loin au niveau de l'Europe, notamment pour tirer les leçons de la crise bovine", *[#Jean Glavany#]*a-t-il lancé.
>
> "La détermination du gouvernement français est de tirer les leçons de cette crise et d'acter le pas dans cette reconversion de l'agriculture (...) vers ce modèle qualitatif que nous attendons", a conclu Jean Glavany.

(c) Reuters Limited 2001.

L'analyse partielle des proto-propositions n'empêche en rien d'effectuer une méso-structuration aboutissant à l'identification de chaînes de co-référence ciblées. La résolution des anaphores nominales et pronominales, respectivement notées [@ … @] *et* [# ... #]*, permet au système d'attribuer le discours rapporté au locuteur.*

7.5.5 Du statut des ressources lexicales dans la structuration en constituants opératoires

La dissociation de la question de la délimitation des constituants de celle de l'identification de leur structure interne permet une structuration interne guidée par le modèle, à l'origine de traitements parcimonieux, ne nécessitant pas de ressources lexicales massives.

Exemple de segmentation en chunks réalisée en désactivant l'accès à la base de formes de l'analyseur syntaxique WimsParser :

[J'étais] [le seul garçon] [dans cette ronde] , où [j'avais amené] [ma compagne] [toute jeune encore] , [Sylvie] , [une petite fille] [du hameau voisin] , si [vive] et si [fraîche] , [avec ses yeux noirs] , [son profil régulier] et [sa peau légèrement hâlée] ! ... [Je n'aimais] qu' [elle] , [je ne voyais] qu' [elle] , - [jusque-là] ! [A peine] [avais-je remarqué] , [dans la ronde] où [nous dansions] , [une blonde] , [grande] et [belle] , qu' [on appelait Adrienne] . [Tout d'un coup] , [suivant] [les règles] [de la danse] , [Adrienne] [se trouva placée seule] [avec moi] [au milieu] [du cercle] . [Nos tailles] [étaient pareilles] . [On nous dit] [de nous embrasser] , et [[la danse] et [le chœur]] [tournaient plus vivement] que [jamais] . [En lui donnant] [ce baiser] , [je ne pus] [m'empêcher] [de lui presser] [la main] . [[Les longs anneaux roulés] [de ses cheveux] [d'or]] [effleuraient] [mes joues] . [De ce moment] , [un trouble inconnu] [s'empara] [de moi] . - [La belle] [devait chanter] [pour avoir] [le droit] [de rentrer] [dans la danse] . [On s'assit autour] [d'elle] , et [aussitôt] , [d'une voix fraîche] et [pénétrante] , [légèrement voilée] , comme [celle] [des filles] [de ce pays brumeux] , [elle chanta] [une de ces anciennes romances pleines] [de mélancolie] et [d'amour] , qui [racontent toujours] [les malheurs] [d'une princesse enfermée] [dans sa tour] [par la volonté] [d'un père] qui [la punit] [d'avoir aimé] . [La mélodie] [se terminait] [à chaque stance] [par ces trilles chevrotants] que [font valoir] si [bien] [les voix jeunes] , quand [elles imitent] [par un frisson modulé] [la voix tremblante] [des aïeules] .

Une fois la base de formes désactivée, seules quelques marques grammaticales sont utilisées : la stabilité de l'analyse en chunks témoigne de l'agilité de l'approche.

Exemple de structuration syntaxique réalisée en désactivant l'accès à la base de formes de l'analyseur syntaxique WimsParser :

[[J'étais le seul garçon dans cette ronde,] [où j'avais amené ma compagne toute jeune encore, Sylvie, une petite fille du hameau voisin, si vive et si fraîche, avec ses yeux noirs, son profil régulier et sa peau légèrement hâlée ! ...]] [[Je n'aimais qu'elle]] [[, je ne voyais qu'elle, - jusque-là !]] [[A peine avais-je remarqué, dans la ronde] [où nous dansions, une blonde, grande et belle,] [qu'on appelait Adrienne] .] [[Tout d'un coup,] [suivant les règles de la danse] [, Adrienne se trouva placée seule avec moi au milieu du cercle] .] [[Nos tailles étaient pareilles] .] [[On nous dit de nous embrasser,]] [[et la danse et le choeur tournaient plus vivement que jamais] .] [[En lui donnant ce baiser, je ne pus m'empêcher de lui presser la main] .] [[Les longs anneaux roulés de ses cheveux d'or effleuraient mes joues] .] [[De ce moment, un trouble inconnu

s'empara de moi] .] [[- La belle devait chanter pour avoir le droit de rentrer dans la danse] .] [[On s'assit autour d'elle,]] [[et aussitôt, d'une voix fraîche et pénétrante, légèrement voilée, comme celle des filles de ce pays brumeux, elle chanta une de ces anciennes romances pleines de mélancolie et d'amour,] [qui racontent toujours les malheurs d'une princesse enfermée dans sa tour par la volonté d'un père] [qui la punit d'avoir aimé] .] [[La mélodie se terminait à chaque stance par ces trilles chevrotants] [que font valoir si bien les voix jeunes,] [quand elles imitent par un frisson modulé la voix tremblante des aïeules] .]

L'analyse de l'exemple se poursuit avec la base de formes toujours désactivée. Seules quelques marques grammaticales sont utilisées : la stabilité de la segmentation en proto-propositions et de l'analyse des relations sujet-verbe témoignent de l'agilité de l'approche.

L'approche guidée par le modèle relativise l'importance des ressources lexicales. Leur utilisation n'implique pas l'amélioration de la qualité d'analyse, problème que je nommais déjà dans ma thèse « la double incomplétude du lexique ». L'incomplétude des formes rend la ressource inutile face à un mot non répertorié (vocabulaire de spécialité, néologisme). L'incomplétude des catégories est génératrice d'erreurs en cas d'emplois non répertoriés. À cette double incomplétude s'ajoute le problème du non-respect d'une orthographe normalisée qui peut engendrer la confusion des formes en cas d'orthographes approchées (désaccentuation, faute de frappe).

Si l'approche guidée par le modèle relativise grandement l'apport des ressources lexicales, elle permet la production automatique de ressources lexicales :

Exemple de production de ressources lexicales verbales, réalisée après désactivation de l'accès à la base de formes de l'analyseur syntaxique WimsParser, sur l'exemple précédent :

aimais voyais dansions appelait trouva tournaient pus effleuraient empara assit chanta racontent punit terminait imitent

En déactivant la base de formes, WimsParser montre sa capacité à produire des ressources lexicales. Si l'exemple met ici en exergue la production de formes verbales, d'autres catégories comme les formes nominales peuvent également être extraites.

Exemple de production de ressources linguistiques avec le logiciel Quotes, qui fonctionne sans base de formes :

« Nous aimerions pouvoir établir des relations profondes avec des PME et de grands groupes français, en tirant parti de leur excellence dans certains domaines et en offrant en retour nos technologies ainsi qu'en leur ouvrant des marchés », déclare Willliam Bullock, président de LMAS.

...

« Nous devons remettre notre copie avant juillet », explique William Bullock.

Quotes ne nécessite pas de base verbale pour fonctionner : l'approche guidée par le modèle permet cependant de déduire que explique et déclare font partie d'une même classe de relateurs verbaux, en l'occurrence les verbes « dire ».

Cette discussion cible peut être à tort le statut des ressources lexicales en traitement des langues. Certes, l'utilisation de ces ressources pose des problèmes de fiabilité. Certes, la capacité à fonctionner sans ressources lexicales massives et à produire des ressources lexicales est intéressante tant sur le plan fondamental que pour des perspectives applicatives. Je pense notamment à la veille, où il est important de savoir détecter et suivre les nouvelles maladies, les nouvelles marques, les nouveaux acteurs, les nouveaux usages, sans supervision humaine.

Cependant il convient de rappeler le côté réducteur de ce commentaire qui fait porter l'attention sur le mot, un des plus petits ordres de grandeur alors que le commentaire s'applique en fait à tous les niveaux de structuration. Le problème de l'utilisation des ressources lexicales correspond en fait au problème de la catégorisation d'une partie de constituant sur des critères purement morphologiques. La production de ressources lexicales est propre à l'ordre de grandeur discuté, le mot. Des ressources tout aussi intéressantes peuvent être produites à d'autres niveaux de structuration. Citons par exemple la capacité de Quotes à produire la liste des propos d'un locuteur rapportés par un journaliste. Pour cette application, il est inutile de construire une liste de locuteurs potentiels. Il est également superflu de disposer d'une liste de verbes « dire », ou d'une liste de fonctions occupées par les locuteurs (ministre, président, directeur,...). C'est cette approche que nous souhaitons promouvoir.

7.5.6 De l'efficacité des marques dans la structuration en constituants opératoires

La fiabilité des marques n'est pas le seul critère à prendre en compte lors de la conception du modèle opératoire. Il convient également de tenir compte de leur efficacité. Par efficacité, il faut entendre la contribution de la marque à la couverture du phénomène étudié. Sur ce plan également, toutes les marques ne se valent pas. Ainsi, les marques introductrices de propositions *de sorte que* et *qui* sont l'une comme l'autre aussi fiables, mais la première est bien moins fréquente que la seconde et en cela, contribue dans une bien moindre mesure à la résolution du problème. De même, les prépositions *dans* et *via* utilisées comme marques introductrices de groupes prépositionnels n'ont pas la même efficacité, le même « rendement », dans la couverture du phénomène sur corpus. Dans une approche de la structuration du discours par langue, l'amélioration de la couverture des phénomènes étudiés passe couramment par l'augmentation du nombre de marques de même nature.

Dans les bas niveaux de structuration, la question de l'efficacité des marques est quelque peu anecdotique. Si l'on considère en effet le cas des prépositions vues comme des marques introductives de groupes prépositionnels, leur effectif réduit et la grande efficacité des plus fréquentes semblent suffisantes pour ne pas s'attarder sur la question de l'efficacité : même si quelques marques sont peu efficaces, voire même totalement inefficaces car absente des corpus traités, elles ne nuisent pas vraiment à la stabilité du système. Si l'on considère le cas des auxiliaires qui sont parfois utilisés comme marques pour l'identification du groupe verbal, on constate également que certaines formes d'auxiliaires n'apparaissent que rarement, par exemple à l'imparfait du subjonctif. Au niveau propositionnel, l'efficacité des marques est comparable à celle des bas niveaux de structuration : l'effectif des marques introductives est réduit et l'efficacité de quelques marques semble « compenser » le fait que d'autres moins fréquentes sont également utilisées. Bien entendu, la proposition étant d'un ordre supérieur au groupe prépositionnel ou au groupe verbal, les marques introductives de propositions sont moins fréquentes et donc, dans l'absolu, moins productives.

Dans les hauts niveaux de structuration, la question de l'efficacité des marques mérite une attention particulière car le phénomène prend une toute autre ampleur. Les marques introductives de constituants discursifs comme *en général*, *en conclusion*, *selon X*, *d'après X*, ont une efficacité systématiquement faible et une diversité potentiellement grande. Leur faible efficacité tient au fait que les constituants ont un empan vaste et qu'elles sont par conséquent moins fréquentes par texte que des marques de plus bas niveau. Leur diversité tient au fait qu'elles dépendent tant du genre textuel que du style personnel. Elle tient également au fait qu'en tant qu'introducteurs de constituants à large empan, leur propre empan tend à être proportionnel à celui des constituants introduits. Cette largeur d'empan favorise à l'intérieur une plus grande variété de formes brutes, difficilement énumérables. D'un point de vue opératoire, l'identification de ces marques requiert classiquement des ressources complexes, à base d'ontologies ou d'automates à états finis (Bilhaut,

2006b). Il s'agit alors d'une mise en œuvre coûteuse pour une efficacité somme toute très faible qu'il nous a semblé nécessaire de remettre en cause.

L'élévation au statut de marque pour des formes peu répandues ou très variables, si elle était déjà discutable dans les bas niveaux de structuration, est-elle bien justifiée dans les hauts niveaux de structuration ? Dans une approche par langue, on constate en effet que le relevé systématique des marques introductrices de segments discursifs aboutit à une sous-segmentation dès lors que l'on change de corpus : l'inefficacité des marques à ce niveau est manifeste. Le relevé systématique des marques introductives de constituants discursifs de haut niveau sur un corpus d'entrainement est laborieux. Les marques sont aussi diverses qu'éparses et ne permettent par conséquent pas une segmentation en constituants opératoires exploitables pour retrouver des constituants discursifs à valeur linguistique.

7.6 Du principe même de structuration fondée sur les constituants opératoires

L'étude de divers composants de traitement des langues mis au point sur corpus m'a amené à introduire le concept de constituant opératoire, défini comme la meilleure approximation d'un constituant à valeur linguistique dans une démarche de structuration ascendante.

L'étude de ce concept transverse permet de mieux comprendre comment calculer un constituant opératoire avec fiabilité et parcimonie, au regard des propriétés du matériau linguistique captable par l'ordinateur. L'étude de ce concept permet également de mieux apprendre à les utiliser pour produire des analyses qui font sens, en réalisant une analyse ascendante partielle suivie d'une analyse descendante guidée par le modèle, ou bien en réalisant une structuration de haut niveau à partir d'une analyse partielle insensible aux difficultés de structuration locale des bas niveaux.

Cette étude met aussi l'accent sur les limites intrinsèques de ce principe de structuration, au regard de l'inefficacité de certaines marques à bas niveau de structuration, de la contribution très discutable des ressources lexicales, et de l'inefficacité de certaines marques frontières dans les hauts niveaux de structuration. Tout en fixant les limites des composants basés sur ce principe de structuration, l'étude suggère que la recherche d'une méthode de structuration complémentaire à celle-ci, centrée sur l'efficacité des marques utilisées, doit être menée afin d'améliorer les capacités d'analyse automatique.

8 Prospective

8.1 Retour sur la méthode de structuration ascendante basée sur l'identification manuelle des marques

Le chapitre 7 *Regard épistémologique* a présenté une méthode de structuration en constituants linguistiques basée sur le concept de constituants opératoires. L'étude présentait le concept transverse de constituants opératoires présents dans mes propres travaux ainsi que dans des travaux connexes de la communauté. Le constituant opératoire prend des appellations variées selon le niveau de structuration concerné, par exemple token, chunk, proto-proposition, cadre de discours opératoire.

Les principes méthodologiques d'une structuration en constituants linguistiques fondée sur des constituants opératoires ont été approfondis. Nous avons abordé la question de la fiabilité des marques exploitées, en dissociant les marques frontières et les marques constitutives d'une part, les marques discriminantes et les marques secondaires d'autre part.

Nous avons explicité une méthode générale de structuration permettant de produire des analyses qui font sens, et reposant sur ces concepts. Cette méthode consiste à réaliser une analyse ascendante partielle, suivie d'une analyse descendante guidée par le modèle. Nous avons par ailleurs montré la possibilité de réaliser avec cette même méthode une structuration de haut niveau, à partir d'une analyse partielle insensible aux difficultés de structuration locale des bas niveaux.

Cette étude a cependant mis l'accent sur les limites intrinsèques de cette méthode de structuration basée sur le concept de constituants opératoires. Ces limites tiennent à l'inefficacité de certaines marques lorsqu'elles sont recherchées à un mauvais niveau de structuration. L'étude conclut à la nécessité d'une recherche portant sur l'identification de marques efficaces à différents niveaux de structuration.

8.2 Vers une méthode de structuration descendante guidée par le modèle et l'induction automatique des marques

Les synthèses des recherches effectuées plus récemment portent sur *l'alignement automatique multilingue*, *la structuration automatique du document*, et *l'analyse automatique de forums de discussion.* Elles partagent des principes de structuration très différents des principes précédents, et apportent des éléments de réponse, qui nous semblent substantiels, aux interrogations soulevées par la méthode précédente. La formalisation de la méthode me semble pouvoir faire l'objet d'un projet de recherche sur cinq ans. Les perspectives d'applications industrielles découlant de cette méthode m'ont poussé à déposer un projet de création d'entreprise. Ce projet a été lauréat national de la création d'entreprises de technologies innovantes, en 2010.

8.2.1 Une méthode de structuration guidée par le modèle

Un des principes de résolution communs aux trois travaux porte sur le fait que la résolution est guidée par un modèle d'interprétation confronté d'emblée au document. Cette approche nous semble

particulièrement pertinente au regard de l'interprétabilité du résultat par l'interprétant. La résolution n'est donc pas ascendante, comme dans la méthode précédente, mais descendante.

Ce positionnement a été particulièrement illustré en 5.5.1 du chapitre sur l'analyse des forums de discussion. La structuration thématique des forums guidée par le modèle garantit d'emblée la bonne formation de la macro-structure en deux parties, thème et rhème.

8.2.2 Une méthode reposant sur l'analyse distributionnelle

L'instanciation du modèle d'interprétation sur un document donné repose sur l'application systématique de la méthode distributionnelle. Nous avons vu au travers des trois synthèses que c'est l'utilisation systématique de cette méthode qui permet de construire un schéma relationnel assez solide pour permettre de relâcher les contraintes portant sur les formes observées.

L'ordre de grandeur du document, la taille de la fenêtre d'observation, et l'empan du phénomène à observer doivent être corrélés. En effet une fenêtre d'observation trop petite empêche de capter des marques ayant un large empan. Réciproquement, une fenêtre d'observation trop grande ne permet pas d'observer correctement la distribution de phénomènes à petit empan.

La question du choix de la bonne taille de fenêtre d'observation, et du bon cadre d'analyse, en fonction de l'empan du phénomène à observer, est l'un des points clés. L'observation est-elle menée au bon endroit et à la bonne résolution ?

Cette question mérite à elle seule de faire l'objet d'un travail de thèse. La piste principale pour atteindre l'objectif d'une sélection automatique de ces paramètres réside dans le modèle d'interprétation dont les propriétés linguistiques et sémiotiques doivent avoir été finement explicitées. La capacité à construire un schéma relationnel conforme aux contraintes de ce modèle est alors gage de bonne sélection des paramètres.

Ce sujet a été abordé dans le chapitre sur l'alignement automatique, section 3.5.1. Une segmentation systématique du document en phrases ne permet pas d'observer certains phénomènes, et il conviendrait d'utiliser d'autres fenêtres d'observation en relation avec la structure du document.

Le sujet a été abordé en section 4.7 du chapitre sur la structuration des documents. La taille de la fenêtre qui permet d'observer les transitions entre chapitres est de 4 pages. Cette taille dépend de l'empan du modèle de transition entre chapitres qui est de 3 ou 4 pages. La taille de la fenêtre qui permet d'observer les transitions entre parties est de 1 page, ce qui correspond à l'empan du modèle de transition entre parties. La capacité du logiciel à détecter les chapitres est meilleure que celle à détecter les parties. Le problème tient au modèle de transition entre partie qui a un mauvais empan.

Le même sujet a été abordé dans le chapitre 5 sur l'analyse de forums de discussion. En 5.5.1, nous avons montré qu'en sélectionnant automatiquement une fenêtre d'observation compatible avec l'ordre de grandeur du document, l'analyse distributionnelle est effectuée à la bonne résolution, ce qui permet d'identifier la frontière entre le thème et rhème.

8.2.3 Une méthode systématisant l'analyse distributionnelle

L'instanciation du modèle d'interprétation sur un document dépend directement de la capacité à mettre en évidence les relations constitutives du modèle. Il n'y a pas d'instanciation partielle du modèle : pas de sujet sans verbe dans le modèle propositionnel, pas de thème sans rhème dans le modèle de structuration thématique. C'est ici l'un des gages de l'interprétabilité des analyses.

La capacité à instancier un modèle d'interprétation dans un cadre d'analyse donné, document, proposition, forum de discussion, réside dans la capacité à construire un schéma relationnel

cohérent avec le modèle d'interprétation, à partir du contenu du cadre. La construction de ce schéma s'effectue à partir du croisement de différentes observations effectuées par analyse distributionnelle.

Ce sujet a été abordé en 3.3.1, dans le chapitre 3 portant sur l'alignement automatique multilingue. Des distributions de formes sont calculées indépendamment dans chacune des traductions. Les formes ayant des distributions similaires forment un schéma relationnel compatible avec le modèle de formes en relation de traduction. Ces formes sont extraites et constituent les alignements. L'ensemble des similarités de distribution interlingues forme un schéma relationnel complexe en relation avec les structures des documents. Ce point mérite d'être approfondi pour mieux comprendre les relations entre le local et le global.

Ce sujet a également été abordé en section 4.7 du chapitre sur la structuration des documents. La distribution des transitions entre parties est mise en relation avec la distribution des relations entre chapitres. C'est la conformité de la structure résultante avec le modèle de structure logique qui permet de valider l'instanciation du modèle.

8.3 Conclusion

La capacité à bâtir un schéma relationnel cohérent avec un modèle d'interprétation dont les propriétés linguistiques et sémiotiques ont été finement explicitées laisse envisager des résultats interprétables de qualité produit automatiquement. Si nous pensons avoir contribué à la mise au point d'une méthode allant dans ce sens, il n'en reste pas moins que sans excellents linguistes, sans excellents sémioticiens, il n'y a de traitement des langues qui vaille.

Ce défi scientifique interdisciplinaire est passionnant. Il sous-tend les trois axes de variation que nous avons introduits au début de ce mémoire :

- la variation en langues, ou comment mettre au point des méthodes d'analyse indépendante des langues. Cet aspect est parfois nommé « multilingue », « alingue », ou « indépendant des langues » dans notre équipe. Quoique nous n'ayons pas décliné nos méthodes sur cet axe pour chacun des points abordés dans le mémoire, il reste présent dans notre pratique;
- la prise en compte du genre des textes, ou comment détecter et tenir compte de régularités de forme ou de style dans l'analyse. Cet aspect est illustré par le traitement de collections homogènes d'articles de presse, de forums de discussion, d'articles scientifiques, de livres;
- la gestion des ordres de grandeur, ou comment analyser de manière raisonnée des documents de taille variée. Nous avons traité de l'analyse syntaxique de phrases, jusqu'à l'analyse de la structure des livres, en passant par l'analyse de la citation dans les dépêches de presse.

Références bibliographiques

ABNEY, Steven. 1991. Parsing by chunks. Principle-based parsing : Computation and Psycholinguistics, pp. 257-278, 1991.

ANTONACOPOULOS, Apostolos, PLETSCHACHER, Stefan, BRIDSON, David, PAPADOPOULOS, Christos. 2009. ICDAR 2009 Page Segmentation Competition. ICDAR 2009: 1370-1374

APIDIANAKI, Marianna. 2008. *Acquisition automatique de sens pour la désambiguïsation et la sélection lexicale en traduction*. Thèse de doctorat. Université Paris-Diderot. Septembre.

BACHIMONT, Bruno. 1998. Intelligence artificielle et écriture dynamique : de la raison graphique à la raison computationnelle. *Au nom du sens*, Petitot et Fabbri (éds), Grasset, pp. 290-319.

BEAUDOUIN, Valérie. 2002. De la publication a la conversation. Lecture et écriture électroniques. *Réseaux*, 2002/6 no 116, p. 199-225. DOI : 10.3917/res.116.0199

BELAÏD, Abdel et TOUSSAINT, Yannick. 2000.Une méthode morpho-syntaxique pour la reconnaissance de tables de matières. Colloque International Francophone sur l'Ecrit et le Document (CIFED'00), Lyon, juillet.

BESAGNI, Dominique et BELAÏD, Abdel. 2004.Citation Recognition for Scientific Publications in Digital Libraries. DIAL 2004: pp. 244-252

BILHAUT, Frédérik. 2006 (a). *Analyse automatique de structures thématiques discursives : Application à la recherche d'information*. Thèse de doctorat. Université de Caen Basse-Normandie.

BILHAUT, Frédérik. 2006 (b). Introducteurs intra-prédicatifs d'univers de discours et leur détection automatique . *Schedae*, prépublication n° 6, fascicule n° 1, pp. 41-50.

BOURDAILLET, Julien et GANASCIA, Jean-Gabriel. 2007. Practical Block Sequence Alignment with Moves. *1st International Conference on Language and Automata Theory and Applications (LATA 2007*), pp. 199-210. Tarragona, España.

BRATITSIS, Tharrenos, & DIMITRACOPOULOU, Angelique. 2007. Interaction Analysis in Asynchronous Discussions: Lessons learned on the learners' perspective, using the DIAS system. In C. A. Chinn, G. Erkens & S. Puntambekar (Eds.), Proceedings of the Computer Supported Collaborative Learning (CSCL) 2007, pp. 87-89.

BROWN P., COCKE J., DELLA PIETRA S., DELLA PIETRA V., JELINEK F., MERCER R. & ROOSSIN P. 1988. A statistical approach to language translation. *Coling Budapest: Proceedings of the 12th International Conference on Computational Linguistics*, 22-27 August 1988, John von Neumann Society for Computing Sciences, Budapest, Hungary; vol. 1, pp. 71-76.

BROWN, Peter F., LAI, Jennifer C. et MERCER, Robert L. 1991. Aligning sentences in parallel corpora. *Proc. 29 th Annual Meeting of the Association for Computational Linguistics*, 18-21 June, Berkley, California, pp. 169-176. http://portal.acm.org/ft_gateway.cfm?id=981366&type=pdf

BRUILLARD, Éric. 2010. Rapport final ERTé CALICO Communautés d'apprentissage en ligne, instrumentation, collaboration. Créteil, École normale supérieure de Cachan / INRP STEF.

CHANOD, Jean-Pierre, CHIDLOVSKII, Boris, DÉJEAN, Hervé, FAMBON, Olivier, FUSELIER, Jérome ,

JACQUIN, Thierry, et MEUNIER, Jean-Luc. 2005. From legacy documents to xml: A conversion framework. 9th European Conference on Research and Advanced Technology for Digital Libraries, Vienna, Austria, September 18-23.

CHAROLLES Michel (1997) L'ENCADREMENT DU DISCOURS : UNIVERS, CHAMPS, DOMAINES ET ESPACES. *Cahier de Recherche Linguistique*, LANDISCO, URA-CNRS 1035, Université Nancy 2, n° 6, 1-73.

CHEN, Stanley F. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 9–16.

CHIDLOVSKII, Boris et FUSELIER, Jérôme. 2005. HTML-to-XML Migration by means of sequential learning and grammatical inference. IJCAI 05 Workshop on Grammatical Inference Applications, Edinburgh, Scotland, 30 July, 2005.

CHURCH K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. Proceedings of the second conference on Applied natural language processing, pp. 136–143, Austin, Texas: Association for Computational Linguistics.

CLOUARD Régis, Elmoataz Abderrahim et Angot François. 1997. PANDORE : une bibliothèque et un environnement de programmation d'opérateurs de traitement d'images. Rapport interne du GREYC, Caen, France, Mars 1997.

COURSIL, Jacques. 2000. *La fonction muette du langage*. Matoury : Ibis rouge.

CRESS, Ulrike. 2008. The need for considering multilevel analysis in CSCL research: An appeal for the use of more advanced statistical methods. International Journal of Computer-Supported Collaborative Learning, 3 (1), pp. 69–84.

CROCHEMORE, Maxime, HANCART Christophe et LECROQ, Thierry. 2001. *Algorithmique du texte*, Vuibert.

CROMIÈRES, Fabien. 2006. Sub-sentential alignment using substring co-occurrence counts, *Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop* , Sydney, Australia, 2006, Association for Computational Linguistics, pp. 13–18.

CROMIÈRES, Fabien. 2010. *Vers un plus grand lien entre alignement, segmentation et structure des phrases*. Thèse de doctorat. Université de Grenoble, Grenoble.

DÉJEAN, Hervé. 1998. *Concepts et algorithmes pour la découverte des structures formelles des langues*. Thèse de doctorat de l'université de Caen.

DÉJEAN, Hervé, GAUSSIER, Éric, GOUTTE, Cyril et YAMADA, Kenji. 2003 Reducing Parameter Space for Word Alignment. *HLT-NAACL 2003 Workshop on Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, pp. 23-26, Edmonton, May-June.

DÉJEAN, Hervé et MEUNIER, Jean-Luc. 2007. Logical Document conversion: combining functional and formal knowledge. *Symposium on Document Engineering*, Winnipeg, Canada, August 28-31, 2007. http://www.xrce.xerox.com/Publications/Attachments/2007-011/2007-011.pdf

DÉJEAN, Hervé et MEUNIER, Jean-Luc. 2010. Reflections on the INEX structure extraction competition. *Document Analysis Systems* 2010: pp. 301-308

DE LAAT, Maarten, LALLY, Vic, LIPPONEN, Lasse et SIMONS, Robert-Jan. 2007. Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. International Journal of Computer-Supported Collaborative Learning, Vol. 2, No. 1. March, pp. 87-103.

DORI, Dov, DOERMANN, David, SHIN, Christian, HARALICK, Robert, PHILLIPS, Ihsin, BUCKMAN, Mitchell et ROSS, David. 1997. The representation of document structure: A generic object-process analysis. In P. Wang and H. Bunke, editors, *Handbook on Optical Character Recognition and Document Image Analysis*. World Scientific, Singapore, pp. 421–456.

DOUCET, Antoine, KAZAI, Gabriella. 2009. ICDAR 2009 Book Structure Extraction Competition. In: IEEE (ed.) *10th International Conference on Document Analysis and Recognition ICDAR 2009*. pp. 1408–1412. Barcelona, Spain.

DOUCET, Antoine, KAZAI, Gabriella, DRESEVIC, Bodin, UZELAC, Aleksandar, RADAKOVIC, Bogdan, TODIC, Nikola. 2011. Setting up a competition framework for the evaluation of structure extraction from ocr-ed books. *International Journal of Document Analysis and Recognition (IJDAR)*, Special Issue on Performance Evaluation of Document Analysis and Recognition Algorithms. 14(1), pp. 45–52, DOI 10.1007/s10032-010-0127-3

DRESEVIC, Bodin, UZELAC, Aleksandar, RADAKOVIC, Bogdan, TODIC, Nikola. 2009. Book Layout Analysis: TOC Structure Extraction Engine. *Advances in Focused Retrieval: 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, (INEX 2009)*, Schloss Dagstuhl, Germany, pp. 164-171.

DUPUY, Jean-Philippe. 2009. Structure de la page Web : texte et paratexte. Revue des Interactions Humaines Médiatisées 9, 1/2008. pages 25-42.

ERJAVEC, Tomaž, IGNAT, Camelia, POULIQUEN, Bruno et STEINBERGER, Ralf (2005). Massive multilingual corpus compilation: Acquis Communautaire and totale. *Journal Archives of Control Sciences*, vol. 15(LI), 2005, no. 4, pp. 529-540.

FAIRON, Cédrick, KLEIN, Jean-René et PAUMIER, Sébastien. 2006. Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête 'Faites don de vos SMS à la science'. Presses universitaires de Louvain, Louvain-la-Neuve. *Cahiers du Cental*, 3.1. 136p.

FOSTER George, ISABELLE, Pierre et PLAMONDON Pierre. 1997. Target-Text Mediated Interactive Machine Translation. *Machine Translation*, pp. 175-194. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.5506&rep=rep1&type=pdf

FOSTER, George, LANGLAIS, Philippe et LAPALME, Guy. 2002. User-Friendly Text Prediction for Translators. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, July, pp. 148-155. Association for Computational Linguistics. http://acl.ldc.upenn.edu/W/W02/W02-1020.pdf

GALE, William A. et CHURCH, Kenneth W. 1991. Identifying word correspondences in parallel texts. *Fourth DARPA Speech and Natural Language Workshop*, San Mateo, California : Morgan Kaufmann, pp. 152-157.

GALE, William A. et CHURCH, Kenneth W. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19 (1), pp. 75-102.

GAUSSIER, Éric. 1998. Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. *Proceedings of the 17th International Conference on Computational Linguistics: COLING-98*, Montréal, Canada, pp. 444-450.

GOSME, Julien, MEKKI, Wigdan, DEBILI, Fathi, LEPAGE, Yves et LUCAS, Nadine. 2010. The GREYC/LLACAN Machine Translation Systems for the IWSLT 2010 Campaign. *Proceedings of IWSLT*, Paris, France, 2010, pp. 59–65.

GOUTTE, Cyril, YAMADA, Kenji et GAUSSIER, Eric. 2004. Aligning words using matrix factorisation. Proceeding ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp. 502–509.

HABERT, Benoît et ZWEIGENBAUM, Pierre. 2003. Classer les mots : sémantique à gros grain et méthodologie harrissienne. Revue de Sémantique et Pragmatique , (12) :25–45.

HEARST, Marti A. 1997. Texttiling : Segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1):33–64.

HUBERT, Frédéric. 2003. *Modèle de traduction des besoins d'un utilisateur pour la dérivation de données géographiques et leur symbolisation par le Web*. Thèse de doctorat. Université de Caen.

ISABELLE, Pierre. 1992. La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie. *Meta : journal des traducteurs*, Volume 37, numéro 4, décembre, pp. 721-737. http://id.erudit.org/iderudit/003228ar

KINTSCH, Walter et VAN DIJK, Teun A. 1978. Toward a model of text comprehension and production. Psychological Review, 85, pp. 363-394.

KOEHN, Philipp, OCH, Franz J. et MARCU, Daniel. 2003. Statistical phrase based translation. *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*. pp. 48-54. Edmonton, May-June. http://acl.ldc.upenn.edu/N/N03/N03-1017.pdf

KRAIF, Olivier. 2001. *Constitution et exploitation de bi-textes pour l'aide à la Traduction*. Thèse de doctorat, Université de Nice Sophia Antipolis.

KUSHMERICK Nicholas. 2000. Wrapper Induction: Efficiency and Expressiveness. Artificial Intelligence, pp. 15-68. http://www.dfki.de/~neumann/esslli04/reader/templatelearning/kushmerick-aij2000.pdf

KUPIEC, Julian. 1993. An Algorithm for finding Noun Phrase Correspondences in Bilingual Corpora. *Proceedings of the 31st Annual Meeting*, Columbus, OH. Association for Computational Linguistics. pp. 17-22.

LANGÉ, Jean-Marc et GAUSSIER, Éric. 1995. Alignement de corpus multilingues au niveau des phrases. TAL. *Traitement automatique des langues*, 36(1-2) pp. 67- 80.

LARDILLEUX, Adrien et LEPAGE, Yves. 2008. A truly multilingual, high coverage, accurate, yet simple, subsentential alignment method, *Proceedings of the xth conference of the Association for Machine Translation in the Americas*, pp. 125-132, Hawai'i, USA.

LARDILLEUX, Adrien et LEPAGE, Yves. 2009. Sampling based multilingual alignment. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pp. 214–218, Borovets, Bulgaria, September.

LARDILLEUX, Adrien. 2010. *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. Thèse de doctorat. Université de Caen Basse-Normandie.

LASCARIDES, Alex et ASHER, Nicolas [1993] Temporal Interpretation, Discourse Relations and Commonsense Entailment, Linguistics and Philosophy, 16(5), pp437-493, Kluwer Academic Publishers, Dordrecht, Holland.

LAVALLARD Anne. 2008. *Exploration interactive d'archives de forums : le cas des jeux de rôle en ligne*. Thèse de doctorat. Université de Caen Basse-Normandie.

LEPAGE, Yves, LARDILLEUX, Adrien et GOSME, Julien. 2009. The GREYC Translation Memory for the IWSLT 2009 Evaluation Campaign: one step beyond translation memory. *Proceedings of the 6th International Workshop on Spoken Language Translation (IWSLT09)*, Japon.

LEJEUNE, Gaël, DOUCET, Antoine et LUCAS, Nadine. 2010. Tentative d'approche multilingue en extraction d'information. *Analyse statistique des données textuelles* JADT (vol 3; Rome, Italie: Lexicometrica), pp. 1259-1268.

LUCAS, Nadine. 2005. Les procédés d'exposition et de développement collectif dans un forum pédagogique : le cas Maxime. *Symfonic*, 20-22 janvier, Amiens, en ligne sur edutice.

LUCAS, Nadine. 2009. *Modélisation différentielle du texte: de la linguistique aux algorithmes*. Mémoire d'habilitation à diriger des recherches. Caen : Université de Caen Basse-Normandie.

LUCAS, Nadine. 2011. Citation interactionnelle et citation du problème dans les forums de discussion en ligne. Citations I: *Citer à travers les formes. intersémiotique de la citation.* L. Rosier et C. Stolz (eds). Bruxelles: Academia Bruylant, sous presse.

MAO Song, ROSENFELD, Azriel et KANUNGO, Tapas. 2003. Document Structure Analysis Algorithms: A Literature Survey. *Proc. SPIE Electronic Imaging*, Santa Clara, California, USA, January 2003, pp.197-207.

MARCOCCIA, Michel. 2010. Les forums de discussion d'adolescents : pratiques d'écritures et compétences communicatives. *Revue française de linguistique appliquée*, 2010/2 (Vol. XV). Éditeur Pub. Linguistiques.

MCNAMEE, Paul, NICHOLAS, Charles et MAYFIELD, James. 2008. Don't have a stemmer?: be un+concern+ed. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08}*, isbn 978-1-60558-164-4, Singapore, Singapore, pp. 813-814, doi : http://doi.acm.org/10.1145/1390334.1390518, ACM, New York, NY, USA.

MOORE, Robert C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. Machine Translation: From Research to Real Users. *Proceedings, 5th Conference of the Association for Machine Translation in the Americas*, Tiburon, California, Springer-Verlag, Heidelberg, Germany, pp. 135-244.

MOORE, Robert C., YIH, Wen-tau, BODE, Andreas. 2006. Improved Discriminative Bilingual Word Alignment *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 513–520, Sydney, July.

MUSLEA, Ion, MINTON, Steven et KNOBLOCK Craig. 2001. Hierarchical wrapper induction for semistructured sources. Journal of Autonomous Agents and Multi-Agent Systems, pp. 93-114.

NAKAMURA-DELLOYE, Yayoi. 2007. Méthodes d'alignement des propositions : un défi aux traductions croisées. *Actes de TALN*, pp. 223-232.

OCH, Franz. J. et NEY, Hermann. 2002. Discriminative training and maximum entropy models for statistical machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, Philadelphia, July 2002; pp. 295-302. http://acl.ldc.upenn.edu/acl2002/MAIN/pdfs/Main074.pdf

OCH, Franz J. et NEY, Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.

OCH, Franz J. et NEY, Hermann. 2004. The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, Vol. 30, No. 4., pp. 417-449. doi:10.1162/0891201042544884. http://www.mitpressjournals.org/doi/pdf/10.1162/0891201042544884

PASCUAL, Elsa et VIRBEL, Jacques. 1996. Semantic and layout properties of text punctuation. *Proceedings of the Association for Computational Linguistics Workshop on Punctuation*, pp. 41–48, June. Santa Cruz, California.

PÉRY-WOODLEY, Marie-Paule. 2000. Cadrer ou centrer son discours ? Introducteurs de cadres et centrage. *Verbum*, 22(1), 59-78

PINATEL, Pascalie. 2003. Coloriage thématique à l'intérieur d'un document: approche contextuelle. Rapport de projet DESS RADI. Université de Caen Basse-Normandie.

RASTIER, François. 2002. Enjeux épistémologiques de la linguistique de corpus. *Journées de Linguistique de Corpus*, Lorient. http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html

ROMERO, Cristóbal et VENTURA, Sebastián. 2010. Educational data mining: a review of the state-of-the-art. *IEEE Trans. Syst. Man Cybern*. C 40(6), pp. 601–618.

ROY, Thibault et BEUST, Pierre. 2007. Ressources termino-ontologiques différentielles personnelles : construction et projection en corpus. *Revue I3 Information - Interaction - Intelligence : Textes et ressources terminologiques et/ou ontologiques : évolution et maintenance*. Numéro hors série 2006, ISBN : 2-85428-762-2, janvier 2007, pp. 35-60. http://users.info.unicaen.fr/~beust/Papiers/T_Roy_P_Beust_RevueI3.pdf

SAKAMOTO, Hiroshi, ARIMURA, Hiroki, and ARIKAWA, Setsuo. 2002. Knowledge Discovery from Semistructured Texts. In Arikawa and Shinohara, eds., *Progress in Discovery Science*, Springer.

SALTON, Gerard, ALLAN, James, and BUCKLEY, Chris. 1994. Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2), pp. 97-108.

SEIFI, Massoud, GUILLAUME, Jean-Loup, LATAPY, Matthieu, LE GRAND, Bénédicte. 2010. Interactive multiscale visualization of huge graphs: application to a network of weblogs. *8th Workshop on Visualization and Knowledge Extraction (EGC 2010)*, Hammamet, Tunisia.

SHAFAIT, Faisal, KEYSERS, Daniel, and BREUEL, Thomas M. 2008. Performance Evaluation and Benchmarking of Six Page Segmentation Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Volume 30, Issue 6. pages 941-954. ISSN:0162-8828.

SIMARD, Michel, FOSTER George et ISABELLE, Pierre. 1992. Using cognates to align sentences in bilingual corpora. *TMI-92*, Montréal, Québec, pp. 67-81.

SIMARD, Michel. 1999. Text-translation alignment: three languages are better than two. *Proceedings of EMNLP/VLC-99*. College Park, MD, pp. 2–11.

STEINBERGER, Ralf, POULIQUEN, Bruno, WIDIGER, Anna, IGNAT Camelia, ERJAVEC Tomaž, TUFIŞ Dan, VARGA, Dániel. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pp. 2142-2147. Genoa, Italy, 24-26 Mai 2006.

TAKEZAWA, Toshiyuki, SUMITA, Eiichiro, SUGAYA, Fumiaki, YAMAMOTO, Hirofumi, and YAMAMOTO, Seiichi. 2002. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World. *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 147–152, Las Palmas de Gran Canaria, Spain.

VALETTE, Mathieu et RASTIER, François. 2008. Prévenir le racisme et la xénophobie. [En ligne], Volume XIII - n°3. Coordonné par Carine Duteil.

VAN DIJK, Teun A. 1985. Structures of news in the press. Dans Teun A. van Dijk, Discourse and communication: New approaches to the analysis of mass media discourse (pp. 69-93). Berlin: De Gruyter.

VAN DIJK, Teun A. 1988. *News as discourse*. Lawrence Erlbaum Associates, Hillsdale N.J.

VARGA, Daniel., NÉMETH, Laszlo, HALÁCSY, Peter, KORNAI, Andras, TRÓN, Viktor et NAGY, Viktor. 2005. Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, pp. 590-596.

VERGNE Jacques. 1999. *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique automatique non combinatoire*, Habilitation à Diriger les Recherches, Université de Caen.

VERGNE, Jacques. 2002. Une méthode pour l'analyse descendante et Calculatoire de corpus multilingues : application au calcul des relations sujet-verbe. Actes de la conférence TALN 2002, Batz. http://www.info.unicaen.fr/%7Ejvergne/TALN_2002/TALN2002_JVergne.doc.pdf

VERGNE, Jacques. 2005. Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. *Conférence internationale sur le document électronique,*

Cide8, Beyrouth, 2005. http://www.info.unicaen.fr/%7Ejvergne/CIDE8/CIDE_JVergne6.pdf

VERGNE, Jacques. 2009. Un chunker multilingue endogène. *Actes de la conférence TALN 2009*, Senlis, 24-26 juin 2009. http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_149.pdf

VÉRONIS, Jean. 2000. Alignement de corpus multilingues. *Ingénierie des langues*, Paris, Hermès, pp. 151-171.

VIRBEL, Jacques. 1989. The Contribution of Linguistic Knowledge to the Interpretation of Text Structure. In J. André, V. Quint et R. Furuta (éds), *Structured Documents*, Cambridge MA, Cambridge University Press : pp. 161-181.

WONG, Kwan Y., CASEY, Richard G. et WAHL, Friedrich M. 1982. Document Analysis System, *IBM journal of Research Development*, 26(6), pp. 647-656.

WU, Dekai. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*. Vol. 23(3), Septembre. MIT Press Cambridge, MA, USA.

Publications et travaux encadrés

Liste complète des publications

GIGUET, Emmanuel. 1995. Categorization according to Language: A step toward combining Linguistic Knowledge and Statistic Learning. *International Workshop of Parsing Technologies (IWPT'95*), Prague - Karlovy Vary, September 20-24.

GIGUET, Emmanuel. 1995. Multilingual Sentence Categorization according to Language. *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL) SIGDAT Workshop "From text to tags : Issues in Multilingual Language Analysis"*, pages 73-76, Dublin, Ireland, March.

GIGUET, Emmanuel. 1996. The Stakes of multilinguality: Multilingual text tokenization in Natural Language Diagnosis. *Proceedings of the 4th Pacific Rim International Conference on Artificial Intelligence (PRICAI) Workshop "Future issues for Multilingual Text Processing"*, Cairns, Australia, August 27.

GIGUET, Emmanuel et VERGNE, Jacques. 1997. Syntactic analysis of unrestricted French. *Proceedings of the International Conference on Recent Advances in Natural Languages Processing (RANLP'97)*, pages 276-281, Tzigov Chark, Bulgaria, September 11-13.

GIGUET, Emmanuel. 1997. Toward an Adequate Model for Automatic Syntactic Parsing. *Poster Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI'97)*, Nagoya, Aichi, Japan, August 23-29.

GIGUET, Emmanuel et VERGNE, Jacques. 1997. Syntactic Structures of Sentences from Large Corpora. *Demonstration Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, Washington, USA, April 1-3. 1997.

GIGUET, Emmanuel et VERGNE, Jacques. 1997. From Part-of-Speech Tagging to Memory-based Deep Syntactic Analysis. *Proceedings of the International Workshop on Parsing Technologies (IWPT'97)*, MIT, Boston, Massachussets, USA, September 17-20.

VERGNE, Jacques et **GIGUET, Emmanuel**. 1998. Regards Théoriques sur le "Tagging". *Fifth annual conference Le Traitement Automatique des Langues Naturelles (TALN 1998)*, Paris, France, June 10-12.

GIGUET, Emmanuel. 1998. *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse de doctorat, spécialité Informatique. Université de Caen, Décembre.

FERRARI, Stéphane, **GIGUET, Emmanuel**, LUCAS, Nadine, VERGNE, Jacques. 2000. Projet LINGUIX, recherche d'informations et traitements linguistiques: le cas des métaphores", *Le document électronique dynamique, 3ème colloque international sur le document électronique (CIDE 2000)*, 4-6 juillet, Lyon, ed. Gaio et Trupin, Caen, pp. 279-293.

GIGUET, Emmanuel, LUCAS, Nadine et COUSIN, Grégoire. 2000. Document structure identification as a means for relevant indexation. *International Conference on Intelligent text processing and Computational Linguistics (CICLING-2000).* Mexico, February.

LUCAS, Nadine, **GIGUET, Emmanuel** et VERGNE, Jacques. 2001. Détection automatique de la citation et du discours rapporté dans les textes informatifs. *Le discours rapporté dans tous ses états : Question de frontières*. Communication à colloque international avec comité de sélection. Bruxelles, novembre.

GIGUET, Emmanuel et LUCAS, Nadine. 2002. Intégration d'Unicode - conception d'un agent de recherche d'information sur Internet. Document numérique, n° spécial "Unicode, écriture du monde?" sous la direction de A. Jacques et H. Hudrisier. Vol. 6 n° 3-4. pp. 225-236.

GIGUET, Emmanuel. 2005. Modélisation de l'activité expérimentale du chercheur en traitement des langues sur corpus multilingues. Journée "Articuler les traitements sur corpus", du 12 février, organisée par Benoît Habert, Serge Heiden et André Salem.

GIGUET, Emmanuel et APIDIANAKI, Marianna. 2005. Alignement d'unités textuelles de taille variable. *Journée Internationale de la Linguistique de Corpus*. Lorient. Septembre.

GIGUET, Emmanuel. 2005a. Multi-grained alignment of parallel texts with endogenous resources. *Proceedings of the Recent Advances in Natural Language Processing (RANLP) International Workshop "New Trends in Machine Translations"*. pages 12-17. Borovets, Bulgaria. 24 septembre.

GIGUET, Emmanuel. 2005b. Linguistic-poor, multi-grained alignment of parallel text sequences. *Workshop "EU Enlargement and Integration Workshop: Exploiting parallel corpora in up to 20 languages"*. 26-27 septembre. Arona, Italy.

GIGUET, Emmanuel et LUQUET, Pierre-Sylvain. 2005. Multilingual Lexical Database Generation from parallel texts with endogenous resources. *PAPILLON-2005 Workshop on Multilingual Lexical Databases*. Chiang Rai, Thaïland. December 12-14.

MANGEOT, Mathieu et **GIGUET, Emmanuel**. 2005. *Multilingual aligned corpora from movie subtitles*. Technical report, Condillac-LISTIC.

LUCAS, Nadine et **GIGUET, Emmanuel**. 2005. UniTHEM, un exemple de traitement linguistique à couverture multilingue. *Conférence Internationale sur le Document Electronique (CIDE 8*). Beyrouth, Liban, pp. 115-132

LUCAS, Nadine, SIDIR, Mohamed et **GIGUET, Emmanuel**. 2006. Analyse de forums dans la formation à distance. *Conférence Internationale sur le Document Electronique (CIDE 9*). Fribourg, Suisse. Europia. pp. 169-180.

GIGUET, Emmanuel et LUQUET, Pierre-Sylvain. 2006. Multilingual Lexical Database Generation from parallel texts in 20 European languages with endogenous resources. *Poster Proceedings of the ACL-COLING-2006 International Conference*. July 16-22. Sydney, Australia.

SIDIR, Mohamed, LUCAS, Nadine et **GIGUET, Emmanuel**. 2007. De l'analyse des discours à l'analyse structurale des réseaux sociaux : une étude diachronique d'un forum éducatif. *Revue Sticef*, vol. 13. numéro spécial forum. ISSN : 1764-7223, mis en ligne le 20/03/2007.

LUCAS, Nadine et **GIGUET, Emmanuel**. 2008. Robust adaptive discourse parsing for e-learning fora. *The 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008)*, Santander, Cantabria, Spain July 1st- July 5th, 2008, P. Diazet al., IEEE. pp. 730-732.

GIGUET, Emmanuel. 2008. *Rapport scientifique du projet Résurgence*. Rapport interne, Groupe de Recherche en Informatique, Image, Instrumentation et Automatique de Caen. Octobre.

GIGUET, Emmanuel, LUCAS, Nadine et Chircu, Catalina. 2008. Le projet Résurgence : Recouvrement de la structure logique des documents électroniques. *JEP-TALN-RECITAL'08 Session "Show & Tell"*, juin, France.

GIGUET, Emmanuel, BAUDRILLART, Alexandre et LUCAS Nadine. 2009. Resurgence for the Book Structure Extraction Competition. *INEX*, Woodlands of Marburg, Ipswich, Queensland, Australia, December 6–10, pp. 136-142.

BLONDEL, François-Marie et **GIGUET, Emmanuel**. 2009. CALICO, une plate-forme pour visualiser et

analyser des discussions. Animation d'un atelier dans le cadre du Colloque international EPAL "Echanger pour apprendre en ligne", Grenoble, France, 4 juin.

GIGUET, Emmanuel, LUCAS, Nadine, BLONDEL, François-Marie, BRUILLARD, Éric. 2009. Share and explore discussion forum objects on the Calico website. Dans "*8th International Conference on Computer Supported Collaborative Learning (CSCL2009)*", June 8-13, 2009, University of Aegean, Rhodes, Greece. Best Technology Design Nomination

GIGUET, Emmanuel, LUCAS, Nadine, BLONDEL, François-Marie, BRUILLARD, Éric. 2009. The Calico Platform: Multilingual Monitoring of Online Discussions. Dans *8th International Conference on Computer Supported Collaborative Learning (CSCL2009) Workshop: "Interaction Analysis and Visualization for Asynchronous Communication: Analysis Methods, Tools, and Research Questions*", June 8-13, 2009, University of Aegean, Rhodes, Greece.

GIGUET, Emmanuel et LUCAS, Nadine. 2009. Creating Discussion threads graphs with Anagora. Dans "*8th International Conference on Computer Supported Collaborative Learning (CSCL2009)*", June 8-13, 2009, University of Aegean, Rhodes, Greece.

GIGUET, Emmanuel et LUCAS, Nadine. 2010. The Book Structure Extraction Competition with the Resurgence Software for Part and Chapter Detection. *INEX 2010 pre-proceedings*, Amsterdam.

GIGUET, Emmanuel et LUCAS, Nadine. 2010. "The Book Structure Extraction Competition with the Resurgence Software at Caen University." *Inex 2009*, edited by Shlomo Geva, Jaap Kamps and Andrew Trotman, pp. 170-178. Heidelberg: Springer.

LUCAS, Nadine et **GIGUET, Emmanuel**. 2010. L'analyse de forums par ThemAgora. JOCAIR. Amiens.

GIGUET, Emmanuel et LUCAS, Nadine. 2011. "The Book Structure Extraction Competition with the Resurgence Software for Part and Chapter Detection at Caen University." *Comparative Evaluation of Focused Retrieval: 8th International Workshop of the Initiative for the Evaluation of Xml Retrieval (Inex 2010)*, ed. by S. Geva, J. Kamps, R. Schenkel and A. Trotman. Berlin / Heidelberg: Springer.

LEJEUNE, Gaël, BRIXTEL, Romain et **GIGUET, Emmanuel**. Deft 2011: Appariement de résumés et d'articles scientifiques fondé sur des distributions de chaînes de caractères. *DEFT 2011*, Montpellier. À paraître.

Thèses

BRIXTEL, Romain. 2011. *Alignement endogène de documents, une approche multilingue et multi-échelle*. Thèse de doctorat. Université de Caen Basse-Normandie, Caen, France. Janvier. (co-direction avec Jacques Vergne et Christine Durieux)

LECLUZE, Charlotte. 2011. *Parallélisation de textes multilingues pour l'extraction automatique de dictionnaires de langues*. Thèse de doctorat. Université de Caen Basse-Normandie, Pertimm, soutenance prévue fin 2011 (co-direction avec Jacques Vergne)

Publications des doctorants

BRIXTEL, Romain. 2007. Extraction endogène de structure pour un alignement multilingue. *TALN-RECITAL'2007* : pp. 367-376. Toulouse, France. June 2007.

BRIXTEL, Romain. 2009. Extraction d'une structure endogène de document pour l'alignement. *Congrès de l'ACFAS 2009*. Ottawa, Canada.

BRIXTEL, Romain, LESNER, Boris, BAGAN, Guillaume et BAZIN, Cyril. 2009. De la mesure de similarité de

codes sources vers la détection de plagiat: le « Pomp-O-Mètre ». *MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication (MajecSTIC 2009)*. Avignon, France.

BRIXTEL, Romain, FONTAINE, Mathieu, LESNER, Boris, BAZIN, Cyril et ROBBES, Romain. 2010. Language-Independent Clone Detection Applied to Plagiarism Detection. *Tenth IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM 2010)*. IEEE Computer Society.Timişoara, Romania.

LECLUZE, Charlotte. 2011. Recherche d'une granularité optimale pour l'alignement multilingue : N-grammes de caractères ou N-grammes de mots ? *Jetou*. Toulouse, France

LESNER, Boris, **BRIXTEL, Romain**, BAGAN, Guillaume et BAZIN, Cyril. 2010. A Novel Framework to Detect Source Code Plagiarism: Now, Students Have to Work for Real!. *SAC '10: Proceedings of the 2010 ACM symposium on Applied Computing*, ACM. 2010.

MESNAGE, Cédric, **BRIXTEL, Romain** et CARMAC, Mark. 2010. Serendipitous Social Shuffle. *Wormrad 2010*. ACM. Barcelona, Espana.

Stages de master recherche en informatique

BRIXTEL, Romain. 2007. *Alignement automatique de corpus multilingues*. Mémoire de Master Recherche. Université de Caen Basse-Normandie, Caen, France. (co-direction avec Jacques Vergne)

BAUDRILLART, Alexandre. 2011. Résolution de la coréférence de longue portée dans les textes. Université de Caen Basse-Normandie, Caen, France, stage en cours. (co-direction avec Nadine Lucas)

Stages de master recherche en sciences de la traduction

LEMOINE, Anne. 2006. *Alignement sémantique de corpus multilingue : une perspective traductologique*, Mémoire de Master conjoint franco-hellénique mention Sciences du langage, spécialité Sciences de la Traduction : Traductologie et Sciences cognitives, Université de Caen Basse-Normandie, Caen, France et Université Ionienne, Corfou, Grèce. (co-direction avec Jacques Vergne)

SACHTOURI, Calliopi. 2006. *Étude comparative des chaînes anaphoriques dans vingt langues européennes*, Mémoire de Master conjoint franco-hellénique mention Sciences du langage, spécialité Sciences de la Traduction : Traductologie et Sciences cognitives, Université de Caen Basse-Normandie, Caen, France et Université Ionienne, Corfou, Grèce. (co-direction avec Jacques Vergne)

LECLUZE, Charlotte. 2007. *Méthode d'alignement sémantique multilingue appliquée à une collection de multidocuments : Un apport aux systèmes d'aide à la traduction*, Mémoire de Master conjoint franco-hellénique mention Sciences du langage, spécialité Sciences de la Traduction : Traductologie et Sciences cognitives, Université de Caen Basse-Normandie, Caen, France et Université Ionienne, Corfou, Grèce. (co-direction avec Jacques Vergne)

TRICHAKI, Marina. 2007. *Étude sur l'apport des cognats à l'alignement des textes*, Mémoire de Master conjoint franco-hellénique mention Sciences du langage, spécialité Sciences de la Traduction : Traductologie et Sciences cognitives, Université de Caen Basse-Normandie, Caen, France et Université Ionienne, Corfou, Grèce. (co-direction avec Jacques Vergne)

Production logicielle

Les logiciels que je présente dans ce chapitre ont été créés soit dans une perspective de mise à l'épreuve de concepts, soit dans une perspective de valorisation de mes travaux. Ils concernent principalement la structuration automatique d'articles et de livres (Résurgence), l'analyse automatique de forums de discussion (la plate-forme Calico et les composants suffixés par agora), l'analyse syntaxique du français et l'analyse morphologique endogène (la plate-forme Wims et les composants préfixés par wims).

2009 – 2011 : **Résurgence4Inex** (Emmanuel Giguet, Nadine Lucas, Alexandre Baudrillart) cherche à construire automatiquement le sommaire de livres scannés libres de droit (300 à 1000 pages). Il s'agit de trouver le découpage hiérarchique des livres en parties, chapitres, sections et sous-sections, en associant titre et numéro de page. Ce logiciel est une branche autonome, dérivée du logiciel Résurgence.

2007 – 2008 : **Résurgence** (Emmanuel Giguet, Nadine Lucas, Catalina Chircu) porte sur la structuration automatique d'articles scientifiques et journalistiques, quelle que soit leur langue. Résurgence cherche à calculer la structure interne des articles en terme de titre, sections, sous-sections, annexes, et à recomposer et associer les objets figuratifs (illustrations, figures, tableaux).

2007 – 2009 : La **plateforme Calico** (Emmanuel Giguet & Pierre Lecavelier) a été créée pour permettre le partage de forums de discussion et d'analyses de forums de discussion au sein du projet Calico. J'ai écrit les spécifications de la plateforme en août 2007, recruté un ingénieur et encadré les développements informatiques. La plateforme est opérationnelle début décembre 2007. J'ai créé la version anglaise et j'ai ouvert la plateforme aux chercheurs extérieurs au projet Calico en mai 2009.

2008 : **Volagora** (Emmanuel Giguet) : Volagora permet à l'utilisateur d'observer les fenêtres temporelles d'activité dans un forum de discussion. Dans un système classique les journées sont supposées commencer à 0h et l'année débuter en janvier. Volagora permet à l'utilisateur de valider ou d'invalider ces présupposés et de caler de manière adéquate le système horaire du forum. Volagora a été conçu en 2008 et mis en ligne en 2009.

2008 : **Colagora** (Emmanuel Giguet, Pierre Lecavelier & Pierre Beust) permet de définir interactivement des thématiques d'intérêt, à partir de mots-clés, et d'observer comment ces thématiques se déploient dans un forum de discussion. Le logiciel est inspiré de Memlabor et Lucia, deux logiciels antérieurs au projet créés au GREYC par Vincent Perlerin. J'ai défini les spécifications du logiciel avec Pierre Beust en juin 2008. Mi juillet, une première version est opérationnelle. Sa mise en ligne date de fin août 2008.

2008 : **Rolagora** (Emmanuel Giguet & Nadine Lucas) a pour ambition d'attribuer des « profils sociaux », appelés « rôles », aux participants d'un forum de discussion, à partir de l'étude des indicateurs fournis par le logiciel Authagora. Cet outil a été conçu fin juin 2008.

2008 : **Authagora** (Emmanuel Giguet) permet d'observer l'activité des participants à un forum de discussion à l'aide de quatre indicateurs : le nombre de message envoyés, le nombre de conversations initiées, le nombre de conversations initiées restées sans suite, le nombre de contributions aux différentes conversations. Ces indicateurs permettent à l'observateur de dégager des profils de participants. Authagora est intégré à Anagora jusqu'en juillet 2008, date à laquelle il devient un module autonome.

2008 : **Showforum** (Emmanuel Giguet & Pierre Lecavelier) permet d'afficher des messages de forums de discussion en liste chronologique, ou en conversation. Les messages peuvent être filtrés par date, auteur, et conversation. L'identité des auteurs des messages peut être anonymisé par une fonction que j'ai créée. Showforum est opérationnel depuis juillet 2008.

2007 : Le **site internet d'analyse de forums** de discussion (Emmanuel Giguet) est créé en mars 2007. Les outils d'analyse du projet Calico, à savoir Bobinette, Anagora et ThemAgora, y sont accessibles. Pour cela, j'ai réécrit le logiciel Bobinette à partir du code source original, transmis par Benjamin Huyn Kim Bang. Les forums peuvent être analysés à la volée sur le site mais pas stockés.

2006 : **Anagora** (Emmanuel Giguet & Nadine Lucas) utilise la polyphonie des forums de discussion (i.e., le fait qu'il y ait plusieurs conversations en même temps) pour proposer une lecture synthétique sous forme de chronogrammes. Anagora a été créé en avril et mai 2006. Il apparaît dans l'article de la revue Sticef vol. 13 en 2007 et fait l'objet d'un article lors de la conférence CSCL à Rhodes, en 2009.

2005 : **ThemAgora** (Nadine Lucas & Emmanuel Giguet) utilise le concept de progression thématique pour produire une représentation compacte, et donc plus rapide à lire, d'un forum de discussion, quelle que soit sa langue. ThemAgora a été créé entre janvier et juin 2005 à partir de la version UniThem du logiciel Thema. Il a été présenté lors de la conférence ICALT à Santander en Espagne, en 2008.

2005 : **WimsAligner** (Emmanuel Giguet) est un logiciel d'alignement sous-phrastique multilingue. Il permet d'extraire des « équivalents traductionnels » d'une collection de documents qui ont été préalablement traduits. Ces équivalents traductionnels correspondent à des traductions de mots, de termes, d'expressions... Ce composant sert par exemple à construire automatiquement des lexiques bilingues.

2004 : **UniThem** (Nadine Lucas & Emmanuel Giguet) utilise le concept de progression thématique pour produire une représentation compacte, et donc plus rapide à lire, d'un article journalistique, scientifique, ou de vulgarisation, et ce, quelle que soit sa langue. Le logiciel a été crée entre septembre et décembre 2004 à partir du logiciel Thema, et présenté lors de la conférence CIDE à Beyrouth au Liban, en 2005.

2001 – 2004 : La **plateforme Wims** (Emmanuel Giguet) est une plateforme d'expérimentations qui permet de faciliter la mise au point, la maintenance et la longévité de mes composants de traitements des langues. Une fois un composant intégré, il peut être testé aisément sur différents corpus gérés par la plateforme, et sur tout ou partie d'un corpus. Les résultats d'analyse sont automatiquement archivés pour permettre le suivi de l'évolution du composant. La plateforme Wims a été conçue entre 2001 et 2004 sur les concepts que j'ai mis au point chez Startem, lors de la création d'une plateforme d'intelligence économique.

2001 – 2004 : **WimsParser** (Emmanuel Giguet) est un composant générique d'analyse syntaxique automatique multilingue. Des ressources ont été mises au point principalement pour le français. Des ressources moins abouties existent également pour l'anglais, l'espagnol, l'italien, le roumain, le portugais, le grec, et l'arabe.

2001 – 2004 : **WimsFreq** / **WimsMorph** / **WimsAgreement** / **WimsBehavior** / **WimsNGram** (Emmanuel Giguet) forment une suite de composants de traitements des langues permettant d'observer certaines caractéristiques linguistiques d'un corpus. WimsFreq présente les mots par fréquence décroissante, WimsMorph extrait des affixes probables, WimsAgreement extrait les suites de candidats affixes, WimsBehavior met en évidence les mots trouvés avant et après une ponctuation, WimsNGram liste les séquences de n caractères par fréquence décroissante.