21es Journées Francophones d'Ingénierie des Connaissances

IC 2010

ICMMX

INGENIUM COGNITIONIS

IC 2010

Actes des 21es Journées Francophones
d'Ingénierie des Connaissances

Présidente du comité de programme

Sylvie Després,
LIM&BIO, Bobigny, France

Président du comité d'organisation

Michel Crampes,
Ecole des Mines d'Alès, Nîmes, France

9 – 11 juin 2010
Ecole des Mines d'Alès, Site de Nîmes, France.

Comité de programme

Présidente : Sylvie DESPRES – LIM&BIO – Bobigny

Membres :

Yamine AIT AMEUR – ENSMA – Poitiers – France
Patrick ALBERT – ILOG – Paris – France
Florence AMARDEILH – MONDECA – Paris – France
Mathieu D'AQUIN – Open University of Milton Keynes – Milton Keynes – UK
Marie-Aude AUFAURE – Laboratoire MAS Ecole Centrale Paris – Paris – France
Catherine BARRY-GREBOVAL – MIS Amiens – Amiens – France
Jean-Paul BARTHES – UTC Compiègne – Compiègne – France
Jean-François BOUJUT – G-SOP INP Grenoble – Grenoble – France
Christian BRASSAC – Codisant Nancy – Nancy – France
Bertrand BRAUNSCHWEIG – ANR Paris – Paris – France
Jean-Pierre CAHIER – Tech-CICO UTT Troyes – Troyes – France
Sylvie CALABRETTO – LIRIS Lyon – Lyon – France
Pierre Antoine CHAMPIN – LIRIS Lyon – Lyon – France
Olivier CORBY – INRIA – Sophia Antipolis – France
Amélie CORDIER – LIRIS – Lyon – France
Michel CRAMPES – EMA – Alès – France
Patrick DROUIN – OLST Montréal – Montréal – Canada
Catherine FARON – I3S – Nice – France
Béatrice FUCHS – LIRIS – Lyon – France
Frédéric FURST – MIS Amiens – Amiens – France
Faïez GARGOURI – ISIM Sfax – Sfax – Tunisie
Alain GIBOIN – INRIA – Sophia Antipolis – France
Monique GRANDBASTIEN – LORIA – Nancy – France
Mounira HARZALLAH – LINA – Nantes – France
Olivier HAEMMERLE – IRIT – Toulouse – France
Ivan HERMAN – CWI – Amsterdam – Pays Bas
Nathalie HERNANDEZ – IRIT – Toulouse – France
Antoine ISAAC – Vrije Universiteit – Amsterdam – France
Pascale KUNTZ-COSPEREC – LINA – Nantes – France
Luc LAMONTAGNE – CERMID – Québec – Canada
Florence LE BER – ENGEES – Strasbourg – France
Michel LECLERE – LIRMM – Montpellier – France
Alain LEGER – Oranges Labs LIRIS – Lyon – France
Jean LIEBER – LORIA – Nancy – France
Lo MOUSSA – Université Gaston Berger – Saint Louis du Sénégal – Sénégal
Alain MILLE – LIRIS – Lyon – France
Jérôme NOBECOURT – LIM&BIO – Bobigny – France
Alexandre PASSANT – DERI Université of Ireland – Galway – Irlande
Sylvie RANWEZ – EMA – Nîmes – France

Pascal SALEMBIER – Tech-CICO UTT Troyes – Troyes – France
Nathalie SOUF – CERIM – Lille – France
Eddie SOULIER – Tech-CICO UTT Troyes – Troyes – France
Yannick TOUSSAINT – LORIA – Nancy – France
Raphaël TRONCY – EUROCOM – Sophia Antipolis – France
Brigitte TROUSSE – INRIA – Sophia Antipolis – France
William TURNER – LIMSI Orsay – Paris – France
Bernard VATANT – MONDECA – Paris – France

Comité de pilotage

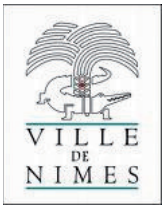
Nathalie AUSSENAC–GILLES – IRIT, CNRS – Toulouse
Bruno BACHIMONT – INA, Paris et UTC – Compiègne
Jean CHARLET – AP-HP et INSERM – Paris
Sylvie DESPRES – LIM&BIO – Bobigny
Fabien GANDON – INRIA – Sophia Antipolis
Marie-Christine JAULENT – SPIM, INSERM – Paris
Gilles KASSEL – MIS, Université de Picardie Jules Verne – Amiens
Philippe LAUBLET – LaLIC, Université de Paris-Sorbonne – Paris
Myriam LEWKOWICZ – Tech-CICO, UTT – Troyes
Nadda MATTA – Tech-CICO, UTT – Troyes
Amedeo NAPOLI – LORIA – Nancy
Yannick PRIE – LIRIS – Lyon
Chantal REYNAUD – LRI, Université Paris 11 & INRIA Saclay – Ile-de-France
Sylvie SZULMAN – LIPN, Université Paris 13 – Paris
Pierre TCHOUNIKINE – LIG – Grenoble
Régine TEULIER – CRG – Paris
Francky TRICHET – LINA – Nantes
Manuel ZACKLAD – Tech-CICO, UTT – Troyes

Comité d'organisation

Ecole des mines d'Alès – Nîmes – France

Michel CRAMPES
Vincent DEROZIER
Gérard DRAY
Gwenaëlle LOMPRE
Michel PLANTIE
Sylvie RANWEZ
Valérie ROMAN
Elisabeth SANSOT

Avec le soutien scientifique et financier de



Avant-propos

Organisées chaque année, les 21es Journées Francophones d'Ingénierie des Connaissances se tiendront à Nîmes du 8 au 11 juin 2010. Placées sous l'égide du GRACQ (Groupe de Recherche en Acquisition des Connaissances), elles constitueront cette année encore un lieu d'échanges et de réflexions de la communauté francophone sur les problématiques de l'ingénierie des connaissances.

L'essor des Sciences et Technologies de l'Information et de la Communication, notamment des technologies du Web, dans l'ensemble de la société engendre des mutations dans les pratiques individuelles et collectives. L'ingénierie des connaissances contribue à cette évolution, en offrant des méthodes pour l'élaboration de modèles de connaissances permettant de rendre ces nouvelles pratiques opérationnelles. Ces modèles relèvent de différentes catégories. Ils décrivent par exemple les connaissances métiers, les processus cognitifs associés au raisonnement et les processus collaboratifs. Des méthodes et des systèmes sont également conçus pour la gestion des connaissances au sens large (acquisition, structuration, représentation, exploitation, exploration, visualisation et diffusion).

L'édition 2010 de la conférence IC s'inscrit dans ces thématiques. Les 25 communications longues sélectionnées parmi 48 soumissions (soit un taux de sélection de 52%) couvrent les thèmes suivants :

- les problématiques de la conceptualisation d'ontologies à partir de textes, l'élaboration de méthodologies de construction et de mise au point d'ontologies ;
- l'exploitation les données en particulier dans le contexte du web des données ; la conception d'outils adaptatifs d'aide à l'utilisateur ;
- la modélisation des pratiques et l'exploitation des traces informatiques d'usages ; l'informatique médicale et ses particularités ;
- les systèmes de représentation des connaissances ;
- l'indexation et la recherche d'information fondées sur des ontologies et de la fouille visuelle.

Près d'un tiers d'entre elles concernent des communications appliquées. L'informatique médicale y est fortement représentée.

Les communications affichées (13 posters) sont rassemblées dans un recueil distinct. Elles constituent des travaux dont la maturité n'est pas suffisante pour une présentation longue mais exposent des problématiques qui devraient susciter des discussions au cours des sessions qui leur sont réservées.

Avant de souhaiter aux participants, une conférence fertile en idées et en discussions, remercions les personnes sans lesquelles ces journées n'auraient pas lieu : les auteurs pour leur contribution, François Bourdoncle, conférencier invité de ces 21es journées,

les membres du comité de programme pour leurs relectures constructives et les organismes soutenant cette manifestation. En outre, comme chaque année l'AFIA décernera le prix de la meilleure communication.
Que soit également remercié le comité d'organisation dont l'enthousiasme, le dynamisme et l'efficacité ont permis une planification parfaite de cet événement. Enfin, remercions les membres du bureau du GRACQ et en particulier Nathalie Aussenac-Gilles et Jean Charlet qui contribuent chaque année à l'organisation de ces journées.

Sylvie DESPRES
Présidente du comité de programme

Conférence invitée

L'Intelligence Collective d'Usage

François BOURDONCLE, co-fondateur de la société Exalead
http://bourdoncle.net/

Alors même que la production de contenus ne cesse de croître sur Internet comme dans l'entreprise, on constate que ces données ne sont pas aujourd'hui réellement exploitables pour faire émerger une "intelligence collective" qui est pourtant l'un des grands enjeux de l'ère numérique. Les technologies du Web Sémantique sont très prometteuses, mais reposent sur quelques présupposés discutables, comme la faisabilité à grande échelle d'une approche "normative" ou encore l'hypothèse implicite que les producteurs de contenus soient nécessairement les mieux à même de formaliser la connaissance qu'ils produisent. Cet exposé discute d'une approche alternative, de type "émergente", qui se base sur un principe très différent selon lequel c'est l'usage particulier qui est fait du savoir collectif qui est le critère déterminant pour la structuration de ce savoir. Cette approche, qui est notamment celle d'une société comme *Exalead*, se base sur des techniques d'apprentissage automatique et fonde sa pertinence sur le fait que la quantité (et la qualité) des corpus disponibles pour ce faire est aujourd'hui suffisante pour atteindre un niveau de qualité égal ou supérieur aux approches normatives.

Session 1. Ingénierie des Connaissances et Textes

La session 1 comporte trois papiers centrés sur les approches textuelles. La première contribution est une réflexion sur la conceptualisation à partir de textes. La deuxième contribution s'inscrit dans une approche statistique appliquée au titrage automatique de textes. La troisième propose un repérage automatique de structures énumératives à partir d'éléments de mise en forme.

Des textes au concept
Propositions pour une approche textuelle de la conceptualisation

Mathieu Valette

ATILF, CNRS, Nancy-Université
mvalette@atilf.fr

Résumé : Il s'agit d'esquisser les conditions théoriques d'une approche textuelle de la conceptualisation. Nous souhaitons illustrer l'hypothèse qu'avant d'accéder au statut de signes dont les signifiés sont normés (*i.e.* les termes), les concepts émergents se manifestent dans les textes sous des formes hétérogènes telles que des groupements récurrents de traits sémantiques qui se stabiliseront – ou non – en unités lexicales nouvelles. Ces coalitions ont valeur de préconcepts ou de *protosémies*. L'enjeu est de décrire et de modéliser ce processus d'émergence pour lui donner, à terme, une place dans une théorie de la terminologie. Au plan pratique, l'objectif est de fournir à moyen termes des outils d'identification et de détection pour la veille et la constitution de terminologies.

Mots-clés : Linguistique, Sémantique textuelle, Formes sémantiques, Protosémie, Conceptualisation.

1 La fouille de données : métaphore minière ou archéologique ?

La terminologie textuelle (Slodzian 2000) s'est consacrée à l'extraction de candidats termes dans les textes pour les expertiser et le cas échéant les valider comme concepts termes. Elle a promu l'idée que les textes sont les lieux de production des termes mais ni les conditions de cette production ni l'éventualité d'une conceptualisation liée aux textes eux-mêmes ne semblent encore avoir retenu l'attention. C'est l'hypothèse générale que nous souhaitons défendre et illustrer, à contre-courant du postulat platonicien dont l'ingénierie des connaissances a hérité, selon lequel le concept préexiste au terme qui le désigne. Il n'est toutefois pas question d'adopter ici une posture logicienne et de considérer que la conceptualisation est un phénomène purement linguistique répondant à des règles de construction propres à une fonction du langage, ni de faire abstraction des conditions psychiques, sociales et interactionnelles de l'élaboration des concepts. Il s'agit d'une part de considérer que la textualité, c'est-à-dire les contraintes propres à la mise en texte, à la formulation linguistique, qu'il s'agisse de contraintes grammaticales, sémantiques, lexicologiques ou liées aux *traditions discursives* (parmi lesquelles les genres et les discours) (Koch 1997) joue un rôle clé dans la formation des concepts, et d'autre part

d'appréhender les textes non plus seulement comme des ressources, des carrières qu'il convient de forer pour en extraire la matière première terminologique ou conceptuelle, mais comme des *archives*, c'est-à-dire la trace objective du processus de création des concepts.

C'est donc dans une lignée davantage foucaldienne et archéologique que minière que nous souhaitons nous inscrire. Pour Foucault, il s'agit notamment d'identifier des *formations discursives*, c'est-à-dire des régularités entre objets, types d'énonciation, concepts et choix thématiques à l'origine de la production idéologique ou scientifique, autrement dit à l'élaboration de connaissances partagées. De là résulte l'importance d'en déterminer les règles de formation (Foucault 1969, 53). De fait, l'activité cognitive qui consiste à créer des concepts, les modifier, les ordonner et les articuler entre eux, se confond souvent avec l'activité de production de texte. On pourrait détourner pour la faire nôtre une proposition de G. Bachelard (1938, 61) et considérer qu'un concept est un groupement d'« approximations successives » – lesquelles sont, selon nous, observables dans des corpus diachroniques.

Nous souhaitons donc ici ouvrir un débat destiné à illustrer l'hypothèse qu'avant d'accéder au statut de signes dont les signifiés sont normés (*i.e.* les termes), les concepts émergents se manifestent dans les textes sous des formes hétérogènes, approximatives, telles que des groupements récurrents de propriétés sémantiques qui se stabiliseront – ou non – en unités lexicales nouvelles. Ces groupements ont valeur de préconcepts ou de *protosémies*, c'est-à-dire d'unités sémantiques antérieures à toute lexicalisation. L'enjeu est de décrire et de modéliser ce processus d'émergence pour lui donner une place dans une théorie de la terminologie. Au plan pratique, l'objectif sera de construire, à terme pour l'heure non échu, des outils d'identification et de détection pour la veille et la constitution de terminologies.

2 Signifiés et concepts, mots et termes

Le mot est un concept linguistique fragile. A la fois imprécis de par ses frontières théoriques et matérielles, et ethnocentrique parce que les langues sans mot sapent tout espoir d'en faire un concept universel, il demeure néanmoins un mode d'aperception du langage parmi les plus intuitifs. Très étudié par les linguistes, il est aussi très utilisé en ingénierie des connaissances. Une approche textuelle du mot pourrait, comme le suggère (Rastier 2001, 182-183) à propos d'une reconception possible du signe, s'inspirer d'un texte où Saussure écrit :

> « […] vous n'avez plus le droit de diviser, et d'admettre d'un côté le mot, de l'autre sa signification. Cela fait tout un. Vous pouvez seulement constater le kénôme et le sème associatif » (Saussure 2002, 93)

Le signe (*sème* signifiant *signe* pour Saussure) est donc « contextuellement défini » selon Rastier et peut-être vu comme un passage vide entre deux contextes, gauche et droite, autrement dit, le signe n'a de valeurs que celles que lui octroient le contexte. La radicalité de cette conception du signe nous agrée, tant elle semble aux

antipodes de l'approche terminologique traditionnelle. En effet, si la lexicologie a recours en maints lieux théoriques et pratiques au contexte pour définir le contenu sémantique du mot (par la collocation notamment)[1], la terminologie quant à elle, rapatrie nécessairement le sens dans le terme. Dans ce cadre général, nous présentons un ensemble de propositions adossé à la sémantique textuelle visant à situer l'étude du lexique dans le paradigme textuel. Plus précisément, notre projet est d'étudier les déterminations textuelles de la création et de la lexicalisation des concepts.

3 Forme sémantique et concept

Adoptons un empirisme de méthode : les mots et les termes, qui en sont un cas particulier, apparaissent dans deux types d'objet matériel ; le texte, objet construit de façon syntagmatique, où ils sont actualisés dans un état qu'on pourrait qualifier de dynamique et la ressource terminologique ou lexicale (thésaurus, lexique, dictionnaire, etc.), objet construit de façon paradigmatique, où ils sont dans un état passif, en attente d'une actualisation. Qu'est-ce qu'un mot, par-delà ces types d'objet ? Il est un signe constitué d'une forme et d'un contenu. La forme est acoustique ou graphique, c'est le signifiant, le contenu est, dans le paradigme structuraliste, une collection de propriétés sémantiques articulées entre elles, qui constituent le signifié. Ces propriétés sont des *sèmes*. Elles sont d'ordre métalinguistique et résultent concrètement d'une analyse ou d'une validation humaine effectuées par un expert. On peut distinguer deux types de collections de sèmes, selon que l'on se trouve dans une problématique du texte ou dans une problématique du dictionnaire.

3.1 Les signifiés comme concept ?

Composé d'un signifié et d'un signifiant, nous représenterons un mot de la façon suivante :

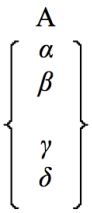


Fig. 1 – Un signifié composé des sèmes α, β, γ, δ associé à un signifiant A

A désigne le signifiant (c'est-à-dire, en pratique, le mot graphique) et α, β, γ, δ entre accolades désignent le signifié proprement dit (composé des sèmes regroupés).

1 Lire (Blumenthal & Hausmann, éd., 2006).

Tous ces traits n'ont pas nécessairement la même qualité ni la même valeur sémantique, ce que figure le saut de ligne entre β et γ. Certains sèmes, par exemple, sont génériques d'une classe donnée. Ainsi, le signifié du mot *« chien »* peut être décrit comme la collection de sèmes présentée dans la partie gauche de la figure 2. Dans la classe sémantique des *Canidés*, le sème /canidé/ est générique, et /domestique/ est un sème spécifique de *« chien »* qui le différencie du renard ou du chacal par exemple. Si l'on construit une classe sémantique des *Animaux de compagnie*, /domestique/ sera, à l'inverse, un sème générique tandis que /canidé/ sera spécifique de *« chien »*, puisque ni le chat, ni le poisson rouge ne sont des canidés. Il est à noter que l'hétérogénéité zoologique de la classe des *Animaux de compagnie* ne préjuge pas de sa cohérence : lorsqu'il s'agit de choisir un animal de compagnie, la sélection s'effectue sur des propriétés propres à la classe (l'encombrement, l'autonomie, etc.) et non sur des critères zoologique par exemple. En cela, ces classes sémantiques, qu'on appelle des *taxèmes*, ne sont pas assimilables à des catégories ontologiques classiques. Un taxème, en effet, est une petite classe sémantique correspondant à une situation d'usage précise et n'a vocation ni à l'universalité, ni à l'intangibilité. Un domaine, dans cette perspective théorique, est composé d'un ensemble de taxèmes correspondant à une pratique déterminée.

Dans la mesure où les sèmes sont des constructions, la liste n'en est ni exhaustive, ni fermée. Elle ne relève pas d'une quelconque valeur de vérité mais d'une *valeur d'usage*. Ainsi, le contenu sémantique de *« chien »* construit par un informateur âgé de quatre ans peut fort bien correspondre au signifié de droite, toujours sur la figure 2.

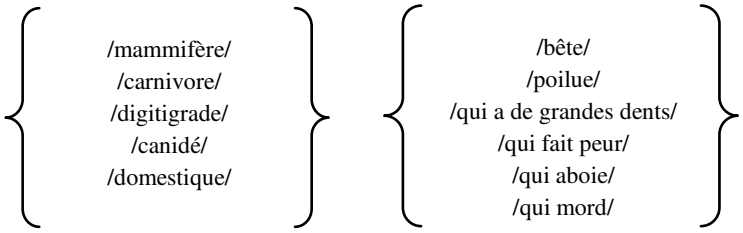


Fig. 2 – Deux signifiés possibles pour *« chien »*.

Aucun de ces signifiés, qui sous-tendraient deux définitions lexicographiques parfaitement distinctes, n'est meilleur qu'un autre. Les propriétés sémantiques diffèrent par leur seul contexte de construction. Le premier est savant, le second se fonde sur une ou plusieurs expériences sensibles ou fantasmées. Dès lors, dans une perspective pratique, on peut se poser la question de la permanence du concept de *« chien »*.

3.2 Les réseaux sémiques

Un texte est, formellement parlant, un alignement de signes (de mots) suivant des règles de construction syntaxiques. La mise en relation dans un texte de plusieurs signifiés donne lieu à de nouveaux regroupements de sèmes, syntagmatiques cette fois-ci, c'est-à-dire des groupements entre sèmes appartenant à des signifiés

différents. Ces groupements syntagmatiques sont occurrentiels car spécifiques à un texte, ou à un ensemble de textes. L'interprétation d'un texte repose sur la reconnaissance et l'identification de ces groupements. On distingue deux types de groupements syntagmatiques, les fonds et les formes sémantiques.

3.2.1. Les fonds sémantiques (isotopies)

Lorsqu'un sème donné se retrouve à plusieurs endroits dans un même texte, il peut s'agir d'une isotopie. Les isotopies se structurent en réseaux et constituent le fond sémantique. Dans le texte qui suit, on observe une isotopie simple, par la récurrence du sème /tabac/ :

> Le plus difficile dans l'arrêt du tabac$^{/tabac/}$, c'est la décision. Et cette décision doit venir d'une certitude que la vie sans tabac$^{/tabac/}$ existe. J'étais un très gros fumeur$^{/tabac/}$. M'imaginer sans cigarette$^{/tabac/}$ était un cauchemar[2].

Minimalement, cela signifie que ce texte traite de tabac. Mais les isotopies peuvent relever d'une description plus fine. Par exemple, elles peuvent être génériques, c'est-à-dire correspondre aux sèmes structurants de classes sémantiques telles que les domaines et les taxèmes. L'isotopie constitue le socle du parcours interprétatif, notamment par les échanges d'informations sémiques qui s'opèrent entre elle et les formes sémantiques. On en donnera une illustration dans le paragraphe 4.3.

3.2.2. Les formes sémantiques

Plusieurs sèmes distincts peuvent être instanciés ensemble dans des textes différents avec une certaine régularité. Ces groupements s'appellent des thèmes, ou des *formes sémantiques*. Nous les représenterons ainsi :

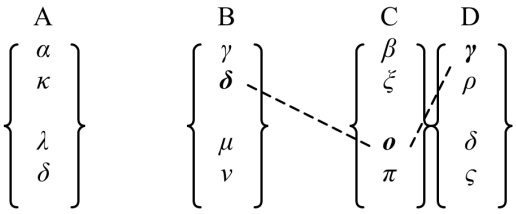


Fig. 3 – Une forme sémantique :
groupement stabilisé des sèmes δ, o et γ dans différents textes

Par exemple dans les textes suivants :

> Le citoyen est libre$^{/liberté/}$ de fumer$^{/fumer/}$ ou de ne pas fumer, de manger de la salade si ça lui chante et des rillettes s'il en a envie[3]

[2] http://sante-medecine.commentcamarche.net/
[3] http://www.le-tigre.net/

> Je suis contre l'interdiction de fumer/liberté/, chacun est libre/liberté/ de son choix/liberté/4

Les sèmes /liberté/ et /fumer/ sont en cooccurrence rapprochée. Ils constituent une petite forme sémantique, c'est-à-dire une unité sémantique dont la lexicalisation n'est ni stable ni synthétique, autrement dit, dont le signifiant est discontinu et variable. Cette unité sémantique correspond, par exemple, à l'unité lexicale *« liberté de fumer »* que l'on peut considérer comme un début de lexicalisation :

> Ils ont la liberté de fumer{/liberté//fumer/}, soit, je serai le dernier à la leur retirer, sauf dans les endroits publics[5].

Les formes sémantiques ne sont pas des périphrases, elles importent en premier lieu pour les modalités qu'elle opèrent sur les signifiés des unités lexicales cooccurrentes. La forme sémantique {/liberté//fumer/} s'est particulièrement développée et tend à se stabiliser depuis la mise en application de la loi dite antitabac en janvier 2008. Elle constitue notamment un des arguments privilégiés de ses détracteurs (ce serait une loi liberticide). Parmi eux, les industriels du tabac ont beau jeu de l'exploiter, mais ils ont soin de restreindre cette liberté de fumer aux seuls adultes, la loi leur interdisant de faire la promotion du tabac auprès des enfants. Or, par un habile jeu rhétorique, leur discours peut fort bien avoir l'effet contraire, c'est-à-dire promouvoir le tabac chez les jeunes. Etudions ces quelques extraits de la documentation présente sur leurs sites Web :

> Notre métier ne consiste pas à inciter des gens à fumer/fumer/. Il consiste à offrir des marques de qualité à des adultes qui ont déjà pris la décision/liberté/ de fumer/fumer/ […]. C'est pourquoi nous sommes convaincus que fumer/fumer/ devrait être le seul fait d'adultes conscients des risques de fumer/fumer/. (JTI) [6]

> Fumer/fumer/ repose sur une décision/liberté/ individuelle qui ne peut être que le fait d'adultes informés des risques liés au tabagisme. (ALTADIS)

> Nous sommes convaincus que le choix/liberté/ de fumer/fumer/ doit être le choix d'adultes avertis et conscients des risques, un choix qui exclut de fait les jeunes, non adultes. (BAT)

> Nous nous engageons ainsi à communiquer de manière responsable avec les adultes qui ont délibérément choisi/liberté/ de fumer/fumer/. (ALTADIS)

> JTI s'engage à fabriquer des cigarettes/fumer/ de qualité pour les adultes qui choisissent/liberté/ de fumer/fumer/ par plaisir. (JTI)

Les industriels du tabac associent à la forme sémantique {/liberté//fumer/} une autre forme dont le pivot est *« adultes »*. Mais dans ces brefs énoncés, *« adultes »* subit un certain nombre de modalités valorisantes, le plus souvent sous la forme d'adjectifs qualitatifs *(« conscients »*, *« informés »*, *« avertis »*, etc.) de sorte que se construit un

[4] http://www.agoravox.fr/

[5] http://www.philo5.com/

[6] Ces exemples et les analyses sont issus des recherches collectives de l'équipe linguistique (ERTIM-INaLCO, Paris) du projet C-MANTIC (ANR-07-MDCO-002).

parcours interprétatif liberté de fumer + adulte$^{/valorisant/}$. Ce discours de valorisation de l'adulte libre de fumer, autrement dit, de l'adulte fumeur, rend possible une lecture spéculaire telle que le jeune, non adulte, est dévalorisé. C'est explicite dans le troisième exemple (« un choix qui exclut de fait les jeunes, non adultes »). On peut donc construire les deux formes sémantiques ci-dessous :

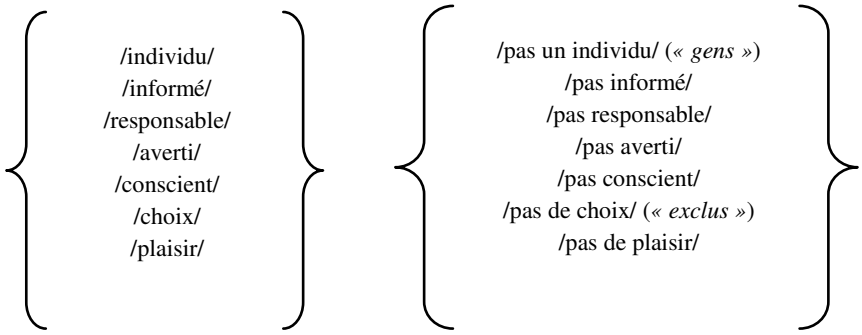


Fig. 4 – Formes sémantiques possibles d'*« adulte »* et de *« non adulte »* dans les textes de l'industrie du tabac

La production de nouveaux sèmes pour le concept existant *« adulte »* participe au succès du discours promouvant le tabac chez les jeunes et les échecs répétés de la lutte contre le tabac : les messages de l'industrie du tabac sont massivement positifs et euphoriques quand ceux des organismes de prévention sont négatifs, dysphoriques[7].

4 Discussion : ajuster les signifiés et les formes sémantiques

4.1 Les déterminations sémantiques du texte

Le sens d'une unité lexicale dépend autant du contexte dans lequel il apparaît que de sa définition première. En complément de l'opposition sèmes génériques *vs* sèmes spécifiques (qui permet d'introduire la notion de classe sémantique), la sémantique interprétative opère une distinction entre des sèmes inhérents et des sèmes afférents. Les sèmes inhérents peuvent être considérés comme plus « définitoires » (par exemple, pour *« chien »*, on aura : /canidé/, /quadrupède/, etc.) que les sèmes afférents qui relèvent de l'usage qui est fait du mot dans les textes. Les sèmes afférents sont donc issus des contextes d'énonciation. C'est le cas dans l'exemple de la figure 4 où le signifié d'*adulte* n'est représenté ici qu'avec des sèmes afférents. Il s'agit de sèmes

[7] C'est du moins un des résultats d'analyse de l'équipe du projet C-MANTIC auquel nous empruntons ce matériau.

hérités d'un ensemble de contextes et susceptibles d'être recontextualisés à l'identique.

De la même manière, le sème /pauvre/ est fréquemment actualisé dans *« population »* lorsqu'il est cooccurrent du lexème *« banlieue »*. Mais il ne s'agit pas d'un sème inhérent : rien dans la banlieue ne la prédispose à accueillir de façon privilégiée une population pauvre. De même, l'expression *« jeunes des banlieues »*, dans le discours journalistique (qui est aujourd'hui largement prescriptif) ou dans le discours politique (qui lui ressemble) ne signifie pas tous les jeunes résidant en banlieue, mais certains jeunes, défavorisés, résidant en banlieue, et même, fréquemment, des jeunes gens *« issus de l'immigration »*. Lorsque le quotidien *Le Monde* titre le 30 septembre 2001 : « Les jeunes des banlieues craignent l'amalgame entre musulmans et terroristes », il enrichit implicitement le signifié de l'unité « jeune de banlieue » d'un sème /musulman/. Ces sèmes, /pauvre/ ou /musulman/ sont des sèmes afférents, « subjectifs » ou « socialement normés », c'est-à-dire circonscrits d'un point de vue historique, géographique et socioculturel. Aux Etats-Unis, les pauvres vivent en centre-ville. En revanche, lorsque « banlieue » s'intègre dans certaines lexies composées telles que « *banlieue de l'ouest parisien* », le sème /pauvreté/ est inhibé par le sème /bourgeois/, afférent à *« ouest parisien »*, quand même la banlieue ouest de Paris est tout aussi hétérogène que la banlieue dans son ensemble.

Les sèmes contenus dans un signifié ne sont pas tous égaux. Leur nature et leur qualité varient en synchronie (tous les sèmes n'ont pas la même valeur dans le signifié) comme en diachronie (un même sème peut évoluer dans le temps, disparaître, etc.). Les sèmes oscillent entre stabilité et instabilité. Le signifié est constitué de sèmes résistants à la variation, sinon permanents et de sèmes instables. Cette variabilité résulte de l'enrichissement ou de l'appauvrissement du signifié à mesure que le mot est actualisé dans les textes. D'une certaine façon, chaque actualisation d'un mot l'enrichit de son contexte d'actualisation, la fréquence de sa participation à un groupement transversal modifie son signifié. Nous dirons, pastichant ainsi une formule de Rastier, que tout mot placé dans un texte en reçoit des déterminations sémantiques, et modifie potentiellement le signifié de chacun des mots qui le composent[8].

Ainsi, le sens résulte d'ajustements entre des signifiés et des formes sémantiques. De la même façon que le cortex visuel traite moins d'informations issues du nerf optique que d'information stockée en mémoire, l'interprétation résulte autant – sinon davantage – d'une activité sémique intense (reconfiguration du signifié, ajustement, convocation des afférences possibles en mémoire, etc.) que du texte lui-même[9].

[8] La proposition initiale de (Rastier 2001, 92) est que « Tout texte placé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent ».

[9] (Rastier & Valette 2009) en présentent quelques exemples dans le cadre de l'évolution sémantique d'une unité lexicale existante (ou *néosémie*).

4.2 La forme sémantique comme *protosémie*

L'élaboration d'un signifié subit des contraintes textuelles et intertextuelles. Les contraintes intertextuelles sont notamment liées aux discours et aux genres, les contraintes (intra)textuelles sont liées à l'économie ou l'organisation sémique du texte[10]. Dans ce contexte, et compte tenu de ce que nous avons présenté dans le paragraphe précédent, nous proposons de considérer la forme sémantique comme le signifié potentiel (ou le sémème potentiel) d'un signe sans signifiant synthétique attitré. Soit l'équivalence hypothétique présentée dans la figure 5.

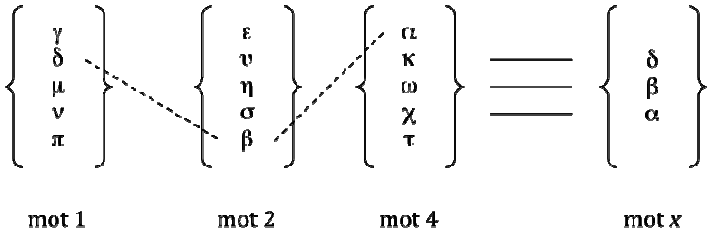


Fig. 5 – La forme sémantique comme protosémie

En effet, signifié et forme sémantique sont tous deux identifiables à un groupement sémique de compacité variable, associé à un signifiant stable et synthétique dans le cas du signifié, discontinu et sans lexicalisation privilégiée dans le cas de la forme sémantique. En bref, nous faisons l'hypothèse que certaines formes sémantiques sont potentiellement des signifiés en devenir, ou les signifiés de proto-signes sans signifiant stabilisé ni synthétique attitré. En d'autres termes, certaines formes sémantiques sont des *protosémies*.

Dans le contexte d'une recherche appliquée, en veille lexicale par exemple, le concept de protosémie a pour incidence la possibilité d'identifier le signifié en cours d'élaboration d'un signe en analysant le complexe sémique dans lequel il s'insère. Pour cela, nous empruntons à (Rastier 2006) les concepts de *diffusion* et *sommation* qui rendent compte des échanges sémiques entre les fonds et les formes sémantiques. La diffusion relève de propagation des sèmes des formes sémantiques vers le fond sémantique. La *sommation* ressortit à la propagation du fond sémantique vers des formes sémantiques. Si le signifié peut être considéré à la manière d'une forme sémantique, il est susceptible de se *diffuser* dans le fond sémantique. Dans ce cas, il nous est possible de restituer une protosémie, c'est-à-dire un signifié-forme sémantique en postulant la *sommation* de ladite forme à partir du fond sémantique que nous identifions par le biais d'*isotopies locales*. C'est ce que nous étudierons dans le paragraphe suivant en relatant une étude récente qui illustre notre propos.

[10] On lira Valette (sous presse) pour un développement.

4.3 Naissance d'un concept

(Reutenauer *et al.* 2009) évaluent les déformations subies par une forme sémantique dont l'élément stable est constitué de l'unité lexicale *« économie réelle »* dans un corpus de presse constitué de 1587 articles, tirés du *Figaro* et de l'*Humanité* entre septembre 2008 et février 2009. Le thème du corpus est la crise économique et financière. Il se présente sous forme de deux versions parallèles : la version lexicale, d'un million d'occurrences de formes, et une image sémique de ce même corpus, composé de 23 millions d'occurrences de candidats-sèmes. A partir d'un calcul de spécificités (implémentation Lexico3, Salem *et al.* 2003), les auteurs mesurent la sensibilité des informations au contexte éditorial des deux quotidiens du corpus. Ainsi, au voisinage d' *« économie réelle », l'Humanité* active les sèmes /bien/ (substantif), /revenu/, /consommation/ et /dépense/, tandis que /ressource/ et /économie/ sont actualisés par *le Figaro*. Les auteurs avancent l'hypothèse aisément corroborable d'une perception plus macroéconomique de l'économie dans *le Figaro* et plus localisante dans *l'Humanité*.

S'appuyant sur une méthodologie similaire, (Reutenauer *et al.* 2010) effectuent une analyse de l'environnement sémique d'un signifié en cours d'élaboration, celui d'*« Outreau »*. Le corpus porte donc sur l'affaire judiciaire dont « Outreau », la ville, est l'éponyme. Divisé en cinq périodes, il est constitué d'articles de presse publiés entre novembre 2001 et avril 2006 comprenant au moins une occurrence du mot étudié. (Lecolle 2007), à laquelle est emprunté le corpus, observe que le sens d'« *Outreau »* évolue du toponyme à « l'erreur judiciaire par excellence ». Comme dans l'étude précédente, le corpus se présente sous deux versions parallèles : la version lexicale de 400 000 occurrences de formes et une image sémique de 10 millions de candidats dont a été extraite une sous-image constituée des seuls candidats correspondant à des formes rendues saillantes par un calcul de spécificités effectué là encore avec Lexico3. Ainsi, les auteurs ont à manipuler une sélection de quelques dizaines de candidats-sèmes seulement.

L'une des analyses effectuées par les auteurs nous intéresse particulièrement. A partir de listes de candidats-sèmes spécifiques à chaque période, des regroupements sémantiques sont réalisés manuellement. Selon nous, ces regroupements peuvent être assimilés à des isotopies. Par exemple, le regroupement de candidats tels que /police/, /procureur/, /écrouer/ s'apparente à une isotopie domaniale //judiciaire// particulièrement présente dans les périodes précoces du corpus. En son sein, on peut observer les traces de différents champs génériques : l'//arrestation// apparaît à travers l'ensemble de candidats-sèmes {/écrouer/, /police/, /arrestation/, /incarcération/, /incarcérer/, /prévenu/}. L'isotopie domaniale //politique// préfigurant le sens d'« erreur judiciaire par excellence », apparaît dans la quatrième période et se renforce ensuite. Inversement, l'isotopie taxémique des //dénominations de crimes// (comportant des candidats tels que /pédophilie/, /meurtre, /viol/) décroît en importance en quatrième période puis disparaît. Les isotopies taxémiques locatives (//lieu d'habitation//, //lieu géographique//) ne sont représentées significativement qu'à

la première période. L'isotopie //fiasco// (constitué de /naufrage/, /drame/, /faillite/, /faute/) est représentative de la cinquième période.

Si nous considérons que les faisceaux d'isotopies locales observables ici sont le résultat d'une diffusion de la forme sémantique d'*« Outreau »* vers le fond, nous pourrions, par *simulation sommatrice* inverse, restituer la forme sémantique, c'est-à-dire le signifié en cours d'élaboration d'*« Outreau »*, lequel est minimalement définissable par un changement de sème générique : le sème domanial //judiciaire// se substitue au sème inhérent toponymique /ville/, avant d'être inhibé pour céder sa place au sème domanial //politique//.

Ainsi, le signe *« Outreau »* qui désignait une ville du Nord, se voit investi, par le contexte, de plusieurs valeurs sémantiques qui le conduisent à devenir un concept, celui de scandale politico-judiciaire.

5 Conclusion

Nous avons souhaité exposer ici à grands traits un ensemble de propositions théoriques destiné à mettre en évidence d'une part l'importance du texte dans la formation des unités lexicales et des concepts correspondants, même dans les situations où l'on préjuge d'une stabilité ontologique *a priori* forte (par exemple le *chien*, l'*adulte,* une entité nommée tel qu'*Outreau*) ; d'autre part les promesses d'une description linguistique et sémantique de ces phénomènes de conceptualisation au prisme des textes. En cela, l'outillage théorique de la sémantique des textes (formes sémantiques, complexe sémique, protosémie, etc.) s'avère, selon nous, pertinent. La mise en œuvre d'une telle méthode permettrait de repérer les variations des concepts dans des textes, en synchronie comme en diachronie, et d'en organiser les différentes facettes.

Toutefois, les études signalées dans cet article relèvent de la linguistique de corpus et ont une visée explicitement exploratoire. Elles ressortissent à une sémantique *enrichie* qui nécessite des temps de traitement et d'analyse des données incompatibles avec des exigences industrielles. Leur finalité est de mieux comprendre les mécanismes nécessaires à l'élaboration, la production et l'interprétation des concepts en situation réelle à partir de textes et d'en tirer partie, à terme, pour des applications ciblées. Il reste évidemment à concevoir les outils et les procédures nécessaires à une opérationnalisation.

Références

BACHELARD, Gaston, (1938) *La formation de l'esprit scientifique*, Paris Vrin.

BLUMENTHAL, Peter, HAUSMANN, Franz Josef, éds. (2006) « Collocations, corpus, dictionnaires », in : *Langue française*, Paris, Larousse.

FOUCAULT, Michel (1969) *L'archéologie du savoir*, Paris, Gallimard.

KOCH, Peter (1997): « Diskurstraditionen: zu ihrem sprachtheoretischen Status und ihrer Dynamik », in Barbara FRANK *et al.* (éd.), *Gattungen mittelalterlicher Schriftlichkeit. Tübingen*, p. 43-79.

LECOLLE, Michelle (2007) « Polysignifiance du toponyme, historicité du sens et interprétation en corpus. Le cas de *Outreau* », Corpus n°6, p. 101-125.

RASTIER, F. (2001) *Arts et sciences du texte*, PUF, Paris.

RASTIER, François (2006) « Formes sémantiques et textualité », in *Langages*, n°163, p. 99-114.

RASTIER, François, VALETTE, Mathieu (2009) « De la polysémie à la néosémie », *Le français moderne*, Salah MEJRI, éd., *La problématique du mot*, n°77, p. 97-116.

REUTENAUER, Coralie, VALETTE, Mathieu, JACQUEY, Evelyne (2009) « De l'annotation sémique globale à l'interprétation locale : environnement et image sémiques d' "économie réelle" dans un corpus sur la crise financière », *Colloque ARCo'09, Interprétation et problématiques du sens*, (9-11 novembre 2009), Rouen.

REUTENAUER, Coralie, LECOLLE, Michelle, JACQUEY, Evelyne, VALETTE, Mathieu (2010), « Sémème au macroscope : genèse et variation sémiques d'une unité lexicale », in : *Actes JADT'2010* (9-11 Juin 2010), Rome.

SALEM, André, LAMALLE, Cédric, MARTINEZ, William, FLEURY, Serge, FRACCHIOLLA, Béatrice, KUNCOVA, André, MAISONDIEU, Aude (2003) « Lexico3 – Outils de statistique textuelle. Manuel d'utilisation. », Syled-CLA2T, Université de la Sorbonne nouvelle – Paris 3.

SAUSSURE, Ferdinand de (2002) *Ecrits de linguistique générale*, Paris, Gallimard.

SLODZIAN, Monique (2000) « L'émergence d'une terminologie textuelle », *Le sens en terminologie*, Ph. THOIRON et H. BEJOINT, éd., Lyon, Presses universitaires de Lyon (Travaux du C.R.T.T.), p. 61-85.

VALETTE, Mathieu (sous presse) « Méthodes pour la veille lexicale », in *Actes de la journée d'étude : "Le dictionnaire électronique. Quelles perspectives pour les sciences humaines et sociales ?"*, Leila Messaoudi (éds.), Publication du laboratoire Langage et société, Université Ibn Tofail Kénitra (disponible sur http://hal.archives-ouvertes.fr/).

Titrage automatique de documents électroniques par extraction de syntagmes nominaux

Cédric Lopez, Violaine Prince, Mathieu Roche

LIRMM, 161 rue Ada, 34392 Montpellier Cedex 5
Lopez@lirmm.fr, Prince@lirmm.fr, Mroche@lirmm.fr

Résumé : Le titrage automatique est un des domaines clé de l'accessibilité des sites WEB tel que défini. Nous proposons dans cet article une approche permettant le titrage automatique de textes (messages de type mails, forum, etc.). À partir de l'étude morpho-syntaxique des titres de notre corpus, nous proposons une approche de titrage automatique. Celle-ci se compose de quatre étapes : l'acquisition du corpus, la détermination des phrases candidates pour le titrage, l'extraction des syntagmes nominaux parmi les phrases candidates et enfin le choix du titre (ChTITRES). Les résultats de l'évaluation par une dizaine d'utilisateurs montrent que les titres déterminés par notre approche sont pertinents.

Mots-clés : Syntagmes nominaux, titrage automatique, statistiques.

1 Introduction

Les titres ont fait l'objet de nombreuses études littéraires et sont vus de différentes manières (Peñalver Vicea, 2003) : « porte qui s'ouvre au lecteur » (Ricardou, 1972), « ensemble de petites unités textuelles » (Frandsen, 1990), « élément le plus important de la plupart des textes » (Furet, 1995), etc. La définition donnée par Le Petit Larousse (2004) est « mot, expression, phrase, etc., servant à désigner un écrit, une de ses parties [...], à en donner le sujet ». Il en résulte que le titre désigne le sujet traité par un groupe de mots bien formé, expression, phrase ou simple mot. Plusieurs groupes de mots bien formés peuvent donc convenir à un titre. Autrement dit, un texte peut avoir plusieurs titres possibles. Il peut varier en fonction de sa taille (en nombre de mots), de sa forme ou bien du sujet mis en avant. Ainsi, le jugement humain sur la qualité d'un titre sera toujours subjectif et plusieurs titres différents pourront être considérés pertinents.

Notons que le titre doit être différencié du résumé, qui est une forme condensée (abrégée, sommaire) d'un texte. Alors que le résumé doit donner un aperçu du contenu du texte, le titre doit désigner le sujet traité dans le texte sans pour autant dévoiler le contenu. Le processus de résumé peut faire appel au titre, par exemple dans (Minel, 2001) où les titres sont systématiquement placés dans le résumé (et apparaissent en gras) ce qui montre l'importance du titre ainsi que la necessité de connaître le titre

pour obtenir un résumé de qualité. Les résumés automatiques fournissent un ensemble de données pertinentes extraites du texte, mais toujours sous forme de phrase(s). Or, un titre n'est que très rarement une phrase. La compression de texte pourrait être intéressante pour le titrage si nous savions compresser fortement un texte pour qu'il n'en résulte que le groupe de mots pertinent. Dans le cas d'une compression classique de texte (par exemple (Yousfi-Monod & Prince, 2006)), un choix parmi les groupes de mots résultant de la compression serait à faire, de manière à ce que le groupe de mots conservé corresponde aux caractéristiques du titre. Encore une fois, le titre (et/ou sous-titre) peut être un élément d'appui à la compression de texte.

De même, le titre doit être différencié de l'index car ce premier ne contient pas toujours les termes clés du texte. Effectivement, le titre peut présenter une reformulation partielle ou totale du texte, ce qui n'est pas envisageable pour un index. Le rôle de l'index est de permettre une recherche facilitée pour l'utilisateur. Encore une fois, la construction d'index peut se servir des titres présents dans le document. Ainsi, si nous parvenons à déterminer des titres pertinents, la qualité de l'index sera grandement améliorée.

Finalement, le titre est donc une entité à part entière, possédant ses propres fonctions et se distinguant nettement des tâches de résumé et d'index. L'objectif du titrage automatique est de proposer un/des titre(s) respectant les contraintes mentionnées dans les définitions. Les méthodes de TALN [1] seront exploitées dans le but de respecter les contraintes qu'un titre doit être un groupe de mots bien formé et qu'il désigne le sujet traité. Un des intérêts du titrage automatique est de proposer un titre pour les documents textuels qui n'en possèdent pas (par exemple, les mails « no objects »), de faire gagner du temps à l'utilisateur en lui proposant un titre automatiquement ainsi que de respecter un des critères de la norme W3C. En effet, le titrage de page Web est un des domaines clés de l'accessibilité des pages web tel que défini par la norme proposée par les associations sur le handicap. Côté lecteur, l'objectif est d'augmenter la lisbilité des pages tout venant obtenues à partir d'une recherche sur mot-clé et dont la pertinence est souvent faible, décourageant les lecteurs devant fournir de grands efforts cognitifs. Côté producteur de site WEB, l'objectif est d'améliorer l'indexation des pages pour une recherche plus pertinente.

Le problème est de savoir quelle est la construction morphosyntaxique d'un titre et si une telle construction peut convenir à tout genre de texte (mail, articles scientifiques, articles journalistiques...). La première idée est qu'un terme clé du texte peut être utilisé en tant que titre, mais la réalité montre que très peu de titres sont conçus par un simple terme. Par ailleurs, la reformulation des éléments pertinents du texte est une grande difficulté du TALN que nous décidons de ne pas exploiter pour le moment. À partir de nos études statistiques portées sur les caractéristiques morphosyntaxiques du titre, nous pouvons définir deux groupes de documents. Dans cet article, d'après la bibliographie et nos études, nous faisons l'hypothèse que les premières phrases d'un document contiennent les informations pertinentes pour un titrage automatique et présentons l'approche ChTITRES permettant la détermination automatique des titres pour un document textuel. Une évaluation des résultats par jugement humain, obtenus sur des données réelles est présentée.

1. Traitement Algorithmique du Langage Naturel.

2 Travaux Antérieurs

Le titrage a pour objectif de représenter pertinemment le contenu des documents en quelques mots. Il peut utiliser des métaphores, l'humour, des jeux de mots ou encore des reformulations [2, 3].

Les titres peuvent avoir plusieurs fonctions, par exemple si nous nous intéressons aux titres journalistiques ou aux titres de mails. D'une part, le titre peut être vu comme objet textuel (Ho-Dac *et al.*, 2004) : polices de caractères, tailles, couleurs, etc. Ceci n'est pas la partie que nous étudierons pour l'instant.

D'autre part, le titre permet a priori d'avoir un aperçu de l'article associé. Ainsi, il est doté d'un contenu sémantique qui a trois fonctions : intéresser/captiver le lecteur, informer le lecteur, introduire le sujet de l'article.Il a été remarqué que les éléments apparaissant dans le titre sont souvent présents dans le corps du texte. Baxendale (1958) a montré que les premières et dernières phrases des paragraphes sont jugées importantes. Les récents travaux de (Belhaoues, 2009; Jacques & Rebeyrolle, 2004) viennent appuyer cette idée et montrent que la proportion de recouvrement des mots de titres est très importante dans les deux premières et deux dernières phrases du texte. Ainsi, une grande partie de l'information permettant la détermination d'un titre se trouve aux extrémités du document. (Vinet, 1993) remarque que très souvent, une définition est donnée dès les premières phrases suivant le titre. En d'autres termes, des mots pertinents apparaîtront dans les premières phrases du texte.

Dans nos travaux, nous commencerons par analyser statistiquement (nombre de mots, présence de noms communs, verbes, etc.) les titres de notre corpus, pour chaque catégorie. Nous mettrons en évidence l'importance de la sélection des syntagmes nominaux pour le titrage. Les résultats portés par les statistiques constitueront une base permettant de déterminer un processus global de titrage automatique, s'appuyant sur des méthodes de sélection statistique et lexicale.

3 Identification des types de textes à titrer

3.1 Protocole d'identification des types

L'analyse statistique des titres est une étape préalable essentielle qui permet de comprendre quel type de titre nous devons attribuer à chaque type de texte. En effet, nous supposons que la forme des titres diffère selon le lecteur visé (enfants, adultes, tout public, etc.) ou encore selon le contenu sémantique du texte. Nous étudions ici cinq catégories de documents : articles/textes Wikipédia (mécaniques, informatique, biologie, biographies, vocabulaire, objets, etc.), articles scientifiques (biologie, physique, linguistique, informatique, etc.), articles journalistiques (*Le Monde*, 1994), mails (divers), listes de diffusion (Listes Ln) et forums de discussion (divers). Pour chaque catégorie, nous avons retenu une centaine de textes en français.

2. Exemple : « A Montpellier, Ségolène fait un retour royal », Midi Libre n^{o}23332.

3. Encore une fois, cela implique que l'on ne puisse pas considérer le titrage comme une tâche de résumé de textes.

L'étiquetage morpho-syntaxique du texte est effectué par TreeTagger (Schmid, 1994). A partir de cet étiquetage, nous avons étudié la fréquence d'apparition de mots de diverses natures dont les noms communs, adjectifs, verbes, adverbes, mots fonctionnels, ponctuations, etc. Cela nous permet de connaître la composition des titres selon les types de textes. D'autre part, nous déterminerons selon quelles proportions les mots du titre sont présents dans le texte. De plus, ce calcul du taux de couverture permet de nous renseigner sur l'emplacement de l'information pertinente dans le texte et indique si le titrage est possible à partir des éléments du texte.

Nous analyserons donc dans la section suivante les caractéristiques morpho-syntaxiques des titres selon les types de textes considérés.

3.2 Analyse et discussion

Table 1 – Documents collectés pour la constitution des corpus.

Nature	% Nom commun	% Entité Nommée	% Verbe	Nombre de mots
Art. scientifiques	97	40	26	9
Art. Wikipedia	87	7	5	3
Art. journalistiques	86	88	25	9
Mails	73	53	6	5
Listes de diffusion	86	99	5	6
Forums de discussion	92	37	15	4

Les résultats (cf. Table 1) montrent que la présence du nom commun dans le titre est primordiale. L'entité nommée (EN) apparaît dans 45% des titres (toutes catégories confondues). Si nous ne tenons pas compte des titres d'articles Wikipédia qui n'utilisent l'EN que dans 7% des cas, la moyenne d'apparition des EN dans les titres est de 60%. Sa présence dans le titre permet de préciser le sens évoqué par les autres termes, précisant (voire fixant) ainsi le sujet.

44% des titres retenus pour notre étude contiennent des adjectifs. La fonction de l'adjectif est de s'adjoindre au nom pour exprimer une qualité (adjectif qualificatif), une relation (adjectif relationnel) ou pour permettre à celui-ci d'être actualisé dans une phrase (adjectif déterminatif). Sa forte présence dans les titres indique la même intention que les EN : préciser la nature du sujet, la plupart du temps par une qualification du nom commun.

Les résultats portés par l'étude statistique des verbes nous permettent d'envisager la formation de deux groupes de documents. En effet, la présence de verbes est forte dans les titres d'articles journalistiques et scientifiques (26%), contrairement aux autres types de textes où les verbes sont très peu présents (6%). Ces résultats peuvent être expliqués par la volonté de l'auteur à représenter au mieux le contenu sémantique du texte. Pour ce faire, les titres longs et l'utilisation de termes complexes sont mis en place. Effectivement, les titres d'articles journalistiques et scientifiques ont des titres d'une taille importante (cf. Tab 3.2). Finallement, ces statistiques peuvent être fonction de la longueur moyenne des phrases dans le texte, en terme de nombre de mots.

Notons qu'une analyse plus détaillée de notre corpus a montré que 50% des titres scientifiques contiennent la conjonction de coordination « et ». La forte présence de ponctuation interne et de coordination marquée par conjonction indique une volonté de bipartition telle qu'elle a été décrite dans (Ho-Dac *et al.*, 2004).

Dans cette optique, les statistiques portées sur les textes Wikipédia montrent que leurs titres ne sont pas "naturels" (i.e. "préformatés") et qu'ils mériteraient une construction plus complexe. Les titres des textes Wikipédia sont très courts (trois mots en moyenne). Ils sont majoritairement constitués d'un nom commun (le mot clé du texte) et d'un adjectif le qualifiant. Dans ce cas, nous devrions plutôt considérer le titre comme un élément simple, pointé par sa description dans le corps de l'article. Remarquons que le format du titre dépend aussi du contexte idéologique de l'auteur. Par exemple, un administrateur de forum préfèrera renommer les titres créés par ses membres afin de proposer une meilleure indexation dans les moteurs de recherche. La liste LN est contrôlée par quelques auteurs uniquement, les titres sont donc fortement influencés par ces auteurs. Le contexte politique des journaux peut aussi influencer sur la forme des titres. Le titrage automatique de documents doit faire face à cette multitude de titres possibles et pertinents que présente le titrage humain, pour un même texte.

3.3 Quel type de titre, pour quel texte ?

Les titres dépendent donc des types de textes et surtout de l'effort de rédaction du corps du texte ainsi que de la présence de verbe. Nous fixons alors deux groupes de documents. Le groupe 1 (G1) contient les textes dont le titre ne présente pas une forte présence de verbe : listes de diffusion, forums de discussion et mails. Le groupe 2 (G2) contient les autres textes, dont le titre présente une syntaxe plus complexe (ce qui explique une longueur des titres plus importante, cf. Tab. 3.2), avec l'emploi de verbe(s) de manière plus fréquente. Ceci implique une meilleure représentation du contenu sémantique.

Dans la suite de l'étude, nous nous intéressons au titrage des documents faisant partie du groupe G1 [4].

4 Approche de Titrage automatique

4.1 Processus Global de Titrage Automatique

D'après les études statistiques précédemment menées, nous proposons un processus global de titrage automatique, composé des quatre étapes suivantes (cf. Fig. 1) :

- Étape 0 : *L'acquisition du corpus* (cf. section Identification des types de textes à étudier).
- Étape 1 : *La détermination des phrases candidates*. Elle s'appuie sur nos statistiques ainsi que sur les travaux précédemment menés. Il s'agit de déterminer les phrases contenant les informations nécessaires au titrage. Nous verrons que très souvent, les termes utilisés dans le titre peuvent se localiser dans les premières phrases du texte.
- Étape 2 : *L'extraction des syntagmes nominaux candidats au titrage*. Elle utilise des filtres syntaxiques tout en s'appuyant sur les études statistiques précédemment menées. En particulier, nous nous intéresserons à la longueur de ces filtres.

4. Le groupe G2 sera étudié plus tard, car ses caractéristiques issues des analyses statistiques montrent qu'ils doivent être traités différemment à cause de leur complexité syntaxique plus élevée.

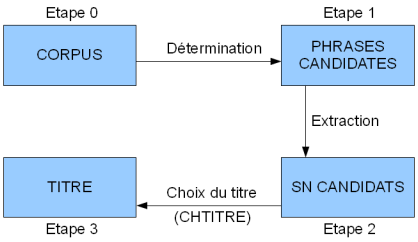


Fig. 1 – Schéma présentant les quatre étapes de l'approche ChTITRES.

– Étape 3 : *Le Choix du Titre (Ch*TITRES*)*. Nous mettrons en œuvre des méthodes statistiques permettant le calcul d'un score et mettant en avant les meilleurs syntagmes pour le titrage. Ces méthodes dépendent du choix des données textuelles à partir desquelles nous déterminerons un titre et de la volonté de mettre en avant (ou pas) les termes discriminants dans le titre.

Nous ne reviendrons pas sur l'acquisition du corpus. Dans la suite de l'article, nous présenterons ces étapes illustrées par des exemples issus de notre programme.

4.2 Extraction des phrases à considérer pour les tâches de titrage

La première étape élémentaire (cf. Schéma Processus Global, Etape 1) consiste à déterminer les données textuelles à partir desquelles nous determinerons un titre. Certains auteurs ont montré que la localisation du titre peut se faire dans les premières phrases du texte (Jacques & Rebeyrolle, 2004). Dans une récente étude (Belhaoues, 2009), il a été conclu que le recouvrement maximal des mots du titre dans le texte, s'obtient en extrayant les sept premières phrases et les deux dernières. Sur notre corpus, 38% des mots contenus dans le titre se retrouvent dans les deux premières phrases et 52% se retrouvent dans le texte complet. Nous proposerons dans la suite de notre étude des approches utilisant au maximum les deux premières phrases.

Les résultats des analyses statistiques ont montré que les titres des documents du groupe G1 contiennent peu de verbes et sont de petite taille (entre deux et six mots environ). Nous commencerons par proposer une liste de syntagmes nominaux extraits du texte, déterminés par leur taille.

4.3 Sélection des syntagmes nominaux maximaux (SN_{max})

L'étape 2 de notre approche (cf. Schéma Processus Global) commence par l'extraction des syntagmes nominaux (SN). Pour cela, nous utilisons Tree Tagger qui permet un étiquetage morpho-syntaxique du texte, bien que nous n'exploitons pas la partie de lemmatisation que cet outil propose. Nous nous sommes appuyés sur les travaux de (Daille, 1996) qui a déterminé des patrons syntaxiques permettant l'extraction de syntagmes nominaux (SN). Par exemple, $Nom1 - Adjectif1$, $Nom1 - Det1 - Nom2$, $Nom1 - Nom2$ etc. Nous avons mis en place un ensemble de 49 patrons syntaxiques permettant l'extraction de syntagmes ayant une taille maximale de 9 mots, taille maxi-

mum constatée cf. Tab. 1 (Par exemple $nom - prp - det - nom - prp - det - nom - prp - nom$). Cette limite de taille est adoptée afin de respecter le résultat des analyses statistiques précédemment exposées. Notons que de nouveaux filtres syntaxiques peuvent être aisément ajoutés.

Voici quelques exemples de SN candidats extraits pour le mail intitulé « Problème avec un étudiant pour passage examen de TP la semaine prochaine » : *un étudiant, un étudiant de FLIN304, mon examen, mon examen de TP, la semaine prochaine, le vendredi, vendredi de 11h30 à 13h, 11h30 à 13h, autres créneaux, les gens, les gens du groupe, un truc, les créneaux, les créneaux de TP.* Finallement, notre travail consiste à sélectionner parmi cette liste de SN candidats, le SN le plus pertinent. Une première présélection permet de ne garder que les SN de taille (en terme de mots) maximale, de manière similaire aux travaux de (Bourigault, 1994), L_{max} et $L_{max} - 1$[5], ceci nous permettant de ne pas élaguer trop rapidement des candidats pouvant être intéressants. Nous les noterons SN_{max}. Notre objectif est de ne pas favoriser les SN binaires afin de respecter les résultats de nos statistiques indiquant que la taille moyenne des titres de G1 est très souvent supérieure à deux mots.

Parmi les SN de la liste précédente, les SN_{max} présélectionnés seront donc : *un étudiant de FLIN304, mon examen de TP, vendredi de 11h30 à 13h, les gens du groupe, les créneaux de TP.*

Si un seul SN_{max} est présélectionné, ce syntagme est le titre. Sinon, afin d'extraire parmi cette présélection le SN le plus pertinent pour l'exploiter en tant que titre (cf. Fig. 1), deux méthodes sont étudiées : T_{MAX} et T_{SOM}. Ceci représente l'étape 3 de notre processus de titrage automatique, l'approche ChTITRES.

4.4 Approche ChTITRES (CHoix du TITRE parmi les SN)

L'étape 3 consiste à sélectionner le SN le plus pertinent pour son utilisation en tant que titre. Nous utilisons des méthodes permettant de calculer l'importance d'un terme dans un texte, notamment la plus emblématique, celle utilisée par (Salton & Buckley, 1988), qui s'appuie principalement sur la fréquence d'apparition du mot dans le document ainsi que sa fréquence d'apparition dans l'ensemble des documents (TF-IDF : Term Frequency - Inverse Document Frequency).

1. **Méthode T_{MAX}.** Pour chacun des mots du SN candidat, le TF-IDF est calculé. Le score pour chaque SN candidat est le TF-IDF maximum rencontré parmi les mots du SN. Avec cette méthode, nous voulons mettre en avant les termes discriminants. Par exemple, pour les syntagmes nominaux « contribution recherche » $SN1$ et « nouvelle relecture » $SN2$, $SN1$ sera retenu, le terme « contribution » étant plus discriminant que les termes « recherche », « nouvelle » et « relecture » dans notre corpus. Il va de soi que cette méthode valorise les EN, celles-ci étant généralement plus discriminantes qu'un autre mot dans le corpus.
2. **Méthode T_{SOM}.** Pour chacun des mots du SN candidat, le TF-IDF est calculé. Le score pour chaque SN candidat est la somme du TF-IDF de chacun de ses

5. La taille moyenne des SN candidats extraits est de 3 mots. Présélectionner les SN de taille supérieure à $L_{max} - 1$ permet ainsi de ne pas tenir compte des SN unaires (sachant que la taille maximale est de 9 mots comme vu précédemment).

termes. Cette méthode privilégie les SN longs. Par exemple, si nous avons les deux syntagmes nominaux « soucis de vibration » $SN3$ et « soucis de vibration avec Saxo » $SN4$ alors $SN4$ sera privilégié puisqu'il contient les mêmes mots que $SN3$ tout en étant plus complet. Cependant, cette méthode permet de distinguer deux SN de même taille : $SN2$ obtient un score plus important que $SN1$ puisque la somme du TF-IDF pour les termes « nouvelle » et « relecture » est plus élevée pour les termes « contribution » et « recherche ».

Dans la suite de notre étude, nous utiliserons ces méthodes sur la première phrase uniquement (T_{MAX1}, T_{SOM1}) ou bien sur les deux premières phrases (T_{MAX2}, T_{SOM2}).

4.5 Sélection lexicale

Les entités nommées [6] (mots ou groupes de mots catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux etc.) peuvent être d'excellents mots-clés permettant de cerner le contenu du texte rapidement. Par exemple, dans un système de questions réponses, QALC (Ferret *et al.*, 2001) utilise les entités nommées pour spécifier le type de la réponse attendue. En effet, comme le mentionne (Zidouni *et al.*, 2009), l'entité nommée représente une description conceptuelle qui fait référence à un objet dont la représentation linguistique est souvent unique.

Si une EN est repérée parmi les trois premiers SN_{max}, alors c'est ce syntagme qui sera retenu par notre approche. Sinon, le SN_{max} retenu sera celui de plus haut score avec T_{MAX} ou T_{SOM}.

5 Expérimentations

Les expérimentations portent sur les documents du groupe G1 : listes de diffusion LN [7], forums de discussion et mails. Pour chacune des trois catégories, dix textes ont été retenus. Ils sont de taille (en nombre de mots), sujets, technicité et effort de rédaction variables.

5.1 Protocole expérimental

L'évaluation a été proposée à 10 experts via une page Web. Trente textes titrés sont proposés aux experts, dix par catégorie de G1. Pour chaque texte, huit titres ont été proposés [8] dont les titres déterminés selon les méthodes T_{MAX1}, T_{SOM1}, T_{MAX2} et T_{SOM2} ainsi que le titre réel (TR). Les trois autres titres ($A1, A2, A3$) sont issus (de

6. Comme (Ren & Perrault, 1992), nous repérons les entités nommées principalement par la présence de majuscules.

7. `http://liste.cines.fr/arc/ln`.

8. Les titres identiques obtenus avec des approches différentes.

manière aléatoire) de la liste des syntagmes nominaux extraits parmi ceux qui n'ont pas été retenus par notre approche.

Pour chaque titre, l'utilisateur doit aprécier sa pertinence en optant pour l'un des critères suivant : Très pertinent $(C1)$, Pertinent $(C2)$, Je ne sais pas $(C3)$, Peu pertinent $(C4)$, Pas pertinent du tout $(C5)$. A chacun de ces critères C_n, est attribué une valeur : -2 pour $C5$, -1 pour $C4$, 0 pour $C3$, +1 pour $C2$ et +2 pour $C1$. La note finale obtenue pour un titre est la moyenne de ces valeurs données par les experts. Pour chaque catégorie de texte (Mails, Liste de diffusion, Forum), un tableau présente les résultats de l'évaluation. Ce tableau contient la moyenne des valeurs correspondant aux critères de notation précédemment exposés, pour chaque titre. Nous comparons nos résultats (titrage automatique) avec les résultats obtenus pour les titres réels.

5.1.1 Mails

Les mails proposés lors de cette évaluation sont des mails personnels, issus de personnes distinctes, de niveau de langue différent et de rédaction plus ou moins soignée. Aucun titre vide (« no object ») n'a été proposé lors de cette évaluation.

Avec une moyenne comprise entre -0.55 et -1.44, les résultats sont jugés peu pertinents (A2, A3) et non pertinents (A1, cf. Tab. 2), ce qui montre bien la qualité de nos méthodes quant au choix du SN_{max} pour le titre.

Le titre réel (TR) obtient une moyenne de 0.57, alors que la méthode T_{MAX2} obtient une moyenne de 0.61. Les titres déterminés par cette dernière méthode sont donc globalement de meilleure qualité que les titres réels. Par exemple, le titre T3 déterminé par T_{MAX2} « Examen de programmation » obtient un score meilleur que le titre réel « Demande d'information ». Pour les mails, il semble donc important de tenir compte des deux premières phrases. En moyenne, toutes les méthodes semblent déterminer des titres pertinents (cf. Tab. 2). Notons tout de même que le score obtenu par la méthode T_{SOM1} est plutôt faible (0.38) comparé au score du titre réel (0.57).

Table 2 – Scores moyens pour le titrage (mails) pour chaque méthode.

Title	TR	T_{SOM1}	T_{MAX1}	T_{SOM2}	T_{MAX2}	A1	A2	A3
T1	1.2	1.6	1.6	1.6	1.6	-1.3	-0.8	7
T2	1.4	-1.3	-1.3	0.8	0.8	-2.4	2.2	-1.8
T3	1.1	-0.7	1.8	-0.7	1.8	-0.8	-1.9	-1.1
T4	-0.1	0.3	0.3	0.6	-0.8	-1.8	-2.3	-1.3
T5	1.9	-0.1	-0.1	-0.2	-0.2	-2.0	-0.3	-1.2
T6	-2	0.4	0.4	0.4	0.4	0.9	0.5	-7
T7	0.6	1.7	1.7	0.0	1.7	0.6	0.4	0.5
T8	1.0	1.8	1.8	1.8	1.8	-1.9	0	-0.2
T9	2	-1.2	-0.1	-0.4	0.5	-1.9	-1.5	0
T10	-1.4	1.3	-1.5	1.3	-1.5	-2.0	-1.8	-1.3
Avg.	**0.57**	**0.38**	**0.46**	**0.52**	**0.61**	**-1.44**	**-0.55**	**-0.64**

5.1.2 Liste de diffusion

Les textes de listes de diffusion proposés aux experts sont issus des archives LN disponibles à l'adresse `http://liste.cines.fr/arc/ln`.

Comme pour les mails, les titres A1, A2 et A3 sont jugés non pertinents, ce qui montre encore une fois la qualité de nos méthodes quant au choix du SN_{max} pour le titre (cf. Tab. 3).

Les résultats témoignent que les titres réels sont très pertinents [9]. Les résultats de l'évaluation pour nos méthodes indiquent que les titres sont globalement pertinent. Les méthodes T_{MAX} permettent la détermination de titres plus pertinents que les méthodes T_{SOM}. La méthode T_{MAX2} permet le titrage le plus pertinent pour cette catégorie de texte. De plus, 50% des titres fournis par T_{MAX2} sont très pertinents avec une moyenne comprise entre 1.5 et 2.

Table 3 – Scores moyens pour le titrage (liste de diffusion) pour chaque méthode.

Title	TR	T_{SOM1}	T_{MAX1}	T_{SOM2}	T_{MAX2}	A1	A2	A3
T1	2.0	0.3	0.3	0.3	0.3	-1.3	-1.3	-0.6
T2	2.0	0.8	0.8	0.8	2	-1.8	-1.5	0.6
T3	2.0	0	0.4	0	0.4	-2.0	-1.7	-1.3
T4	1.9	-0.3	-0.3	-0.3	-0.3	-1.5	-1.3	0.8
T5	1.6	0.4	0.4	0.4	0.4	-1.6	-1.2	-0.3
T6	1.6	0.3	0.3	0.3	0.1	-2.0	1.2	-0.2
T7	1.4	-1.1	-1.1	0.4	1.7	-2.0	-0.2	-1.5
T8	2.0	1.6	1.6	1.6	1.6	-1.6	-1.5	-1.3
T9	1.6	-0.9	1.5	-0.9	1.5	-1.8	-1.1	-1.3
T10	1.9	1.7	1.7	1.7	1.7	-0.1	-1.7	-0.7
Avg.	**1.8**	**0.28**	**0.56**	**0.43**	**0.81**	**-1.57**	**-1.03**	**-0.58**

5.1.3 Forum

Les textes proposés pour l'évaluation de cette catégorie sont extraits de forums rencontrés aléatoirement sur l'Internet (forum de mécanique, numismatique, biologie etc.). Encore une fois les titres A1, A2 et A3 sont jugés non pertinents (cf. Tab. 4). Quatre titres de forum ont été choisi formatés (T7 à T10), c'est-à-dire que les administrateurs ont renommés les titres de messages. Ceci explique le bon résultat pour les titres réels (1.15). Nos quatre méthodes ont des résultats montrant que le titrage est pertinent même s'ils sont plutôt faibles pour T_{MAX2}. La méthode T_{SOM1} obtient le meilleur résultat avec un score de 0.88. Ceci peut s'expliquer par le fait que les messages de forum sont généralement courts et contiennent l'information principale dans la première phrase. Il semble ici que la prise en compte de phrases supplémentaires apporte plus de bruit que d'information pertinente nécessaire au titrage.

Par exemple, pour T6, le titre réel est « Service à domicile ». Les experts ont jugés plus pertinent le titre extrait par nos quatre méthodes : « Société de service à domicile ».

5.1.4 Discussion

En général, nos quatre méthodes permettent de déterminer des titres pertinents. La disparité des résultats peut s'expliquer par le fait que l'expert compare tous les titres qui lui sont proposés par rapport au titre qu'il juge le plus pertinent. Ainsi, même si deux

9. Ceci peut s'expliquer par le fait que la rédaction de ces titres est soignée, appliquée. De plus, les titres de Liste Ln sont formatés.

Table 4 – Scores moyens pour le titrage (forums) pour chaque méthode.

Title	TR	T_{SOM1}	T_{MAX1}	T_{SOM2}	T_{MAX2}	A1	A2	A3
T1	-0.2	1.7	1.7	1.7	1.7	-1.5	-1.7	-0.4
T2	0.5	1.5	1.5	-0.2	-0.2	-0.9	-1.0	-1.3
T3	0.9	1.9	1.9	1.9	1.9	-0.1	0.9	-1.4
T4	1.3	0.6	-0.7	0.6	-0.7	-1.4	-1.2	-1.5
T5	1.4	0.9	0.9	-0.4	-0.4	-0.6	-1.3	-0.6
T6	0.4	1.8	1.8	1.8	1.8	-0.2	-0.4	0.6
T7	1.8	-0.9	-0.9	-0.9	-0.9	-0.8	0	-0.8
T8	1.8	-0.3	-0.3	-0.3	-0.4	-1.7	-1.3	0
T9	1.8	0.1	0.1	0.1	-0.1	-1.4	-0.2	-1.1
T10	1.8	1.5	1.5	1.5	1.5	-1.4	-1.2	-1.4
Avg.	**1.15**	**0.88**	**0.75**	**0.58**	**0.42**	**-1.00**	**-0.74**	**-0.79**

titres sont très pertinents, un des deux sera privilégié en lui associant l'étiquette « Très pertinent » et l'autre se verra associer dans la plupart des cas, l'étiquette « Pertinent ». L'évaluation montre qu'il est préférable d'utiliser les méthodes T_{MAX2} pour titrer les

Table 5 – Scores moyens pour chaque méthode.

Titling	TR	T_{SOM1}	T_{MAX1}	T_{SOM2}	T_{MAX2}	A1	A2	A3
Avg.	1.17	0.51	0.59	0.51	0.61	-1.33	-0.77	-0.67

mails et les listes de diffusion. La méthode T_{SOM1} semble plus appropriée pour le titrage des messages de forum. Dans la catégorie Forum, les résultats indiquent qu'il est préférable de ne tenir compte que de la première phrase. Cependant, de manière générale, l'application de nos méthodes prenant en compte les deux premières phrases offrent souvent de meilleurs résultats (pour deux catégories sur trois). Finallement, nos quatre méthodes d'extraction permettent d'extraire le SN_{max} le plus pertinent parmi les SN_{max} candidats. En effet, les résultats de l'évaluation montrent que les titres A1, A2 et A3 sont toujours peu pertinents (voire pas pertinents du tout) alors que nos méthodes déterminent des titres pertinents (voire très pertinents). Les titres construits par nos méthodes sont donc de bonne qualité (même s'ils obtiennent des résultats légèrement plus faibles que les titres réels, pour deux catégories sur trois, cf. Tab. 5). Notons que les résultats sont faibles pour les titres réels de mails ce qui montre un véritable intérêt de notre approche : proposer automatiquement un titre de mail qui est au moins aussi bon que le titre réel (cf. Tab. 5.1.1). Notre approche permet donc de construire des titres automatiquement, ce qui constitue un réel gain de temps pour l'expert.

6 Conclusion

Nous avons vu que la qualité des titres calculés automatiquement dépend fortement du soin apporté à la rédaction du texte [10]. Néanmoins, l'approche ChTITRES [11] propose

10. Si on suppose que les mails et les messages de forum ont un niveau de rédaction de faible qualité, contrairement aux articles journalistiques ou scientifiques.

11. Disponible à l'adresse `http://www.lirmm.fr/~lopez/`.

des titres pertinents quelque soit le type de texte du groupe G1 (Mails, Forums, Liste de diffusion). Les résultats montrent tout de même que des améliorations peuvent être apportées. Même si une partie des performances de notre approche dépend du Tree Tagger, il nous semble possible d'améliorer nos résultats. En effet, nous avons vu que selon la méthode employée, les résultats peuvent être plus ou moins intéressants selon le type de texte à titrer. Un futur travail pourrait porter sur l'étude des résultats de titrage selon une méthode construite à partir d'une combinaison de méthodes que nous avons proposé ici. Bien sûr, nous ne laissons pas de côté le titrage des textes du groupe G2. Celui-ci nécessite une analyse syntaxique approfondie que nous menerons dans nos prochains travaux. Nous étudierons aussi le sous-titrage. Selon nos statistiques, les titres du groupe G2 devront être construits en tenant compte de la présence plus significative de verbe(s).

Références

BAXENDALE B. (1958). Man-made index for technical literature - an experiment. *IBM Journal of Research and Development*, p. 354–361.

BELHAOUES M. (2009). Titrage automatique de pages web. *Stage Recherche, Université Montpellier II*.

BOURIGAULT D. (1994). Lexter, un logiciel d'extraction de terminologie. application à l'acquisition des connaissances à partir de textes. *Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.*, p. 120–130.

DAILLE B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. *The Balancing Act : Combining Symbolic and Statistical Approaches to language*, p. 29–36.

FERRET O., GRAU B. & AL. (2001). Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse. *TALN*.

HO-DAC L.-M., JACQUES M.-P. & REBEYROLLE J. (2004). Sur la fonction discursive des titres. *S. Porhiel and D. Klingler (Eds). L'unité texte, Pleyben, Perspectives.*, p. 125–152.

JACQUES M. & REBEYROLLE J. (2004). Titres et structuration des documents. *Actes International Symposium : Discourse and Document*, p. 125–152.

PEÑALVER VICEA M. (2003). Le titre est-il un désignateur rigide ? *Dialnet, Vol. 2*, p. 251–258.

REN X. & PERRAULT F. (1992). The typology of unknown words : An experimental study of two corpora. *COLING 92*.

SALTON G. & BUCKLEY C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*, p. 513 à 523.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49.

VINET M.-T. (1993). L'aspet et la copule vide dans la grammaire des titres. *Persee*, **100**, 83–101.

YOUSFI-MONOD M. & PRINCE V. (2006). Compression de phrases par élagage d' arbre morpho-syntaxique. *TSI : Technique et Science Informatiques 25, 4*, p. 447–456.

ZIDOUNI A., GLOTIN H. & QUAFAFOU M. (2009). Recherche d'entités nommées dans les journaux radiophoniques par contextes hiérarchique et syntaxique. *CORIA 2009 - Conférence en Recherche d'Information et Applications*, p.2̃.