Méthode de conception d'une application de veille et d'Analyse Linguistique Assistée par Ordinateur

Richard DELAPLACE (*,***), Marguerite LEENHARDT (*,**), Li-Chi WU (*,**)
rdelaplace@le-semiopole.fr, mleenhardt@le-semiopole.fr, lcwu@le-semiopole.fr

(*) Le Sémiopôle, 66 rue Marceau, 93100 Montreuil (FRANCE),
(**) SYLED/CLA2T, ILPGA Université Paris 3 Sorbonne Nouvelle (FRANCE),
(***) CRIM/ERTIM, Institut National des Langues et Civilisations Orientales (FRANCE).

Mots clefs :
Veille stratégique, statistique textuelle, ingénierie des connaissances, modélisation des connaissances, multilinguisme
Keywords:
Strategic foresight, textual statistics, knowledge engineering, multilinguism, knowledge modeling
Palabras clave :
Escudriñar estratégico, estadísticas textuales, ingeniería del conocimiento, multilinguismo, formalización del conocimiento

Résumé

Nous présentons une solution de veille stratégique sociétale, développée par la société Le Sémiopôle. La veille sociétale, segment de la veille stratégique en plein essor, répond aux nouveaux besoins exprimés par les marques et les institutions : analyser les opinions exprimées à leur égard dans les communautés d'opinion sur le web. En effet, analyser l'avis des internautes est devenu déterminant, car ils s'expriment de façon plus visible et mieux médiatisée. Le principe de conception de cette solution remet l'analyste expert au sein des processus de production et représente une évolution dans les processus métiers de la veille sociétale en ligne. L'intervention humaine, nécessaire pour des étapes de validation qualité, intervient tout au long de la constitution du corpus de données d'analyse. La place du veilleur dans le processus de décision est alors mise au premier plan. La robustesse du système de récolte de données multilingue constitue un atout opérationnel important dans la phase actuelle, où les territoires de veille sociétale en ligne tendent à se globaliser : il devient impératif de pouvoir étudier des communautés d'internautes s'exprimant en différentes langues. Enfin, la fluidification des étapes de traitement jusqu'à l'analyse en elle-même permet de tirer un profit nettement plus élevé de l'expertise du linguiste, pour aller vers un système d'analyse hautement qualitatif. Cela est mis en œuvre dans le processus d'Analyse Linguistique Assistée par Ordinateur, mis au point par Le Sémiopôle.
Les problématiques de récolte d'information textuelle en contexte multilingue et le processus d'Analyse Linguistique Assistée par Ordinateur, couplées dans le cadre de la veille sociétale en ligne, sont présentées successivement. Les fonctionnalités implémentées dans le système que nous présentons introduit de nombreux bénéfices opérationnels, parmi lesquels :

- gain de temps dans le processus d'analyse de corpus, grâce à l'intégration de fonctions de calcul textométrique ;
- flexibilité du système, qui permet l'enrichissement de corpus avec des trames d'annotation dédiées à des besoins d'analyse particuliers, l'étude des opinions ;
- optimisation du processus de production des études, grâce à l'export de données volumétriques dans des formats compatibles avec des logiciels tiers.

Introduction

L'émergence d'Internet, corrélée à l'utilisation croissante des outils de publication de contenus en ligne et du web communautaire durant la dernière décennie, a profondément modifié les logiques de traitement et d'accès à l'information. Cet état de fait rend nécessaire la mise au point de stratégies d'acquisition et d'analyse de données face à différents verrous technologiques. Dans la phase actuelle, cela révolutionne les parcours de l'information et le pouvoir de l'opinion publique en ligne. La veille sociétale, segment de la veille stratégique, répond aux nouveaux besoins exprimés par les marques et les institutions : analyser les opinions exprimées à leur égard dans les communautés d'opinion sur le web. En effet, on assiste ces dernières années à une remise en question profonde du schéma traditionnel 'top-down' (communication descendante) de la communication de marque et de la communication institutionnelle. Cela au profit d'une dynamique 'bottom-up' (communication ascendante), dans laquelle l'avis des internautes – à la fois audience des médias, interlocuteurs des institutions, aussi consommateurs de la marque –, est devenu déterminant parce qu'il s'exprime de façon plus visible et mieux médiatisée. Les problématiques de veille sociétale qui en découlent sont variées et complexes et tirent de nombreux bénéfices à intégrer les outils et méthodologies de l'analyse des discours en ligne

Les solutions techniques actuellement existantes dans ce secteur exploitent encore peu les innovations technologiques en Traitement Automatique des Langues (TAL), tout comme les méthodes d'analyse de données textuelles développées en Textométrie. Dans l'objectif d'optimiser la conception des systèmes d'analyse qualitative de données web, utilisés en veille stratégique sociétale, notre contribution vise à :

- opérationnaliser la tâche de récolte d'informations textuelles à partir de structures erratiques, en contexte multilingue ;
- développer des séquences de traitement génériques pour fluidifier les étapes de traitement, depuis l'extraction de données identifiées comme pertinentes, jusqu'à l'analyse proprement dite.

1 Problématiques de formalisation, variété des données sur le web multilingue

Une problématique surgit lorsque l'on souhaite mettre en place une chaîne de traitement dédiée à l'analyse qualitative des contenus textuels sur le web : faire face à de grands ensembles de données hétérogènes, faiblement structurées et en constante expansion. La récupération de données multilingues pose quant à elle le double problème de la gestion des encodages et du traitement spécifique des pages de textes rédigés dans des langues différentes.

1.1 L'hétérogénéité des supports d'informations sur le web

Un premier facteur d'hétérogénéité est celui de la variété des types de support : des sites aux plateformes de réseau social, en passant par les forums, les blogs, les portails d'information, les webzines ou encore les services de micro-blogging, la structure qui contient les informations, le *contenant*, diffère fortement. Les métalangages utilisés pour décrire l'affichage des textes étant faiblement contraints, il est difficile de localiser les informations pertinentes à récupérer. Un autre facteur important est le processus de production des contenus : les processus rédactionnels dans lesquels les internautes peuvent intervenir, par exemple en laissant un commentaire, complexifient la tâche d'extraction d'informations pertinentes.

Les informations récupérées à partir de différents supports de production en ligne doivent être structurées dans des trames de contenants homogènes, pour effectuer une analyse sur corpus. Parmi les types de données recherchées, on peut par exemple considérer différents types de contenu textuel – titre, chapeau, date et heure, légendes, paragraphes. Plusieurs objectifs d'analyse des informations sur Internet rentrent dans le paradigme de la veille stratégique, ces objectifs mettant en rapport

des types de supports différents, en particulier dans le cadre d'une veille sociétale[1]. La robustesse des systèmes joue un rôle déterminant dans la gestion de l'ensemble de ces variables.

1.1.1 De la pertinence et du bruit

Le traitement des informations hétérogènes implique la mise au point de stratégies d'extraction d'information pertinentes, c'est-à-dire ne contenant pas d'information *bruitée*, par exemple les encarts publicitaires d'une page si seuls les contenus textuels sont ciblés pour l'extraction. Considérons un article de presse en ligne et l'ensemble des commentaires produits suite à sa publication comme des informations pertinentes à extraire. Un certain nombre d'informations complémentaires doivent être récupérées, telles que la date, l'auteur, la rubrique ou encore les mots-clés. On peut y ajouter les informations multimédia, telles que les images, les sons ou les vidéos, dont on peut obtenir les liens URL. Enfin, des informations contextuelles plus complexes peuvent également être récupérées comme, par exemple, la relation de réponse entre deux commentaires via la disposition du texte – indentation du commentaire répondant à un premier – ou via des interpellations marquées par les pronoms 'tu' ou 'vous' ou la mention du nom du destinataire du message.

Figure 1 - Exemple d'interpellation (encadré bleu) et d'imbrications (flèches rouges)

Parmi les informations bruitées qu'on peut observer dans l'exemple ci-dessus, on relève par exemple les artefacts « Répondre » et « Alerter » ; bien entendu, les éléments du menu, les publicités et autres contenus sans rapport avec les informations que l'on souhaite extraire, comptent aussi parmi les informations bruitées.

L'outil informatique a ses limites et peut difficilement déterminer seul quelles sont les informations pertinentes et quels sont les éléments de bruit. Cependant, la collecte de corpus manuelle, intégralement réalisée par l'humain, induit un coût important en temps et en main d'œuvre. Pour contribuer à pallier cette

[1] D'autres contextes applicatifs prendraient en compte les contenus multimédia, de type image, son ou vidéo.

problématique et réduire les coûts et le temps de production des corpus, nous avons réduit les tâches manuelles à la seule sélection des structures dont les contenus sont pertinents. Notre approche a consisté à automatiser la récolte des données pertinentes à partir d'un tri manuel préalable et en fonction d'une liste exhaustive des objets à récupérer, indépendamment de la structure qui les contient. Nous avons adopté une stratégie fondée sur le XPath[2]. D'après nos travaux, les différents supports présents sur la toile ne changent que peu fréquemment leurs structures HTML : il n'est pas rare, pour un même site, qu'une même description XPath puisse être appliquée à plusieurs pages. Il s'agit donc d'une approche économe dans la mesure où une même source n'aura pas nécessairement besoin d'être décrite plusieurs fois.

1.1.2 Les résultats de l'extraction

Nous montrons ci-après des résultats de tests d'extraction de contenus à partir de journaux en ligne, en français et en coréen, réalisés le 26 juillet 2010 sur les quotidiens LeFigaro.fr et Joins.com. La stratégie adoptée a l'avantage d'être applicable quelles que soient les langues utilisées, comme nous pouvons le constater ci-dessous avec les exemples d'extraction à partir de pages en langue occidentale et en langue asiatique.

Tableau 1 - Etapes de l'extraction appliquée à une page de journal en ligne en français

Exemple de page d'article à extraire : LeFigaro.fr	Exemple de description XPath de la structure à extraire	Exemple de structure XML de l'article extrait
	<article> <racine>.//*[@id='article']</racine> <pageSuivante></pageSuivante> <source></source> <horodatage>/div[@class='infos']/span[@class='sign']/text()</horodatage> <dateMiseAJour>/div[@class='infos']/span[@class='sign']/text()</dateMiseAJour> <auteur>/div[@class='infos']/span[@class='sign']/span[@class='auteur']//text()</auteur> <traducteur></traducteur> <titre>/h1/text()</titre> <sousTitre></sousTitre> <chapeau>/h2/text()</chapeau> <corpsDuTexte>/div[@class='texte']/p/text()</corpsDuTexte> <motsCles>/div[@class]/p/span/a/text()</motsCles> <liens>/div[@class='texte']/p/a/@href</liens> <like></like> <votes></votes> <valeurNote></valeurNote> <baseNote></baseNote> <retweets></retweets> <vue></vue> </article>	

[2] XPath est un langage d'indication de chemin dans du contenu plus ou moins structuré, de type XML, à l'aide duquel on exprime une requête afin d'extraire un élément quelconque pointé par le chemin indiqué. Pour davantage de précisions sur ce langage, voir [5], en particulier le chapitre 9.

Tableau 2 - Etapes de l'extraction appliquée à une page de journal en ligne en coréen

Exemple de page d'article à extraire : Joins.com	Exemple de description XPath de la structure à extraire	Exemple de structure XML de l'article extrait
	<article> <racine>//*[@id='CC670']</racine> <pageSuivante></pageSuivante> <source>//*[@id='articleTitNews']/h2/span//text()</source> <horodatage>/div[@class='btmBox']/p/text()</horodatage> <dateMiseAJour>/div[@class='btmBox']/p/text()</dateMiseAJour> <auteur></auteur> <traducteur></traducteur> <titre>//*[@id='articleTitNews']/h2/text()</titre> <sousTitre></sousTitre> <chapeau></chapeau> <corpsDuTexte>//*[@id='articleBody']//text()</corpsDuTexte> <motsCles></motsCles> <liens></liens> <like></like> <votes></votes> <valeurNote></valeurNote> <baseNote></baseNote> <retweets></retweets> <vue></vue> </article>	

Ce processus permet également de tirer le juste profit de l'expertise du veilleur, mieux valorisé puisqu'il touche aussi bien à la sélection des sources pertinentes qu'à la sélection des contenus pertinents, en minimisant le temps consacré aux tâches manuelles d'extraction pour privilégier la tâche intellectuelle de réflexion et de sélection. Du point de vue de la compétence métier de veilleur, notre procédure « équipe » ce dernier d'outils et de compétences techniques lui permettant de mieux appréhender l'objet de son travail : les contenus, leur sélection, leur comparaison et leur analyse. Nous obtenons donc au final un corpus structurellement homogène pour un moindre coût en temps et en main d'œuvre. De plus, la constitution de corpus multilingues, comparables ou parallèles, est facilitée : notre solution permet donc également la constitution de ressources linguistiques hautement qualitatives.

1.2 Les langues sur le web et les contenus multilingues

De nombreux travaux sont consacrés, depuis 2004 notamment, à l'adaptation de technologies de traitement automatique des langues peu présentes sur le web. Dans le même temps, le multilinguisme s'est accru sur Internet, en particulier avec l'arrivée des internautes s'exprimant dans des langues orientales, arabes et asiatiques notamment. La richesse et la variété des langues de production des contenus constituent un facteur de complexité supplémentaire. Les problématiques de gestion des encodages ne constituent cependant plus un verrou technologique à l'heure actuelle. Par ailleurs, un système de veille stratégique en ligne doit gérer les cas de contenus multilingues, sur une même page web par exemple. Les systèmes actuels assurent une prise en charge robuste des langues occidentales à alphabet latin. Les outils de traitement automatique des langues orientales ont quant à eux fait l'objet d'un effort de recherche important : l'extraction d'information ne représente donc plus un verrou technologique en soi, mais des leviers restent à mettre en place pour passer de l'extraction « standard » à l'extraction qualitative, qui s'appuie sur un processus supervisé par l'humain et non sur une procédure d'apprentissage automatique, dont les résultats seraient de moindre qualité pour des coûts plutôt élevés. D'autre part, les besoins de veille stratégique se globalisent : analyser les opinions des internautes suite à un lancement produit dans plusieurs pays est un cas de figure typique. L'enjeu est ici de gérer de façon fluide la récupération des contenus et les premiers décomptes – nombre de formes, d'occurrences, d'hapax, par

exemple – sur les corpus multilingues. La difficulté découle de l'articulation, dans un même flux de traitement, d'un moteur de récolte de contenus textuels et d'un moteur de traitement textométrique. La robustesse du moteur de traitement dépend de sa capacité à optimiser l'étape de segmentation du fil textuel en fonction de la langue, aspect sur lequel nous revenons plus en détails dans la section suivante. D'autre part, si des leviers existent pour la gestion des encodages, cet aspect demeure cependant non trivial. Dans le cas du japonais par exemple, différents encodages doivent être pris en compte. Les conventions d'écriture informatique ne correspondant pas à celles des langues à alphabet latin, des encodages spécifiques, Shift-JIS, EUC-JIS et ISO-2022 ont été mis au point. L'UTF-8 permet de régler les problèmes de compatibilité qui découlent de cette variété d'encodages pour le japonais[3]. Si des techniques existent pour la détection automatique de l'encodage, certains cas doivent être gérés obligatoirement par l'humain, par exemple à cause de pages web mal formées ou comportant de mauvaises informations d'encodage. Les langues occidentales et asiatiques sont actuellement les mieux représentées sur les plateformes de réseau social et de micro-blogging. L'imbrication des langues dans les contenus diffusés et enrichis par les usagers est courante. Il est donc impératif de pouvoir déterminer correctement la langue de production des contenus en amont de l'analyse sur corpus. Les sites de presse en ligne et les pages web multilingues présentent une problématique analogue. La détection automatique de la langue est une technologie robuste, pleinement opérationnelle pour les cas rencontrés sur les supports de type presse en ligne, entre autres. Cette tâche doit être supervisée par l'humain dans certains cas, en particulier dans les supports communautaires où les systèmes de détection automatique obtiennent de moins bonnes performances, les usages spécifiques du « parlécrit », comme les néologismes, les phénomènes d'abréviation ou la présence de smileys, rendant leur tâche plus difficile à accomplir.

En somme, l'intervention humaine est nécessaire tout au long du processus de constitution du corpus, afin d'effectuer différentes tâches de validation de qualité. L'analyste veilleur interagit donc de façon étroite avec le système afin d'éviter la récupération de données bruitées : il est garant de la qualité du corpus d'étude et son rôle est valorisé tout au long de la chaîne de production, depuis la sélection de la matière première – la donnée textuelle brute, ici – jusqu'à son analyse.

2 Récolte d'informations textuelles multilingues et Analyse Linguistique Assistée par Ordinateur

Développer une passerelle pour fluidifier les étapes de traitement, depuis l'extraction de données signifiantes jusqu'à l'analyse proprement dite, cela sous-tend une gestion robuste des problèmes de segmentation dans les différentes langues traitées. Il s'agit en effet à ce niveau de l'applicatif, de communiquer avec un moteur de traitement textométrique, qui permet d'analyser les textes dans une trame de contenants homogènes. Nous prenons l'exemple concret des problématiques de segmentation de deux langues asiatiques fortement présentes en ligne, le Chinois et le Japonais.

2.1 La relativité de la notion de « mot » : problématiques de segmentation en Chinois et en Japonais

Les traitements textométriques opèrent à partir d'une trame de contenants au sein desquels le fil textuel est segmenté en unités lexicales. Une telle représentation permet d'analyser les données textuelles de façon fine. Cependant, parvenir à isoler des segments pertinents dans le fil textuel lorsque le système doit être opérationnel sur des données multilingues n'est pas trivial. En effet, si la définition naïve du mot correspond à une chaîne de caractères précédée et suivie d'un blanc dans la plupart des langues occidentales à alphabet latin, isoler des segments textuels pertinents dans des données en Chinois ou en Japonais pose des problèmes d'ambiguïté morphologique importants car la notion de mot graphique n'existe pas. Dans le cas du Chinois, la réalité linguistique recouverte par la notion de mot fait débat. En effet, comme le Japonais, le Chinois mobilise un système d'écriture syllabique, sans blanc graphique pour segmenter le fil textuel. Si certains indices morphologiques fournissent des paramètres fiables pour la segmentation du Japonais en unités lexicales, le Chinois requiert quant à lui la prise en compte des

[3] Cependant, l'UTF-8 demeure peu usité par les internautes nippons, notamment pour des raisons d'identité nationale.

positions syntaxiques des unités. De récents travaux sur l'évaluation des systèmes de segmentation du Chinois ont d'ailleurs montré que les modèles CRF, les chaînes de Markov et les algorithmes d'information mutuelle obtiennent de meilleures performances. Les systèmes hybrides, qui fondent leur segmentation sur le couplage de critères morphosyntaxiques et de dictionnaires de formes de référence, sont les plus robustes pour traiter les textes en Japonais [14].

2.1.1 De la complexité de la segmentation du chinois et du japonais

Les langues comme le français possédant le séparateur graphique dans un texte posent peu de problèmes. Pour ces langues, les problèmes de reconnaissance des unités se localisent dans les mots graphiques qui ne coïncident pas avec des unités linguistiques. Il s'agit non pas d'une segmentation, mais plutôt d'une reconstitution de mots discontinus. Dans le cas du chinois ou du japonais, l'absence de tout séparateur entre les unités lexicales nécessite tout d'abord une phase de segmentation des phrases. Une segmentation correcte est cruciale car cela influence directement la qualité et les performances des outils de TAL intégrant la segmentation dans le prétraitement des corpus. Les deux problématiques majeures de cette première analyse linguistique du chinois sont l'ambiguïté lexicale et la reconnaissance des mots inconnus. Premièrement, il existe deux types d'ambiguïté en chinois : l'ambiguïté de combinaison et l'ambiguïté d'intersection intérieure [4].

A) l'ambiguïté de combinaison

Si la chaîne AB est une lexie et A et B sont aussi des lexies indépendantes, AB est une chaîne d'ambiguïté de combinaison.

1) 學生會打籃球 xue shen hui da lan qiu.

Cette phrase peut être segmentée de deux façons différentes :

1a) 學生會_打_籃球[4] *Un syndicat sait jour au basket-ball.
syndicat des étudiants (n) – jouer (v) – basket-ball (n)

1b) 學生_會_打_籃球 Un étudiant sais jouer au basket-ball.
étudiant (n) – pouvoir (v) – jouer (v) – basket-ball (n)

Ces deux structures syntaxiques sont correctes, mais la chaîne 學生會 peut être découpée en deux façons différentes. Seule la phrase 1b), étant donné le sens de la phrase, est segmentée correctement.

B) l'ambiguïté d'intersection intérieure

Si une chaîne de caractères telle que XYZ dont les sous-chaînes XY et YZ sont tous des lexies, XYZ est une chaîne d'intersection intérieur.

2) 乒乓球拍卖完了 ping pang qiu pai mai wan le.

Cette phrase peut aussi être segmentée de deux façons différentes :

2a) 乒乓球_拍卖_完_了 Les balles de ping-pong sont vendues aux enchères.
Les balles de ping-pong (n) – vendre aux enchères (v) – finir (v) – (particule)

2b) 乒乓球拍_卖_完_了 Les raquettes de ping-pong sont vendues.
Les raquettes de ping-pong (n) – vendre (v) – finir (v) – (particule)

[4] Les traits bas que nous avons mis manuellement servent à séparer entre deux unités lexicales pour mieux les distinguer.

Dans cet exemple, les deux segmentations sont syntaxiquement et sémantiquement correctes. Seule une analyse pragmatique ou contextuelle permet de choisir la segmentation adéquate. Deuxièmement, abordons le problème de la reconnaissance des mots inconnus, tels que les noms de personnes, lieux, institutions. De nouveaux mots apparaissent sans cesse, ils ne sont donc pas exhaustivement référencés dans un dictionnaire. C'est l'une des principales raisons des difficultés rencontrées dans la segmentation du chinois. L'approche basée sur un dictionnaire est la méthode la plus fondamentale et la plus expérimentée dans la segmentation du chinois. Différents critères sont pris en compte, comme celui de la longueur des mots, d'appariement du sens dans une phrase, de la fréquence des mots, etc. Etant donné qu'aucun des dictionnaires ne peut être exhaustif, la segmentation à l'aide d'une analyse linguistique peut être utilisée en complément. Les approches statistiques sur les Modèles de Markov Cachés, les modèles CRF, etc. sont développées dans de récents travaux [10, 11, 12, 16] : moins dépendantes des ressources linguistiques, elles présentent l'avantage de la reconnaissance des mots inconnus et des désambiguïsations.

Une langue agglutinante comme le japonais, différente du chinois - une langue isolante - pose aussi des problèmes de segmentation. Le japonais possède trois types d'écriture différents[5] et permet une segmentation en fonction du changement de type de caractère. Le changement de type de caractère correspond à peu près à la frontière entre deux unités, ceci est souvent utilisé pour l'analyse des mots non référencés dans un dictionnaire. Un certain nombre de particules grammaticales servent à spécifier la fonction (sujet, complément, etc.) du syntagme qui la précède [3]. Plusieurs éléments d'une phrase sont ainsi repérés comme l'exemple suivant.

3) 兄は新聞を買います。ani-wa shinbun-wo kai-masu Mon frère achète un journal.
3a) 兄は_新聞を_買い_ます。)
frère-(sujet) journal-(COD) acheter-(forme d'un verbe)[6]

Une des méthodes les plus efficaces pour la segmentation des textes japonais en écriture mélangées repose sur l'exploitation des chaînes de Markov. Cette approche est adoptée dans les analyseurs morphologiques les plus utilisés comme Juman [6] et Chasen [9].

Le traitement automatique des textes chinois et japonais, en particulier concernant la gestion de l'absence de blanc typographique et l'utilisation différentes typologies morphologiques ont déjà franchi certains obstacles dus à la complexité du système d'écriture des deux langues, en particulier dans le champ des recherches en textométrie multilingue.

2.1.2 L'importance d'une solution adéquate à la nature linguistique du texte segmenté

Si la qualité de la segmentation est un levier essentiel, l'adéquation de la méthode de segmentation à l'objectif applicatif est un aspect fondamental pour la pertinence des résultats. De nombreux outils de segmentation, appelés segmenteurs, ont été développés depuis le premier système de segmentation automatique du chinois apparu en 1983, conçu par l'Institut aéronautique de Pékin. Il est nécessaire de choisir un segmenteur qui soit le mieux adapté au type de textes que l'on souhaite traiter : des textes de spécialité – par exemple des corpus de textes juridiques – ou des textes courants – articles de presse –, ne posent pas les mêmes problèmes. Le repérage des formes lexicales courtes permet la segmentation des formes lexicales, dite *segmentation fine*. La *segmentation grossière*, qui produit des formes lexicales longues, est plutôt adaptée à l'extraction des noms propres et, plus généralement, des entités nommées. Ce dernier cas de figure est fréquent, par exemple lorsqu'il s'agit de produire un dictionnaire thématique ou d'examiner des rapports d'une forme lexicale à une autre afin de pouvoir retrouver l'intégralité de l'information.

---

[5] Les deux systèmes d'écriture japonaise sont nommés *hiragana* et *katakana*, et les caractères chinois, appelés *kanji*. Les mots pleins, les radicaux sont souvent écrits en caractères chinois. Les *hiragana* servent à écrire des mots d'origine japonaise ou les mots n'ayant qu'une fonction grammaticale tel que les particules enclitiques. L'utilisation de *katakanas* est limitée à transcrire des mots étrangers ou des onomatopées.

[6] La traduction entre parenthèses est un mot grammatical qui désigne la fonction du mot précédent.

Le choix de segmenteurs adéquats est donc fondamental en contexte industriel, sur deux plans au moins : ils participent de l'optimisation des prétraitements et analyses implémentés dans des applications entrepreneuriales de veille stratégique et permettent de réduire le temps du traitement sur de grandes masses de données. Une étude récente sur trois segmenteurs chinois connus a d'ailleurs montré que le segmenteur Hylanda[7] est plutôt adapté au traitement des termes spécialisés comme le domaine juridique, il segmente des formes en unités longues. Le segmenteur conçu par Stanford University[8] est plus indiqué pour les textes courants et la segmentation des formes en unités courtes [15].

2.2 Les méthodes textométriques et le processus d'Analyse Linguistique Assistée par Ordinateur

Un moteur de traitement textométrique permet le passage d'une suite de mots non représentative à un ensemble d'informations significatives, en exploitant les contenus isolés dans la chaine textuelle. De nombreuses dimensions de comparaison sont alors à la disposition de l'analyste, qui peut par exemple :

- comparer la distribution des contenus isolés à l'échelle du corpus,
- comparer la variation des contenus dans différents contenants (parties du corpus),
- associer des caractéristiques distributionnelles ou linguistiques aux contenus textuels (annotations, association de ses particularités à chaque forme).

Les différentes fonctionnalités disponibles visent à extraire du corpus les régularités parlantes pour les analystes. Plusieurs calculs de statistique textuelle sont mobilisables pour l'analyse textométrique d'un corpus. Parmi eux, nous rappelons le principe de quatre fonctions d'exploration particulièrement adaptées à l'analyse textométrique de corpus en veille sociétale :

1) Calcul des Spécificités : méthode statistique visant à projeter pour un sous-ensemble donné d'un corpus, les objets dont la présence est représentative de celui-ci.
2) Analyse Factorielle des Correspondances (AFC) : représentation graphique de la distance des objets comparés sur la base d'algorithme de similarités.
3) Calcul des Segments Répétés (SR) : ensemble d'objets ordonnés dont les occurrences dans le co-texte d'un corpus lui suppose un signifié particulier.
4) (Poly)cooccurrence : ensemble d'objets non ordonnés dont la cooccurrence contextuelle indique l'existence d'un réseau sémantique.

Ces outils, entre autres, constituent l' « équipement » de l'analyste veilleur, qui lui permet de mettre en valeur son expertise du domaine et sa compréhension des méthodes d'analyse linguistique.

2.2.1 Résonance textuelle et exploitation

Nous rappelons le principe de résonance textuelle, qui permet « de considérer les variations conjointes des différentes unités textuelles [dans différents] volets du corpus »[9]. Comme l'explique [13], dont nous nous permettons de reprendre le schéma original de ce principe ci-dessous, la résonance textuelle procède d'une relation d'induction entre deux espaces du corpus : on observe en fait la projection de segments textuels dans chacun de ces espaces, afin d'y déceler des correspondances. La puissance de cet outil d'exploration et d'analyse réside dans la portée variable de la relation d'induction : on peut l'appliquer aussi bien localement – entre deux sections d'un même espace textuel – que globalement – entre deux espaces textuels différents. Ce type d'analyses permet, par exemple,

[7] L'entreprise Hylanda à Tianjin fait des études sur le traitement automatique de la langue chinoise dans la fouille de textes. Elle développe également des produits de nouvelles technologies. Son segmenteur a été mis en application par plusieurs moteurs de recherche.
[8] Le groupe de spécialistes du traitement des langages naturels de l'Université Stanford. Il développe des outils de TAL, comme le taggeur part-of-speech, le parseur, le segmenteur pour le chinois, l'outil pour reconnaître des entités nommées, etc.
[9] Nous reprenons ici les mots de Salem, dans son article intitulé *Introduction à la résonance textuelle*, in *Actes des JADT 2004*.

l'étude des conversations écrites médiatisées par ordinateur, aussi bien sur un plan synchronique que diachronique. Les cas d'application sont nombreux ; on peut encore citer la mesure du degré de proximité entre deux textes ou l'analyse thématique des conversations dans les espaces communautaires. Autrement dit, la résonance textuelle est un outil méthodologique puissant pour la veille sociétale : il s'agit de combiner comparaisons fines, analyse de l'évolution de l'information et observation contextuelle des correspondances linguistiques entre les textes.

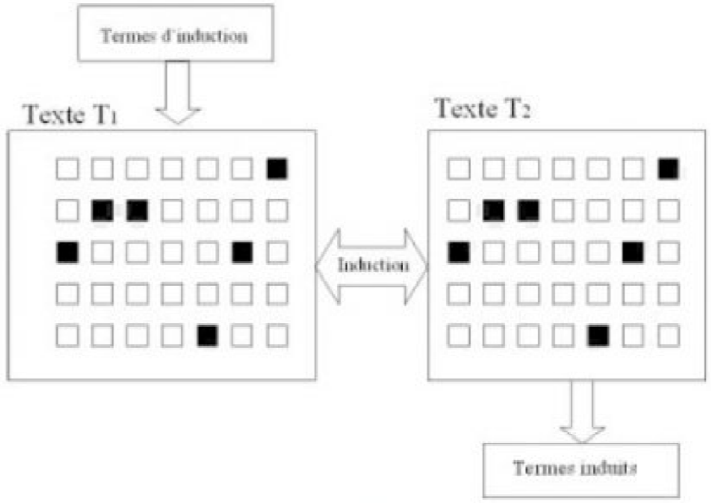


Figure 1.
Schéma général de la résonance textuelle
entre deux ensembles de textes

Figure 1 - Schéma original de la résonance textuelle, donnée par [13]

Les recherches en Linguistique Appliquée sur les échanges médiatisés par ordinateur, proches des problématiques que nous traitons en veille sociétale, bénéficient à l'heure actuelle des méthodes d'analyse textométrique et constituent un champ d'application privilégié. D'autre part, si les calculs statistiques mobilisés en textométrie sont applicables indépendamment de la langue, l'analyse de corpus en contexte multilingue comporte des aspects spécifiques. Le traitement des langues agglutinantes impose par exemple de recourir à des filtres de segmentation spécifiques, pour palier l'ambiguïté des graphies, comme on l'a vu plus haut. Ce type de calculs permet d'obtenir des cartographies textuelles riches, en ce qu'on peut par exemple identifier la diffusion d'une thématique au travers de plusieurs espaces textuels qui peuvent être de langues différentes.

2.2.2 La place de l'analyste

A un autre niveau de complexité linguistique, l'analyste doit nuancer l'interprétation des résultats à l'aune des usages linguistiques socialement normés. Par exemple, la redondance lexicale, synonyme de répétition en langage courant, n'est pas acceptable en français, mais d'usage en anglais ou en japonais. Les processus métiers de veille sociétale sont donc enrichis par les analystes, qui interviennent sur des tâches d'analyse qualitative et mobilisent leur savoir-faire d'expert linguiste. Le

processus métier d'Analyse Linguistique Assistée par Ordinateur (ALAO) consiste à réintégrer l'expertise linguistique dans l'ensemble du processus de production des études qualitatives. L'analyste décide l'application de prétraitements, tels que la correction orthographique ou l'étiquetage syntaxique, avant de débuter l'analyse. Il peut également enrichir la description du corpus, en particulier avec l'annotation des opinions dans les textes. Plus en aval de l'analyse, il mobilise des méthodes d'analyse textuelle, sémantique ou conversationnelle, selon le corpus. L'analyste est avant tout formé à l'utilisation et à la compréhension des fonctions de calcul, pour garantir la qualité de leur interprétation et donc de l'analyse. Son rôle est donc actif : l'analyste veilleur interagit avec le système dans l'ensemble des étapes, de l'acquisition à l'analyse, en passant par le prétraitement des données. La qualité des analyses produites est plus fiable, en particulier parce que le veilleur connaît et comprend les fonctions de calcul qu'il mobilise et est donc à même d'en livrer une interprétation raisonnée, reposant sur des arguments tangibles.

3 Etudes de cas : veille sociétale et analyse des discours en ligne

Nous présentons deux études de cas, pour illustrer les sorties produites par la chaîne de traitement développée par Le Sémiopôle. Rappelons avant tout les fonctionnalités principales de cette chaîne de traitement d'Analyse Linguistique Assistée par Ordinateur :

- récolte de contenus textuels sur le web en contexte multilingue (crawling) ;
- prétraitements du corpus ;
- export de résultats volumétriques dans un format tabulaire ;
- calculs textométriques ;
- visualisation graphique de résultats de calculs textométriques ;
- enrichissement du corpus grâce à l'ajout de différents niveaux d'annotation.

En somme, notre méthode de conception opérationnalise une chaîne de production, en prenant en charge au sein d'une plateforme unique l'ensemble des étapes de production et d'analyse qualitative d'un corpus.

La démocratisation de l'Internet et le fort engagement des utilisateurs dans le web impliquent une démultiplication des acteurs du discours médiatique, qui a pour conséquence une remise en question profonde du schéma traditionnel 'top-down' (communication descendante) de la communication de marque et de la communication institutionnelle, ces dernières années. Le schéma qui prévaut est maintenant celui du 'bottom-up' (communication ascendante), dans lequel l'avis des internautes, qui constituent à la fois l'audience des médias, les interlocuteurs des institutions aussi bien que les consommateurs de la marque, est aujourd'hui devenu déterminant parce qu'il s'exprime de façon plus visible et mieux médiatisée. Les problématiques de veille sociétale qui en découlent sont variées et complexes et tirent de nombreux bénéfices à intégrer les outils et méthodologies de l'analyse des discours en ligne, depuis l'analyse qualitative de l'audience de médias en ligne (première étude de cas) à l'étude des discours d'internautes et l'analyse de leurs opinions sur une marque ou un produit (seconde étude de cas).

La première étude de cas est liée à l'analyse des retombées médiatiques sur le web, sur l'actualité d'un parti politique français. Le corpus se fonde sur la syndication de plus de 804 articles publiés dans cinq quotidiens nationaux en ligne, Le Monde, Le Figaro, Libération, Le Nouvel Obs et Le Point, entre novembre 2008 et août 2009. L'intégralité des commentaires associés à chaque article a été intégrée au corpus, qui se compose donc également de 200 496 commentaires écrits par 23 371 auteurs différents, pour un total de 4,8 millions de mots. La standardisation variable des sites, comme les différences de structures de contenants, sont des problèmes typiques d'une telle tâche. La récupération des commentaires a par exemple demandé une procédure de description rigoureuse des contenants de chaque journal, pour éviter la présence de données bruitées. Dans le cadre de notre processus métier de veille, l'intervention humaine pour la validation qualité est rendue obligatoire à ce niveau, en amont de la phase d'extraction des données. Nous donnons un exemple de graphique généré à l'aide d'un tableur, en utilisant la fonctionnalité

d'export de données volumétriques au format .csv. D'un point de vue méthodologique, il est fort intéressant de comparer l'évolution des commentaires d'internautes, du nombre d'auteurs de commentaires et du nombre d'articles au cours du temps. Ces données permettent par exemple d'estimer concrètement les périodes auxquelles les internautes ont été plus ou moins réactifs à l'actualité.

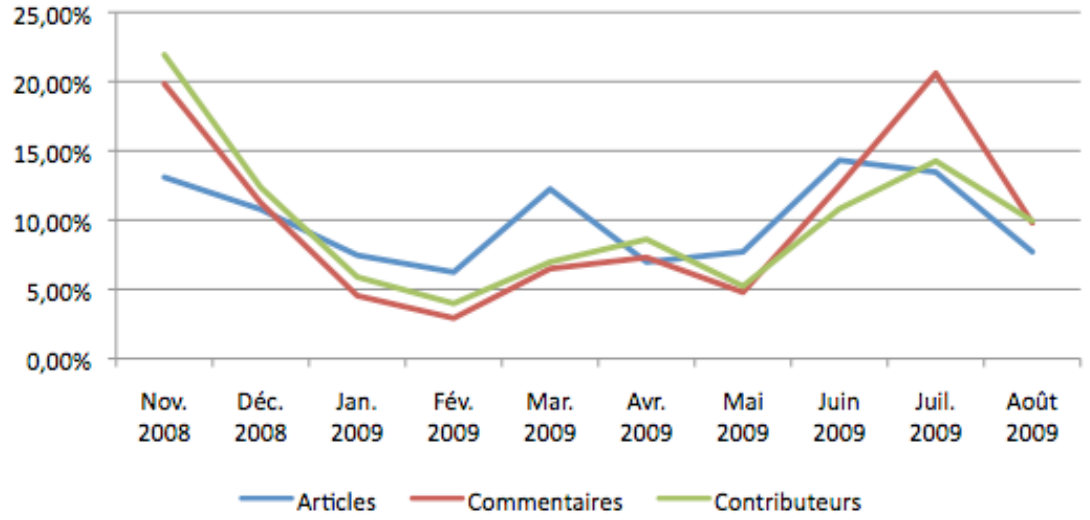


Figure 2 - Volumétries des publications, des commentaires et des contributeurs dans le corpus. Les courbes sont générées par un logiciel de tableur informatique, à partir de l'export de résultats volumétriques fournis par le système au format .csv

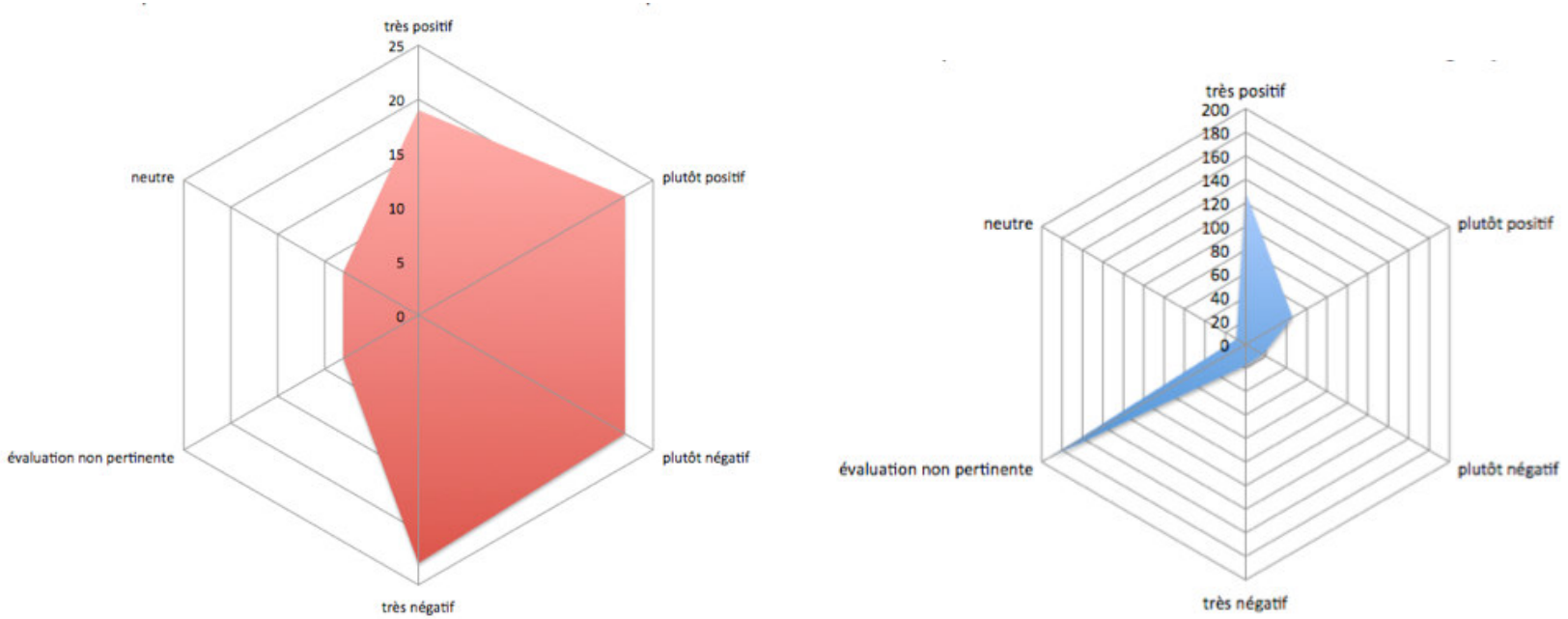
La seconde étude de cas est une analyse de réputation menée dans le cadre d'un lancement produit, qui comporte un volet d'étude des opinions dans les commentaires d'internautes sur différents supports en ligne. Le corpus d'étude de 40 000 mots contient des contenus textuels en français et en anglais britannique. Dans un premier temps, les tendances majeures de chaque sous-corpus par langue sont identifiées en appliquant un calcul d'*analyse factorielle des correspondances*[10]. Cette fonctionnalité, qui fournit un résumé des proximités lexicales entre les parties du corpus considérées, est intégrée dans notre solution d'analyse. L'analyste génère donc tout au long de son travail des représentations visuelles de qualité, directement exploitables dans le livrable final à remettre au client. L'objectif était ici de mettre en avant les tendances principales des commentaires d'internautes, faits en réaction à des publications annonçant la sortir d'une nouvelle gamme de maroquinerie chez une grande maison de mode hexagonale.

[10] Pour davantage de précisions sur l'AFC, voir [7], en particulier le chapitre 3.

Figure 3 - Résultat d'une AFC pour positionner les commentaires d'internautes produits sur des supports différents ; exemple du corpus français, en contexte d'analyse des retombées autour d'un lancement produit.

Les données rassemblées en corpus ont été récoltées dans différents supports en ligne (portails, webzines, presse en ligne, forums, webservices vidéo), en français et en anglais britannique. La figure 2 permet d'apprécier le positionnement des discours en fonction de la proximité linguistique des textes, dans le sous-corpus français. Pour obtenir ce résultat, nous avons exploité l'AFC. Cette analyse donne ici à voir le fait que les commentaires des internautes français postés sur les supports de type 'Portail', 'Webzine', 'Presse en ligne' et 'Forum' entretiennent de fortes proximités linguistiques : en l'occurrence, des analyses complémentaires – en particulier le calcul des spécificités et des polycooccurrences en fonction des différents supports – ont confirmé que les commentaires étaient liés à trois thématiques caractéristiques du lancement produit : la marque, le produit et l'égérie de la campagne publicitaire. Bien que les commentaires postés sur les supports de type 'Blog' s'opposent à ces premiers supports (distance sur l'axe vertical), ils partagent cependant une grande partie de leur vocabulaire avec eux (proximité sur l'axe horizontal). Les analyses complémentaires indiquent une proximité au niveau des thématiques marque et produit : l'égérie de la campagne est sous-spécifique des discours des internautes sur les supports 'Blog'. On effectue ensuite des calculs volumétriques à partir des discours annotés en fonction des opinions exprimées –

dont la détection se fonde sur une grille adaptant le modèle de l'Appraisal Theory[11]. Le système de veille mis en place permet à l'analyste d'ajouter facilement une trame d'annotation de contenus sur les données textuelles, afin d'enrichir la description du corpus en identifiant des phénomènes linguistiques complexes et impossibles à traiter automatiquement à l'heure actuelle. Il est également possible d'opérer des décomptes sur les annotations et, de la même façon que pour les données volumétriques, d'utiliser l'export au format .csv, pour garantir la compatibilité avec des outils de tableur.



Extrait 2 – Répartition des évaluations (opinions) dans les commentaires d'internautes dans différents supports du web français. A gauche, la répartition par orientation des opinions dans les supports de type webzine ; à droite, dans les supports de type blogs.

Nous avons ensuite pu analyser les spécificités des discours dans ces supports, afin de caractériser les thématiques associées aux différentes modalités d'évaluation – donc les polarités plus ou moins positives des opinions des internautes. L'étude des opinions a, dans un premier temps, permis de contraster les tendances des avis exprimés par les internautes sur le produit et/ou la gamme de maroquinerie en question. Dans les supports de type 'Webzine', les internautes en ont eu une perception négative. Parallèlement, dans les supports 'Blogs', si le nombre d'opinions exprimées est plus important, la majorité d'entre elles sont non pertinentes, c'est-à-dire ne portent sur aucune des thématiques caractéristiques du lancement produit. En l'occurrence, il est ressorti de l'analyse que l'opération blogueurs déployée sur le web français dans le cadre de ce lancement produit a davantage contribué à mettre en avant les blogueurs associés à la marque, plutôt que le produit lui-même.

[11] Le modèle de l'Appraisal Theory, introduit par [8], a été le cadre de plusieurs travaux d'implémentation d'algorithmes dans le champ de l'*opinion analysis*, comme par exemple chez [1] et [2].

Perspectives

La solution de veille retenue associe les technologies de récolte de données textuelles et les moteurs de traitement textométrique.
Ses fonctionnalités présentent de nombreux bénéfices opérationnels, parmi lesquels :

- le gain de temps dans le processus d'analyse de corpus, grâce à l'intégration de fonctions de calcul textométrique ;
- la flexibilité du système, qui permet l'enrichissement de corpus avec des trames d'annotation dédiées à un besoin d'analyse particulier, par exemple l'étude des opinions ;
- le gain de temps dans le processus de production des études, grâce à l'export de données volumétriques dans des formats compatibles avec des logiciels tiers.

Elle remet de plus l'analyste expert au sein des processus de production et constitue une évolution dans les processus métiers de la veille sociétale en ligne. L'intervention humaine reste nécessaire pour des étapes de validation qualité, tout au long de la constitution du corpus. La robustesse du système de récolte de données multilingue est un atout opérationnel important dans la phase actuelle, où les territoires de veille sociétale en ligne tendent à se globaliser. Enfin, la fluidification des étapes de traitement jusqu'à l'analyse en elle-même permet de tirer un profit nettement plus élevé de l'expertise du linguiste. Un tel flux de travail est mis en œuvre par le processus d'Analyse Linguistique Assistée par Ordinateur, mis au point par Le Sémiopôle.

Références

[1] BLOOM K., STEIN S. et ARGAMON S., Appraisal extraction for news opinion analysis at NTCIR-6, *Proceedings of NTCIR-6*, 2007, p 279-289

[2] BLOOM K., GARG N. et ARGAMON S., Extracting appraisal expressions, *Proceedings of NAACL HLT*, 2007, p 308-315

[3] HAGUENAUER C., *Morphologie du japonais moderne*, Vol. I : Généralités, mots invariables, Paris : C. Klincksieck, 1951

[4] HOU M. 侯敏 et Sun J-J. 孫建軍, Hanyu zidong fenci zhong de qiyi wenti 漢語自動分詞中的歧異問題 (Les problèmes d'ambiguïté de la segmentation automatique du chinois), *Applied Linguistics 語言文字應用*, 1996, (1), p 68-72

[5] HAROLD E. R. et MEANS W. S., *XML in a Nutshell*, O'Reilly (Ed.), 2001

[6] KUROHASHI S. et NAGAO M., *Japanese morphological analysis system JUMAN*, Kyôto University, 1998

[7] LEBART L. et SALEM A., *Statistique textuelle*, Dunod, Paris, 1994

[8] MARTIN J.R. et WHITE P.R.R., *The language of evaluation: appraisal in English*, Palgrave, London, 2005

[9] MATSUMOTO Y., KITAUCHI A., YAMASHITA T., HIRANO Y., MATSUDA H., TAKAOKA K. et ASAHARA M., *Morphological analysis system Chasen 2.2.9.*, Nara Institute of Science and Technology, 2002

[10] PENG F., FENG F. et MCCALLUM A., Chinese segmentation and new word detection using conditional random fields, *Processing of COLING*, 2004, p 562-568

[11] PONTE J. M. et CROFT W. B., Useg: A retargetable word segmentation procedure for information retrieval, *Symposium on Document Analysis and Information Retrieval*, 1996, Vol. 96

[12] SPROAT R. et SHIH C., A statistical method for finding word boundaries in Chinese text, *Computer Processing of Chinese and Oriental Languages*, 1990, 4(4), p 336-351

[13] SALEM A., Introduction à la résonance textuelle, In *Actes des JADT 2004 (7èmes Journées internationales d'Analyse Statistique des Données Textuelles)*, 2004, p 986-992

[14] VERT J.-P., *Panorama de la recherche en traitement automatique du langage écrit au Japon*, MBA thesis, Corps des mines, Paris, 1998

[15] WU L., Outils de segmentation du chinois et textométrie, *RECITAL (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, Montréal, 2010

[16] ZHANG H., LIU Q., CHENG X., ZHANG H. et YU H., Chinese lexical analysis using hierarchical hidden markov model, *Proceedings of the second SIGHAN workshop on Chinese Language processing*, Sapporo, Japan, 2003, p 63-70