RECITAL 2010, Montréal, 19-23 juillet 2010

Outils de segmentation du chinois et textométrie

Li-Chi WU[1]

(1) SYLED, Université Sorbonne Nouvelle Paris III, 13 rue de Santeuil, 75005 Paris, France

lucielichi@gmail.com

Résumé La segmentation en mots est une première étape possible dans le traitement automatique de la langue chinoise. Les systèmes de segmentation se sont beaucoup développés depuis le premier apparu dans les années 1980. Il n'existe cependant aucun outil standard aujourd'hui. L'objectif de ce travail est de faire une comparaison des différents outils de segmentation en s'appuyant sur une analyse statistique. Le but est de définir pour quel type de texte chacun d'eux est le plus performant. Quatre outils de segmentation et deux corpus avec des thèmes distincts ont été choisis pour cette étude. À l'aide des outils textométriques *Lexico3* et *mkAlign*, nous avons centré notre analyse sur le nombre de syllabes du chinois. Les données quantitatives ont permis d'objectiver des différences entre les outils. Le système Hylanda s'avère performant dans la segmentation des termes spécialisés et le système Stanford est plus indiqué pour les textes généraux. L'étude de la comparaison des outils de segmentation montre le statut incontournable de l'analyse textométrique aujourd'hui, celle-ci permettant d'avoir accès rapidement à la recherche d'information.

Abstract Chinese word segmentation is the first step in Chinese natural language processing. The system of segmentation has considerably developed since the first automatic system of segmentation of the 1980's. However, till today there are no standard tools. The aim of this paper is to compare various tools of segmentation by through statistical analysis. Our goal is to identify the kind of texts for which these segmentation tools are the most effective. This study chose four segmentation tools and two corpora, marked by distinct themes. Using two textometric toolboxes, *Lexico3* and *mkAlign*, we focused on the number of syllables in Chinese. The quantitative data allowed us to objectify disparities between tools. The Hylanda system turns out to be effective in the segmentation of specialized terms and the Stanford system is more appropriate for general texts. The comparative study of segmenters shows the undeniable status of textometrical analysis which is able to quickly access information retrieval.

Mots-clés : Textométrie, comparaison des segmenteurs chinois, nombre de syllabes

Keywords: Textometry, comparison of Chinese segmenters, number of syllables

1 Introduction

Les méthodes d'analyse des textes sur ordinateur sont répandues depuis longtemps dans les travaux sur les langues occidentales. Mais l'étude textométrique du chinois n'a commencé que dans les années 1980. Les premières études quantitatives concernaient la lexicologie comme par exemple la production du Dictionnaire des fréquences des mots chinois contemporains (Modern Chinese Frequency Dictionary). De nombreux travaux sur des livres spécifiques ont été publiés à la même époque, spécialement des ouvrages sur le chinois classique. Dans la majorité des cas, les calculs de ces études ont été faits manuellement, les chiffres statistiques ne seraient donc pas garantis sans erreur. C'est ainsi qu'a émergé la recherche sur les textes qui a mené vers les études statistiques des textes chinois.

Notre travail a pour objectif d'effectuer une comparaison de quatre outils de segmentation, également appelés segmenteurs. L'étude est basée sur une analyse textométrique et nous nous sommes concentrée sur le nombre de syllabes en chinois. La comparaison des segmenteurs a pour but de définir les spécificités pour chaque segmenteur en analysant les types de textes les plus adaptés.

L'étude textométrique en chinois s'est développée tardivement, certainement à cause de facteurs liés au système de l'écriture traditionnelle chinoise. L'informatisation de cette langue s'est en effet révélée beaucoup plus complexe que celle du système basé sur l'utilisation des alphabets latins. La mise en place de technologies permettant la saisie et l'affichage des caractères chinois a permis de dépasser la complexité de ce système d'écriture. La norme internationale du codage de caractère Unicode fournit désormais la possibilité de représenter des textes dans toutes les langues, indépendamment du système informatique ou des plates-formes.

Les progrès considérables des équipements informatiques nous apportent une très grande liberté d'accès à l'information. Les applications du traitement automatique des langues sont de plus en plus variées : la traduction, le résumé de textes, la fouille de textes, l'extraction d'information, etc. Le chinois possède une typographie différente des langues occidentales en raison de son système d'écriture. Un texte chinois est représenté par une chaîne de caractères continue, sans blanc typographique[1]. Pour qu'un ordinateur effectue une analyse correcte, la première étape primordiale est de segmenter les textes en unités lexicales (« tokenisation », découpage d'un texte en mots). Or, il n'y a pas de consensus entre les Chinois et différentes segmentations sont acceptées. Le premier système de segmentation automatique a été réalisé en 1983 par l'Institut aéronautique de Pékin. Par la suite, beaucoup d'outils de segmentation du chinois ont été développés, mais il n'y a pas d'outil standard. Une même phrase peut être découpée de façon différente selon l'outil utilisé. Il est donc crucial de choisir un outil de segmentation adéquat permettant l'accès direct à l'information recherchée.

2 Outils de segmentation et corpus

Étant donné que l'écriture chinoise crée des difficultés dans le traitement automatique des langues, il est nécessaire d'avoir une norme de la segmentation des mots chinois. Une norme de segmentation du chinois comporte en général deux parties : segmentation des unités lexicales et annotation des catégories grammaticales. En 1993, la République populaire de Chine a conçu *la norme de la segmentation des mots*

[1] L'absence d'espace entre les mots était pratiquée dans l'antiquité grecque et romaine ainsi qu'au début du Moyen Âge européen. Dans le cas des langues européennes, le lecteur devait d'abord repérer les syllabes puis les mots. Dans le cas du chinois, les syllabes sont marquées, un caractère correspondant à une syllabe, et il ne reste qu'à assembler les caractères en mots.

chinois contemporains pour le traitement informatique (信息處理用現代漢語分詞規範 xinxi chuli yong xiandai hanyu fenci guifan) pour le traitement automatique du chinois. Cette norme propose des principes et des règles de segmentation des mots chinois, qui ne sont pas toujours opératoire et parfois difficiles à appliquer. Depuis, de nombreuses normes de segmentation du chinois ont été créées par différents organismes en Chine continentale ou en dehors du territoire, afin d'avoir des règles de segmentation améliorées. Elles sont soit appuyées sur cette *norme* d'État, soit créées par l'organisme en question. Deux de ces segmenteurs que nous avons étudiés (ICTCLAS et SF_PKU, cf. 2.1) sont fondés sur la norme d'État . Ils effectuent une segmentation similaire que notre analyse va mettre en évidence.

2.1 Description des segmenteurs

Nous avons utilisé dans notre étude les quatre segmenteurs les plus connus dans la segmentation du chinois.

1. Hylanda *Zhongwen zhineng fenci*

   Le segmenteur Hylanda est une application commerciale. Il utilise des méthodes comme le nombre maximum antérieur de segments (forward maximum matching, FMM), nombre maximum postérieur de segments (backward maximum matching, BMM), etc. (Liang, 1984). Son programme annote les catégories grammaticales des mots segmentés[2]. La caractéristique de Hylanda est de reconnaître des entités nommées : des noms propres de personnes, des noms de lieux géographiques, des noms des organismes, etc., et spécialement des noms propres dans le domaine de la mécanique.

2. Chinese Lexical Analysis System

   Le segmenteur ICTCLAS (Zhang et al., 2003) a été créé par la Chinese Academy of Science et a été mis à jour plusieurs fois[3]. Il possède des fonctions comme l'annotation lexicale, la reconnaissance d'entités nommées et de nouveaux mots et leur intégration dans un dictionnaire défini par l'utilisateur. ICTCLAS s'appuie sur un grand lexique et utilise un modèle de Markov[4]. L'étiquetage grammatical se réfère principalement au corpus annoté du *Quotidien du peuple* de l'Université de Pékin (Yu et al., 2000) car ce corpus est utilisé comme corpus d'apprentissage de la segmentation.

3. Stanford Chinese Word Segmenter

---

[2] 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. La version de l'outil que l'on a trouvée dans le site de l'entreprise est une version d'essai sans annotation des catégories grammaticales la quantité de texte d'essai est donc limitée. http://www.hylanda.com/server/ (page consultée le 6 janvier 2010)

[3] Le segmenteur d'ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) a été mis au point par Kevin Zhang à l'Institute of Computing Technology, Chinese Academy of Sciences. Il en existe plusieurs versions, nous avons utilisé la version 2008 *zhenghe ban* (整合版) qui a été améliorée par rapport aux anciennes versions pour la recherche universitaire. Le téléchargement est disponible dans un forum de discussion spécialisé pour la linguistique de corpus. http://www.corpus4u.org/attachment.php?attachmentid=426&d=1220683589 (page consultée le 8 janvier 2010)

[4] Le Modèle de Markov Caché Hiérarchique (Hierarchical Hidden Markov Model, HHMM) est un modèle statistique utilisé dans le traitement automatique des langues. Il est appliqué à l'extraction d'informations, la reconnaissance vocale, etc.

Le segmenteur Stanford qui s'appuie sur la norme de l'Université de Pennsylvania (Xia, 2000) a été produit par le groupe de spécialistes du traitement des langues naturelles de l'Université Stanford. Cet outil utilise le modèle des champs aléatoires conditionnels pour étiqueter les données (Tseng et al., 2005). Il propose deux modèles de segmentation sans annotation des catégories lexicales, l'une s'appuyant sur la norme du corpus annoté de l'Université de Pékin, ou SF_PKU et l'autre s'appuyant sur celle de Penn Chinese Treebank[5], ou SF_CTB.

2.2 Préparation du corpus

Afin d'initier cette étude de la segmentation en textométrie, deux échantillons de test contenant un petit nombre d'unités lexicales ont été choisis. Nous avons utilisé deux corpus de différents domaines possédant un nombre de caractères similaires correspondant à 16 000 sinogrammes : le corpus de la Constitution de la République Populaire de Chine[6], désormais *Constitution*, et le corpus des conférences de presse du Ministère des Affaires Étrangères de Chine[7], désormais *Presse*. La taille totale des deux corpus segmentés par les outils étudiés est entre 8 300 et 9 800 occurrences, ce qui correspond à approximativement entre 1 000 et 1 600 formes différentes (cf. 3.2 pour plus de détails). Pour chaque corpus, nous obtenons quatre segmentations différentes du même texte au moyen des quatre segmenteurs. Les textes chinois ont été sauvegardés en format texte brut avec le jeu de caractères GB2312, qui est destiné à représenter les caractères simplifiés[8].

Dans un premier temps, les segmentations obtenues pour les deux corpus ont été alignées afin de faciliter l'analyse. Pour cela, nous avons eu recours à l'outil d'alignement mkAlign[9], ce qui nous a permis de comparer en lexicométrie les deux textes. L'alignement a permis d'obtenir des textes où chaque groupe aligné est signalé par le symbole dièse « # » comme séparateur. Les quatre textes ont été regroupés dans un même fichier et séparés par des balises.

---

[5] Le Penn Chinese Treebank contient des corpus segmentés, étiquetés de POS de 500 milliers de mots chinois. Les ressources des corpus proviennent de l'agence de presse Xinhua, Sinorama news magazine et Hong Kong News. http://www.cis.upenn.edu/~chinese/ctb.html (page consultée le 20 janvier 2010)

[6] Les textes électroniques ont été recueillis sur le site de l'agence de presse chinoise Xinhua. http://news.xinhuanet.com/newscenter/2004-03/15/content_1367387.htm (page consultée le 6 janvier 2010)

[7] Nous avons rassemblé les textes électroniques des dialogues entre le porte-parole et des journalistes sur six conférences de presse du 11 juin au 30 juin 2009 dans le site officiel du Ministère des Affaires Étrangères de la République populaire de Chine. http://www.fmprc.gov.cn/chn/gxh/wzb/fyrbt/jzhsl/default.htm (page consultée le 18 janvier 2010)

[8] GB2312 est un jeu de caractères utilisé en Chine. Il attribue un code de 16 bits pour un sinogramme simplifié, soit deux octets. Mais certains caractères rares ne peuvent pas être représentés avec ce système. GB18030 a donc été créé et il supporte les caractères tant du chinois simplifié que du chinois traditionnel. Big 5 est un jeu de caratères utilisé à Taiwan et à Hong Kong pour les caractères traditionnels.

[9] Le programme mkAlign, créé par Serge Fleury de l'Université Paris III, permet d'afficher et de corriger simultanément un alignement de deux textes de même langue ou de langues différentes. http://tal.univ-paris3.fr/mkAlign/ (page consultée le 17 février 2010)

3 Étude des outils

3.1 Exploration préliminaire

Le module *variation*[10] de mkAlign (Fleury, Zimina, 2009) permet de repérer toute variation d'un texte source par rapport à un texte cible ou dans deux types de segmentations d'un même texte, comme c'est le cas ici. Les différences de segmentation sont mises en évidence au moyen de la coloration. Les numéros des paragraphes, signalés par le séparateur # (figure 1) sont notés dans la première colonne. La visualisation du corpus nous permet d'avoir un aperçu des deux textes et d'examiner leurs différences et leurs similitudes.

Figure 1 : Variations des textes de segmenteurs Hylanda et celles d'ICTCLAS

Nous avons calculé le nombre de formes différentes selon les quatre segmentations. Six paires de comparaisons ont été faites en s'appuyant sur trois types de distinctions prédéfinies : l'ajout (case verte), la modification (case bleue) et la suppression (case rouge). Le nombre le plus élevé (case bleue) de formes différentes segmentées pour chaque paire, est obtenu avec les segmenteurs Hylanda et SF_CTB pour *Constitution.* Pour *Presse*, il est obtenu par Hylanda et ICTCLAS. On en déduit qu'ils possèdent de nombreuses formes de segmentations différentes. Les segmenteurs ICTCLAS et SF_PKU possèdent, au contraire, le moins de formes de segmentations différentes. Nous faisons l'hypothèse que ICTCLAS et SF_PKU sont les plus similaires dans la segmentation pour des textes de droit et des textes de presse.

3.2 Accroissement de vocabulaire

L'étude de l'apparition de nouvelles formes graphiques du corpus *Constitution* confirme les différences quantitatives entrevues entre les quatre types de segmentations. La courbe d'accroissement de vocabulaire calculée simultanément pour les quatre volets du corpus (figure 2) montre que la croissance du vocabulaire du segmenteur Hylanda augmente plus rapidement que celles des trois autres. L'interruption de la courbe de Hylanda avant les autres indique que le texte comporte moins d'occurrences. La courbe (rouge) correspondant à l'apparition de nouveaux mots chinois est située au-dessus de celles qui correspondent à l'apparition des mots dans les textes segmentés par ICTCLAS, SF_CTB et SF_PKU. Ceci confirme que le texte segmenté par Hylanda comprend le plus grand nombre de formes graphiques. La courbe (jaune) située au-dessous témoigne que le texte segmenté par SF_CTB possède moins de formes graphiques. Les courbes d'ICTCLAS (verte) et de SF_PKU (bleue) se superposent quasiment. Nous supposons que leurs segmentations sont similaires. Nous pourrions avancer l'argument que cela provient du fait que ICTCLAS et SF_PKU utilisent la même norme, *la norme* de l'État de Chine, à savoir celle fonctionnant selon le corpus annoté de l'Université de Pékin.

[10] *Variation* permet de repérer les variations dans deux versions d'un même texte ou dans deux textes différents en les comparant avec l'outil d'alignement mkAlign.

Des paliers créés par le ralentissement de l'accroissement du vocabulaire au cours du récit pourraient être mis en rapport d'une courbe à l'autre. Au ralentissement qui survient sur la courbe du segmenteur SF_CTB (abscisse 2 000) correspond un ralentissement sur celle du segmenteur ICTCLAS et SF_PKU (abscisse 1 900) et sur celle de Hylanda (abscisse 1 800). À celui qui survient pour le texte de SF_CTB (abscisse 5 600) correspond également un ralentissement dans le texte de ICTCLAS et de SF_PKU (abscisse 5 000) et celui de Hylanda (abscisse 4 500).

Quant au corpus *Presse* (figure 3), les courbes de l'accroissement de vocabulaire se superposent quasiment dans les premières cinq cents occurrences. L'interruption de la courbe de Hylanda avant les trois autres, comme pour le corpus *Constitution,* confirme que le texte comporte moins d'occurrences. C'est également le texte segmenté par SF_CTB qui possède le plus d'occurrences pour *Presse*. Comme le montre la figure 2, le nombre d'occurrences entre les quatre segmenteurs pour *Presse* est très proche.

La courbe d'ICTCLAS et celle de SF_PKU sont également très proches comme nous l'avons déjà vu dans *Constitution*. Sur la figure 3, nous pouvons voir que les quatre courbes suivent la même progression avec peu de décalage entre elles par comparaison aux courbes de la figure 2. Cette similarité indique que les quatre textes de *Presse* sont segmentés de façon similaire au niveau des occurrences et au niveau des formes graphiques, à l'inverse du corpus *Constitution*.

Grâce aux représentations graphiques, la distinction entre les segmenteurs apparaît clairement. De plus, le genre du texte influence la segmentation. En effet, le texte de presse a été segmenté de façon semblable par les quatre segmenteurs, alors que nous avons mis en évidence de grandes différences dans les versions segmentées du corpus de droit.

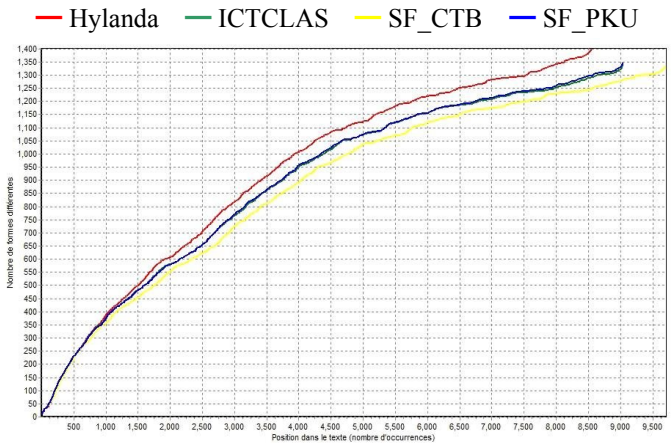

Figure 2 : Accroissement de vocabulaire dans les quatre volets de *Constitution*

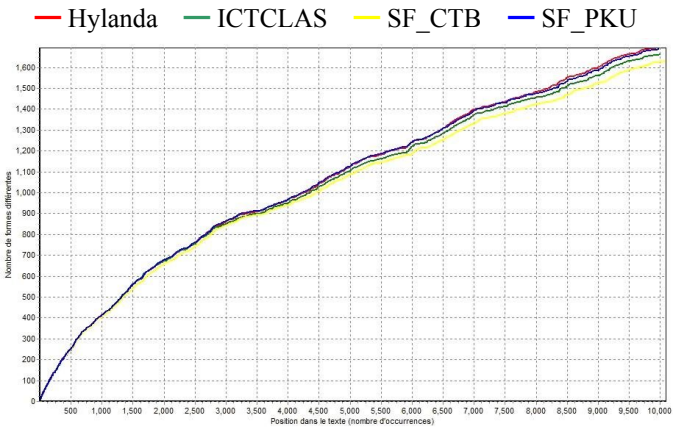

Figure 3 : Accroissement de vocabulaire dans les quatre volets de *Presse*

3.3 Nombre de syllabes

Le chinois est une langue monosyllabique. Cela est vrai pour le chinois ancien ou archaïque dans une forme traditionnelle de la langue écrite du style noble (*wenyan*) avant l'apparition du chinois vernaculaire (*baihua*). Le chinois contemporain a tendance à passer du monosyllabisme au dissyllabisme, voire polysyllabisme (Wang, 2000). Les deux corpus aux thèmes différents (l'un provient d'un domaine spécialisé, l'autre d'un domaine général) segmentés par les quatre outils, nous ont poussée à faire une étude sur le nombre de syllabes. D'anciens travaux ont indiqué que le nombre de syllabes est influencé par plusieurs facteurs : phénomènes phonétiques, sémantiques, la formation des mots, la communication de langue, développement de la société, etc. (Alleton, 1994 ; Huang, Yang, 1990).

3.3.1 Formes fréquentes

Dans cette étude, nous nous sommes appuyée sur les cent premières formes les plus fréquentes de chaque texte. Les monosyllabes et dissyllabes sont les plus nombreux au sein des deux corpus. Dans *Constitution*, les dissyllabes sont plus nombreux que les monosyllabes (figure 4). Les polysyllabes (trois syllabes ou plus) sont beaucoup moins nombreux. On note tout de même que les pentasyllabes sont particulièrement remarquables dans Hylanda. Le segmenteur ST_CTB possède peu de quadrisyllabes et aucun pentasyllabe.

Dans le corpus *Presse* (figure 5), les monosyllabes sont plus nombreux que les dissyllabes par rapport au texte *Constitution*. Mais ils sont dominants dans le corpus. Au contraire, les quadrisyllabes sont beaucoup moins nombreux, un seul quadrisyllabe apparaît dans le segmenteur ICTCLAS et aucun de pentasyllabes.

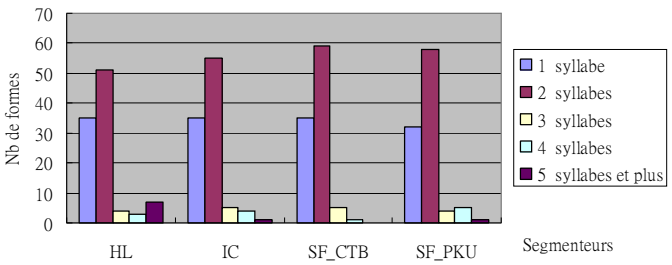

Figure 4 : Répartition des formes par segmenteur sur les cent premières formes les plus fréquentes de *Constitution*

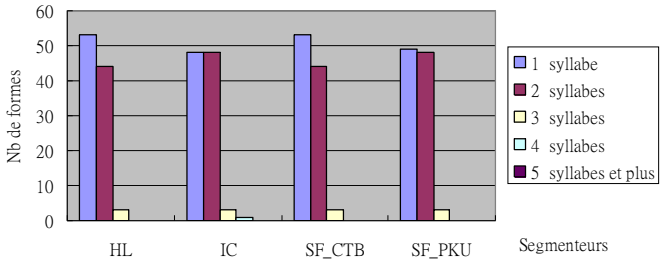

Figure 5 : Répartition des formes par segmenteur sur les cent premières formes les plus fréquentes de *Presse*

3.3.2 Le nombre de syllabes

Afin de calculer le nombre de syllabes dans l'ensemble des corpus dans chaque segmenteur, nous avons eu recours à la fonction « groupe de forme » de *Lexico3*[11]. Les groupes de formes sont des unités textuelles définies par l'utilisateur à l'aide d'outils automatiques. Cela permet de regrouper les occurrences de formes graphiques différentes mais liées par une propriété commune dans le texte, comme la flexion, la dérivation, etc.

L'analyse de la fréquence des mots comprenant plus de trois syllabes montre que plus le nombre de syllabes augmente plus la fréquence de ces mots-là diminue. Il existe donc un lien entre la fréquence d'un mot et son nombre de syllabes. Zipf (1949) parle de « principe du moindre effort » qui est que le nombre de syllabes tend à être inversement proportionnel à la fréquence d'utilisation d'un mot. Autrement dit, que les mots les plus couramment utilisés sont les plus courts. Nous avons obtenu les deux graphes présentés par les figures 6 et 7 selon ce principe du moindre effort. Les graphes montrent que la répartition de la longueur des mots correspond au principe de Zipf en faisant abstraction des dissyllabes, de plus en plus fréquents en chinois contemporain. Les quatre courbes de *Presse* sont très semblables : elles se présentent comme un graphe harmonieux. Rappelons que les monosyllabes sont plus nombreux que les dissyllabes (cf. figure 5), ce qui n'est pas le cas ici dans l'ensemble du corpus, nous avons examiné la liste des cents premières formes les plus fréquentes, elles sont les mots grammaticaux « 的 *de* (de)[12] », « 了 *le* (particule

[11] Lexico3, outil d'analyse des données textuelles, est développé par l'équipe universitaire SYLED-CLA2T (Systèmes Linguistiques Enonciation et Discours, Centre de Lexicométrie et d'Analyse Automatique des Textes). Le logiciel a été conçu par André Salem, professeur de l'Université Paris III. http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/ (page consultée le 6 janvier 2010).

[12] Le caractère chinois est suivi de la transcription *pinyin* en italique et de la traduction en français entre parenthèses.

aspectuelle) »[13], les verbes monosyllabiques « 是 *shi* (être) », « 有 *you* (avoir) », « 要 *yao* (vouloir) », les conjonctions de coordination « 与 *yu* (et) », « 和 *han* (et) », les prépositions « 向 *xiang* (à, pour) », « 在 *zai* (à) », « 对 *dui* (pour) », les pronoms « 我 *wo* (je) », « 你 *ni* (tu) », la négation adverbiale « 不 *bu* (ne … pas) », etc. Ce sont des mots courants dans un texte général, mais plutôt rares dans un texte du domaine spécialisé comme *Constitution*. Dans ce dernier, on trouve plutôt des mots pleins (des dissyllabes sont majoritaires dans le chinois contemporain, cf. figure 4), au contraire, les mots vides y sont peu fréquents.

Les courbes de *Constitution* sont dissemblables (figure 6). Les plus grandes différences sont relevées entre les monosyllabes et les dissyllabes ainsi qu'entre les trissyllabes et les pentasyllabes. Par contre, les courbes d'ICTCLAS (rose) et de SF_PKU (bleu turquoise) se superposent quasiment, les fréquences des mots pour un nombre de syllabes donné est quasi similaire.

Le choix de deux domaines différents pour chacun des deux corpus a permis de mettre en évidence l'influence du type de texte d'une part sur la répartition des mots et d'autre part sur la variation du nombre de syllabes des mots. Les textes du domaine spécifique sont plus remarquables en ce qui concerne la différence entre le nombre de syllabes par rapport aux textes généraux comme *Presse*.

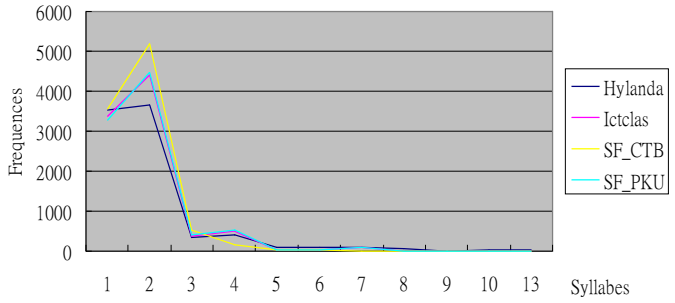

Figure 6 : Effectif des mots en fonction du nombre de syllabes dans *Constitution*

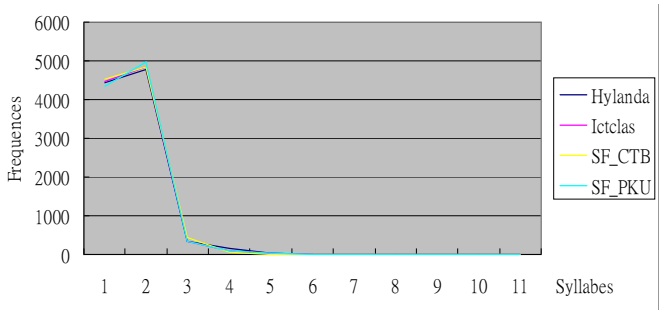

Figure 7 : Effectif des mots en fonction du nombre de syllabes dans *Presse*

3.3.3 Analyse par syllabe

Nous proposons maintenant une étude plus approfondie des sous-parties du corpus. La fonction groupe de formes de *Lexico3* permet d'acquérir une chaîne de caractères contenant le nombre de syllabes à rechercher au moyen d'une expression rationnelle[14]. La figure 8 paramétrée par termes de spécificités[15] permet de faire une synthèse de la ventilation du nombre de syllabes des mots découpés du corpus *Constitution.* La spécificité de telle ou telle syllabe en fonction d'un segmenteur donné apparaît également dans cette figure 8. Les formes de plus de cinq syllabes sont en nombre relativement élevé dans le texte segmenté par

[13] Comme les mots chinois sont invariables, les verbes n'ont aucune conjugaison. Pour exprimer le temps ou l'aspect en chinois, on emploie des particules. Il existe trois particules « 过 », « 了 », et « 着 » qui marquent respectivement l'expérience vécu, l'action accomplie et une action qui se prolonge dans la durée. Les trois particules d'aspect sont toujours précédées de verbes.

[14] Le motif de l'expression rationnelle pour trouver une syllabe (ou un caractère chinois) dans une chaîne de caractères est « ^.{2}$ », c'est-à-dire que l'on cherche une chaîne de caractères qui débute par n'importe quel caractère qui contient deux octets et qui termine cette chaîne. Un caractère chinois contient deux octets dans le codage de caractères que nous utilisons, pour chercher deux syllabes, le chiffre 2 est remplacé par le chiffre 4, etc. Les deux corpus ne contiennent pas de caractères non chinois, de plus, les ponctuations chinoises ont été retirées lors de la recherche des syllabes.

[15] La méthode de « spécificité » montre les mots les plus caractéristiques dans un corpus ou dans une partie du corpus. Cette méthode est proposée par Pierre Lafon (1980, 1984). Elle mesure « les variations de la fréquence dans un corpus découpé en parties, en fonction d'un seuil choisi par l'analyste, il indique si la fréquence observée dans telle ou telle partie peut-être considérée comme normale ou non. »

Hylanda alors que l'on en trouve très peu dans le texte de SF_CTB. Les quadrisyllabes apparus dans le texte de SF_PKU et d'ICTCLAS arrivent en second. Les dissyllabes sont en grand nombre dans le texte de SF_CTB, au contraire, il en existe un petit nombre dans le texte de Hylanda par rapport à SF_CTB. Les monosyllabes sont relativement plus nombreux dans Hylanda.

Quant à *Presse* (figure 9), les formes possédant plus de quatre syllabes sont relativement importantes dans Hylanda, spécialement pour les quadrisyllabes. Le segmenteur ICTCLAS est plus apte à détecter les formes de cinq syllabes et plus. Elles sont au contraire moins nombreuses dans SF_CTB. Les trisyllabes sont remarquables dans SF_CTB. Quant aux dissyllabes, ils sont plus nombreux dans SF_PKU, mais la proportion de monosyllabes est relativement moins importante.

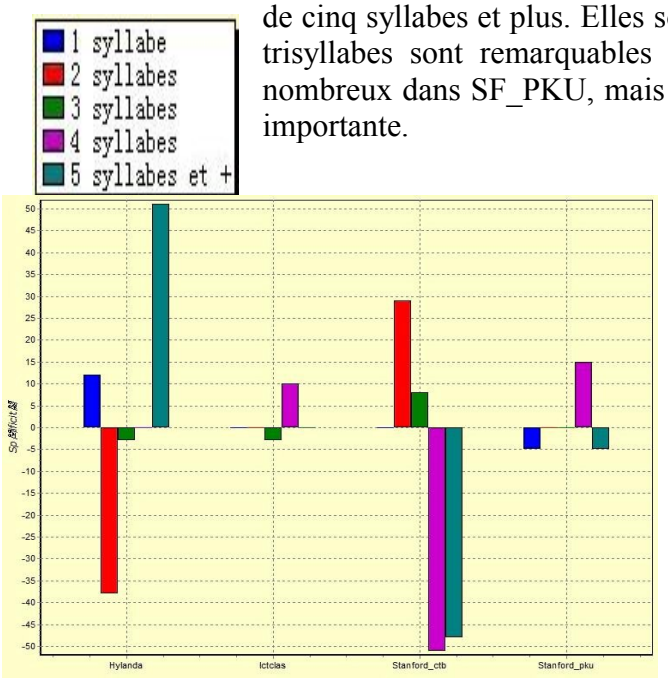

Figure 8 : Ventilation des mots d'une syllabe à plus de cinq syllabes dans *Constitution*

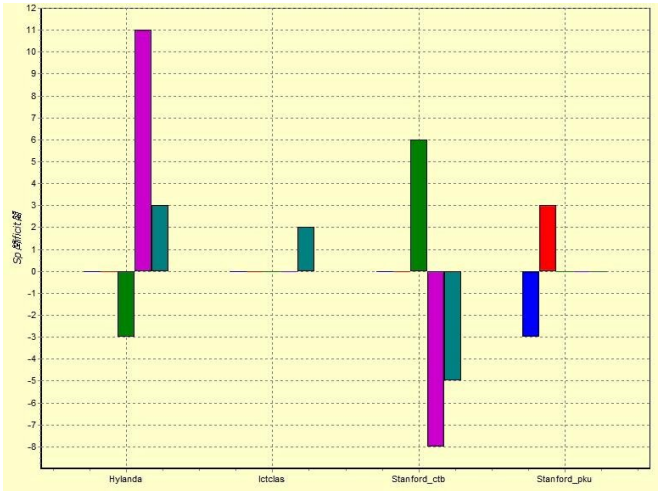

Figure 9 : Ventilation des mots d'une syllabe à plus de cinq syllabes dans *Presse*

3.3.4 Pentasyllabes

Hylanda montre une proportion très importante de polysyllabes (cinq syllabes et plus) dans les figures 8 et 9 classés par termes de spécificités. Cela nous pousse à envisager une observation plus soigneuse. La concordance fournie par *Lexico3* représente des termes spécialisés polysyllabiques pour le texte Hylanda en grand nombre dans *Constitution*, p. ex. 中华人民共和国 *zhonghua renmin gongheguo* (République populaire de Chine) ; 全国人民代表大会 *quanguo renmin daibiao dahui* (assemblée nationale populaire) 最高人民法院 *zuigao renmin fayuan* (cour suprême de justice) ; 全国人民代表大会常务委员会 *quanguo renmin daibiao dahui changwu weiyuanhui* (comité permanent de l'assemblée nationale populaire). Hylanda segmente de façon appropriée les termes spécialisés du corpus *Constitution.* Cela pourrait aider spécifiquement à la recherche de la terminologie : Hylanda paraît donc plus performant dans ce domaine que les trois autres segmenteurs.

Nous avons procédé selon la même méthode pour le segmenteur SF_CTB, étant donné qu'il a un taux très bas de quadrisyllabes et de pentasyllabes en opposition à un fort taux de dissyllabes. Les noms propres segmentés correctement par Hylanda sont ici découpés à l'intérieur de la chaîne de caractères en plusieurs formes graphiques, p. ex. la forme 中华人民共和国 (République populaire de Chine) est découpée en trois formes comme 中华_人民_共和国 (Chine_peuple_république[16]). Les formes de quatre ou cinq syllabes

[16] Le symbole tiret bas « _ » sert ici à indiquer la frontière d'une unité lexicale.

sont simplement les expressions temporelles, p. ex. 一八四 0 年 *yi ba si ling nian* (l'année 1840, littéralement, 1840 suivi du mot année) et les numérotations des articles de la Constitution, p. ex. 第一百零一 *di yi bai ling yi* (article 101, littéralement un préfixe servant à former les nombres ordinaux suivi du nombre 101). Ce ne sont pas des termes spécifiques du corpus. Par ailleurs, parmi les polysyllabes segmenté par Hylanda, certains qui sont des termes non spécifiques du domaine ont attiré notre attention. Ce sont des collocations, c'est-à-dire la combinaison de deux termes ou plus qui sont fréquemment utilisés. Par exemple, 不幸遇难者 *buxing yunanzhe* (des victimes) est composé de 不幸 *buxing* (malheur) et 遇难者 *yunanzhe* (victime). Ce phénomène pourrait être abordé dans une étude subséquente.

D'après cette étude textométrique de deux corpus en quatre segmentations, Hylanda apparaît comme un outil pertinent dans la segmentation des noms propres et plus particulièrement dans un domaine spécifique. La segmentation de SF_CTB serait plutôt fine, c'est-à-dire que la longueur moyenne des segments est plus limitée. Les deux autres segmenteurs peuvent être qualifiés d'intermédiaires, aucune spécificité n'ayant été mise en évidence.

4 Résultats

4.1 Processus d'évaluation

Notre objectif est de déterminer quel type de texte est le plus adapté pour chaque segmenteur. Afin d'évaluer nos analyses, nous avons segmenté manuellement (étant native chinoise) en se référant au Dictionnaire du Chinois Moderne (现代汉语词典 *xiandai hanyu cidian*), dictionnaire d'autorité dans la langue chinoise. Ensuite, nous avons comparé cette segmentation manuelle avec les quatre segmentations sur les deux corpus. De plus, nous avons comparé les noms propres, spécialement les noms propres de personnes en chinois et la traduction littérale des noms étrangers et également les termes spécialisés du domaine du corpus. Les formes segmentées par les outils qui sont présentes et identiques dans la version manuelle sont considérées comme pertinentes alors que les autres sont soit une segmentation différente, effectuée selon les règles de l'outil, soit une segmentation erronée. La segmentation manuelle est basée sur l'introspection de la personne native et sur sa connaissance de la langue, et privilégie le sens complet d'une forme en tenant compte du domaine du texte. Par exemple, dans le corpus *Constitution,* 中华人民共和国 (République populaire de Chine) est segmenté comme une forme lexicale au lieu d'être découpée en trois formes comme 中华_人民_共和国 (Chine_peuple_république).

4.2 Présentation des résultats

Le tableau 1 présente la précision des unités lexicales segmentées par les segmenteurs par rapport à la segmentation manuelle. La bonne performance doit être interprétée en fonction du contexte et de la segmentation manuelle effectuée. La proportion de formes segmentées pertinentes est plus importante dans le corpus général que dans le corpus spécialisé. Dans les deux corpus, la segmentation de Hylanda est la plus proche de la segmentation manuelle. SF_CTB est le plus éloigné du découpage manuel pour le corpus spécialisé *Constitution*. Au contraire, il atteint une performance assez bonne de segmentation pour le corpus général *Presse*. ICTCLAS et SF_PKU sont intermédiaires et n'ont pas de trait distinctif. Ils ont une précision assez proche pour les deux corpus.

	Constitution	*Presse*
Hylanda	93,5 %	96,7 %
ICTCLAS	92,4 %	93,1 %
SF_CTB	87,5 %	95,1 %
SF_PKU	91,9 %	93,6 %

Tableau 1 : Évaluation de la segmentation des segmenteurs pour les deux corpus (précision)

La segmentation des mots inconnus est toujours une tâche difficile dans le TAL. Nous avons également évalué ces deux corpus en comptant les noms propres qui y sont présents (tableau 2). Étant donné les textes de loi du pays comme *Constitution*, de nombreux termes d'institutions de l'État ou d'organisations ayant des termes locaux sont apparus (88,3 % pour Hylanda, 42,2 % pour ICTCLAS, 0 % pour SF_CTB, 33 % pour SF_PKU). Notons que ces termes spécialisés sont entre quatre et treize syllabes. Hylanda a une bonne performance, alors que ICTCLAS et SF_PKU ont des résultats assez faibles. En revanche, SF_CTB n'est pas du tout spécialisé dans la segmentation des textes de loi. Par contre, SF_CTB et ICTCLAS montrent un très bon résultat dans *Presse*, aussi bien pour les noms propres de personnes chinois que pour la translittération des noms étrangers[17]. Au contraire, Hylanda est plutôt faible dans la segmentation des noms propres de personne dans *Presse*. Cette évaluation manuelle met en évidence l'utilité de l'étude de ces segmenteurs. Cette expérience de petite taille sur deux corpus révèle un trait distinctif entre les segmenteurs. Il serait intéressant d'étendre notre étude à d'autres phénomènes linguistiques chinois en évaluant ces segmenteurs sur des corpus plus volumineux.

Segmenteurs	*Constitution*	*Presse*	
	Noms propres du domaine	Noms propres de personnes chinois	Noms propres de personnes étrangers
Manuel	327	63	32
Hylanda	289	14	19
ICTCLAS	138	61	28
SF_CTB	0	63	30
SF_PKU	108	18	21

Tableau 2 : Nombre de noms propres segmentés dans les deux corpus

[17] Les noms propres de personne en chinois sont composés en premier le patronyme monosyllabique en majorité, ou dissyllabiques suivi du prénom correspondant à la longueur d'une ou de deux syllabes. Ils sont formés de longueur de deux à quatre syllabes, généralement de trois syllabes. Par ailleurs, les femmes mariées portent le nom de famille de leurs maris suivi du nom de jeune fille puis du prénom, quatre syllabes sont majoritaires. La translittération des noms étrangers est interprétée soit par un patronyme seulement, soit par un prénom suivi du patronyme. Ce dernier est inséré un point médian pour séparer un prénom et un patronyme comme « Jacques René Chirac » est translittérée en chinois « 雅克·勒内·希拉克 ».

5 Conclusion

Cette comparaison de segmenteurs sur deux corpus de thèmes différents parvient à une bonne qualité d'analyse. Notre étude basée sur le nombre de syllabes du chinois a permis de distinguer un segmenteur plus performant pour les textes spécialisés, et un autre segmenteur plus pertinent pour les textes généraux. Les deux autres segmenteurs sont apparus relativement similaires, ce qui est justifié étant donné qu'ils sont fondés sur *la norme* de l'État de Chine. Leur performance est intermédiaire par rapport aux deux premiers. L'évaluation de la comparaison de ces quatre segmenteurs au moyen de la segmentation manuelle affirme que la méthodologie est pertinente dans le cadre de l'étude.

L'étude sur le nombre de syllabes en chinois ouvre des portes dans la recherche en textométrie sur la comparaison des outils de segmentation. Une étude approfondie sur la variation du nombre de syllabes pourrait déterminer si celle-ci est liée à la linguistique chinoise.

L'exploration textométrique des textes chinois a déjà franchi certains obstacles dus à la complexité du système d'écriture de la langue chinoise. Les résultats favorables de cette étude nous amènent à approfondir le phénomène de collocation et d'entités nommées dans la segmentation et à nous demander si la catégorie grammaticale est un trait pertinent dans la segmentation de la langue.

Références

ALLETON, V. (1994). Le nombre de syllabes d'un mot est-il pertinent en chinois contemporain ? *Cahiers de linguistique - Asie orientale*, *23*(1), 5-11.

FLEURY, S., ZIMINA, M. (2009). mkAlign, Manuel d'utilisation. EA2290 SYLED/CLA2T Université Sorbonne Nouvelle - Paris 3.

HUANG, Z., YANG, J. (黃志強, 楊劍橋) (1990). Lun hanyu shuangyinjiehua de yuanyin 論漢語詞彙雙音節化的原因 (Étude du dissyllabisme des mots chinois). *復旦學報 (社會科學版) Fudan Journal (Social Sciences Edition)*, (1), 98-101.

LAFON, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, (1), 127-165.

LAFON, P. (1984). *Dépouillements et statistiques en lexicométrie*. Travaux de linguistique quantitative. Genève : Slatkine.

LIANG, N. (梁南元) (1984). Shumian hanyu de zidong fenci yu yige zidong fenci xitong - CDWS 書面漢語的自動分詞與一個自動分詞系統 - CDWS (Written Chinese automatic distinguishing word & a automatic distinguishing words system - CDWS). *北京航空航天大學學報 (Journal of Beijing University of Aeronautics and Astronautics)*, (4), 97-104.

TSENG, H., CHANG, P., ANDREW, G., JURAFSKY, D., MANNING, C. (2005). A conditional random field word segmenter. In *Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, 168-171.

WANG, H. (王化鹏) (2000). Lun xiandai hanyuci de shuangyinjiehua ji qi fazhan guilü 論現代漢語詞的雙音節化及其發展規律 (On the disyllabe superiority of modern Chinese and its developing laws). *北方論叢 (The Northern Forum)*, *164*(6), 120-125.

WANG, L. (王力) (1984). Zhongguo yufa *中國語法 (Grammaire chinoise)*. Shandong : Shandong chubanshe.

XIA, F. (2000). *The segmentation guidelines for the Penn Chinese Treebank (3.0)* (Technical Report IRCS Report 00-06). University of Pennsylvania.

YU, S., ZHU, X., DUAN, H. (俞士汶, 朱學鋒, 段慧明) (2000). Daguimo xiandai hanyu biaozhu yuliaoku de jiagong guifan 大規模現代漢語標注語料庫的加工規範 (The guideline for segmentation and part of speech tagging on very large scale corpus of contemporary Chinese). *中文信息學報 (Journal of Chinese Information Processing)*, (6), 58-64.

ZHANG, H., LIU, Q., CHENG, X., ZHANG, H., YU, H. (2003). Chinese lexical analysis using hierarchical hidden markov model. In *Proceeding of the Second SIGHAN Workshop on Chinese Language Processing*, 63-70.

ZHU, D. (朱德熙) (1982). Yufa jiangyi *語法講義 (Lectures on grammar)*. Pékin : Shangwu yinshu guan.

ZIPF, G. K. (1949). *Human behavior and the principle of least effort : an introduction to human ecology*. Cambridge MA : Addison-Wesley.