inalco
Institut national
des langues
et civilisations orientales

inalco
ERTIM
Equipe de recherche
textes, informatique,
multilinguisme

Comparaison des résultats de trois outils de segmentation en mots du Chinois sur un corpus issu de forums sur le diabète

Master

Traitement Automatique des Langues

Parcours : Ingénierie Multilingue

par

Catherine THOMAS

Directeur de mémoire :

François Stuck

Encadrant :

Driss Sadoun

Année universitaire 2015/2016

Remerciements

Mes remerciements vont d'abord à l'équipe des enseignants du Master Plurital cohabilité par Paris3 Sorbonne Nouvelle, Paris X et INALCO. Outre leurs qualités pédagogiques certaines et leur disponibilité, j'ai particulièrement apprécié leur capacité à constituer et encadrer des promotions d'étudiants d'une extrême diversité. Bien mieux, les enseignants s'appuient sur cette hétérogénéité des profils pour aligner leur niveau d'exigence au point le plus haut et stimuler le travail en équipe, la qualité et la fréquence des échanges entre les étudiants. Les enseignants donnent beaucoup plus que des cours, je les remercie pour leur bienveillance et leur engagement.

Merci encore à mes camarades d'études, ceux de ma promotion et ceux des années précédentes et même ceux de l'année suivante avec qui j'ai été en contact. A un moment ou un autre, en direct ou par les informations mises à disposition sur leurs blogs ou par leurs conseils, ils m'ont aidée à progresser dans une matière, à passer un cap, à résoudre un problème, à ouvrir une porte sur un domaine. Merci de leur générosité, merci pour cet esprit d'entraide jamais démenti. Merci aussi pour les moments de détente, les soirées et les fêtes, avec et sans prétexte.

Enfin, à la suite de mon stage sur le projet MultiTal, je veux remercier les membres du laboratoire ERTIM, ainsi que les personnels de la rue de Lille. Plus particulièrement, je tiens à remercier les membres du projet MultiTal, qui m'ont immédiatement intégrée à la réflexion sur le projet et apporté tout le support dont j'ai eu besoin pour mener à bien les tâches confiées. Mais je n'oublie pas les autres membres du laboratoire, y compris les doctorants, toujours disponibles pour répondre à mes questions, me passer une documentation qui m'avait échappée, voire servir de testeur volontaire sur des outils ou des méthodes. Merci à tous enfin pour la joyeuse convivialité qui a toujours accompagné la rigueur du travail scientifique.

Résumé

Ce travail rend compte de la comparaison entre les résultats de trois outils de segmentation automatique en mots du Chinois. L'intérêt est de montrer comment les trois outils diffèrent dans leur interprétation de particularités linguistiques de la langue chinoise, entre eux et par rapport à une segmentation humaine. Pour cela un corpus brut a été constitué par extraction d'un forum chinois traitant du diabète ; ce corpus de taille limitée contient des tournures et du vocabulaire spécifiques à la langue chinoise et au domaine. L'analyse de quelques exemples linguistiquement représentatifs montre qu'aucun des outils de segmentation n'est entièrement fiable et que les erreurs entre eux ne sont pas homogènes. En conclusion, une démarche pragmatique est proposée pour guider le choix d'un outil de segmentation en mots comme préalable aux traitements automatiques à réaliser sur des textes chinois.

Mots clés : comparaison, segmentation, chinois, mandarin, forum, diabète, choix outil de TAL, démarche pragmatique

Table des matières

1 Introduction

1.1 Pourquoi travailler sur la segmentation du Chinois ?

Lors de sa soutenance de mémoire de M2, une doctorante chinoise a présenté une des difficultés rencontrée : l'outil qu'elle avait utilisé pour segmenter en mots ses textes chinois n'avait pas reconnu un des mots, constitué de deux caractères, dont elle voulait analyser l'usage. Quelques mois plus tard, nous nous rencontrons à nouveau et j'explique ma participation au projet MultiTal (« resume-multital.pdf » 2016):

- Je dois chercher, tester et référencer des outils de TAL pour la langue chinoise

- Excellent ! Alors qu'est ce que tu me conseilles comme segmenteur de mots pour le chinois ? Lequel est le meilleur? J'en voudrais un qui ne fasse pas d'erreur !

- …. ? heu ?

Impossible pour moi de répondre à cette question, qui en entraîne de nombreuses autres : quelle segmentation pour quel usage, faut-il prendre une perspective de linguiste ou de traitement automatique, comment définir une erreur de segmentation ?

A partir de cette réflexion, j'ai décidé de comparer les résultats obtenus par trois outils de TAL du chinois, pas pour trouver "le meilleur" mais pour pouvoir aider d'autres talistes à choisir en connaissance de cause selon leur objectif. De plus, pour rester cohérent avec l'objectif du projet Multital de simplifier au maximum la mise en œuvre des outils pour un utilisateur final dépourvu de compétences informatiques approfondies, j'ai mené l'analyse avec des moyens très accessibles et davantage dans une perspective linguistique qu'informatique.

1.2 Repérer les comportements différents de trois outils

Pour opérer cette comparaison, pas question donc de calculer rappel, précision et f-mesure des trois outils sur un corpus annoté standardisé. En effet ces calculs ont déjà été faits par les développeurs des outils et surtout, ils n'apporteront pas de réponse aux problématiques d'analyse des textes contenant des particularités linguistiques telles que vocabulaires de spécialité, néologismes ou éléments dialogiques. C'est pourquoi je vais tenter de comparer les segmentations réalisées par les outils sur un corpus non étiqueté constitué de posts de forums chinois de diabétiques. Mon objectif est de repérer les types de comportement de chaque segmenteur sur des cas où la segmentation est ambiguë ou difficile à automatiser. Ce seront des indices qui pourront guider le choix du segmenteur en fonction du type d'analyse que l'utilisateur voudra réaliser.

1.3 Convention de notation

Dans la suite du document, chaque fois qu'un caractère chinois est utilisé, il sera suivi de sa transcription en pinyin, le système de romanisation officiel de la république populaire de Chine, et de sa signification principale. Pour davantage de lisibilité, le pinyin est noté sans les tons de prononciation ; le caractère, sa transcription et sa signification sont séparés par des barres obliques, la transcription et la signification sont en italique.

Exemple : Le nom en chinois de l'INALCO contient 8 caractères, ce qui est de bon augure ; en 5 et 6ème positions on trouve le mot dissyllabique 文化 */ wenhua / civilisation*.

2 Problématique de la segmentation des mots en Chinois

2.1 La segmentation des mots en chinois : une question non tranchée

Contrairement au Français ou à l'Anglais, les mots des textes écrits en chinois ne sont pas séparés par des espaces, les caractères sont tous écrits les uns à côté des autres. Au début des études de chinois à l'Inalco, les textes du manuel débutant (Rabut, Yongyi, et Hong 2012) ne sont pas vraiment écrits en Chinois : pour ne pas dérouter les étudiants français, chaque mot est séparé par un espace. Assez rapidement et sans préavis, les espaces disparaissent des textes d'apprentissage et les étudiants ne s'en aperçoivent pas immédiatement, car ils ont pris l'habitude de reconnaître les mots au fil de la lecture. Mais ce mécanisme de compréhension des textes n'est pas encore à la portée de l'analyse réalisée par les outils de TAL. En effet, ainsi que le note d'entrée (Magistry 2012) "Pour la plupart des langues utilisant l'alphabet latin, un découpage sur les espaces est une bonne approximation d'une segmentation en unités lexicales". Mais, pour le chinois, "l'étape de tokenisation préalable à beaucoup de systèmes d'analyse automatique est de ce fait plus délicate. […] on parle d'étape de segmentation en mots." De plus, s'il existe des standards de segmentation du chinois, il n'existe pas forcément de consensus sur une méthode. Certes, la République Populaire de Chine a publié en 1993 "la norme de la segmentation des mots chinois contemporains pour le traitement informatique" (刘源, 谭强, et 沈旭昆 1994) mais outre que les principes proposés sont difficiles à mettre en œuvre, d'autres travaux sont venus depuis apporter de nouveaux regards sur la question de la segmentation. En particulier pour la segmentation de textes issus des médias sociaux, (Yushi et Zheng 2015) utilisent conjointement un modèle statistique de CRF (champs aléatoires conditionnels) et l'algorithme MMSEG (Maximum Matching Segmentation) de (Chih-Hao 2000) qui s'appuie sur un lexique.

Après avoir rappelé quelques notions clés sur l'écriture et la grammaire chinoise, je présente dans ce chapitre les recommandations de la norme de segmentation du Chinois ISO 24614-2 (Organisation internationale de normalisation, Comité technique ISO/TC 37, et Sous-comité SC 4 2010) enfin, je présente quelques travaux de recherche en matière de méthodes non supervisées.

2.2 Notions clés sur la langue chinoise et son écriture

En plus de ne pas ménager d'espaces entre les caractères, l'écriture chinoise a quelques particularités qu'il est bon de connaître et de garder en tête avant d'aborder son traitement automatique.

2.2.1 Une langue, deux formes d'écriture

Les réflexions sur la simplification de l'écriture et de la langue ont commencé dès la fin du XIX° siècle en Chine, mais l'instabilité politique qui a perduré pendant la première moitié du XX°siècle n'a pas permis d'établir et de diffuser une méthode. En 1958, le gouvernement de Chine Populaire a

introduit une réforme de l'écriture des caractères chinois visant à simplifier les graphismes utilisés ; ce sont environ 2000 caractères qui ont ainsi reçu une forme simplifiée officielle. Pour d'évidentes raisons politiques, cette réforme n'a pas été diffusée à Taiwan, Hong-Kong et Singapour. Les caractères officiels utilisés en Chine continentale sont appelés *caractères simplifiés*, les caractères anciens toujours utilisés ailleurs sont dits *caractères traditionnels*.

Forme traditionnelle	Forme simplifiée	Transcription / Sens
幾	几	Ji / combien, quelques
說	说	Shuo / parler, dire
麼	么	Me / quoi

Table 1: Exemples de caractères traditionnels et simplifiés

Depuis les années 1980 et la politique de réforme et d'ouverture, il y a davantage d'échanges entre la Chine Populaire et les autres Chine : on assiste en parallèle au retour des caractères traditionnels en Chine Populaire et à l'adoption des caractères simplifiés à l'extérieur. Ces deux graphies peuvent ainsi cohabiter dans un corpus. Il existe des outils de conversion entre les deux graphies, mais ils n'ont pas une fiabilité totale car il n'y pas de bijection entre les deux jeux de caractères. Pour traiter un corpus contenant les deux formes de caractères, on peut soit réaliser une conversion avant le traitement, soit utiliser des outils qui savent traiter les deux jeux de caractères.

2.2.2 Particularités de la ponctuation asiatique

La ponctuation asiatique utilise des signes différents de la ponctuation occidentale. Curieusement, il arrive que des systèmes qui reconnaissent les caractères chinois butent sur la ponctuation spécifique ; dans ce cas, il faut normaliser les corpus en remplaçant la ponctuation chinoise par des équivalents occidentaux. Notons ci-dessous les signes spécifiques au chinois [1] , sans oublier que les signes communs comme " :" deux-points ou " ;" peuvent différer de certaines langues occidentales car ils ne sont pas précédés par un espace insécable.

1 http://chine.in/guide/ponctuation_1984.html

Nom	Signe typographique	Signification
句号 jù hào	。	Point final d'une phrase
顿号 dùn hào	、	Séparation entre les termes parallèles et de même importance dans une phrase, énumération.
引号 yǐn hào	“ ” ‘ ' ㄩ ㄣ ㄴ ㄱ	Citation ou dialogue, ironie
省略号 shěng luè hào	……	Ellipse laissant la fin de la phrase en sous-entendu. Six points qui prend la place de deux caractères
破折号 pò zhé hào	——	1. Introduit une explication ou interprétation, fonction quasi-parenthèse. 2. Le développement ou l'opposition dans la logique
书名号 shū míng hào	《》 〈〉	Nom d'une publication, d'un ouvrage
间隔号 jiàn gé hào	·	1. Séparation entre les chiffres exprimant le mois et la date, notamment dans les événements historique 2. Séparation entre le nom et le prénom dans la transcription d'un nom non sinophone

Table 2: Marques de ponctuation spécifiques au Chinois

2.2.3 Le chinois, une langue sans flexion ni conjugaison

Il n'y a ni flexion ni conjugaison en Chinois : dans les "Éléments fondamentaux de la phrase chinoise" (Darrobers 1998) , les auteurs soulignent "l'invariabilité radicale des éléments de la phrase". Du point de vue du TAL, cette particularité peut se révéler un avantage, puisqu'elle supprime les étapes de lexicalisation et normalisation de la casse. Mais elle entraîne aussi la nécessité de penser autrement l'analyse des phrases.

Le nombre peut être marqué sur les noms d'êtres animés par l'ajout d'un suffixe (们 / *men*) sans que ce soit systématique ou bien par des spécificatifs comme 点 / *dian / quelques* ou 些 / *xie / plusieurs*. Concernant les verbes, le temps, le mode et l'aspect sont exprimés soit par des expression temporelles (昨天晚上 / *zuotianwanshang / hier soir*) situées avant le verbe dans la phrase, soit par des particules (过 / *guo / expérience réalisée et terminée,* 了 / *le / action accomplie*) placées après le verbe ou à la fin de la phrase, soit par des procédés syntaxiques comme la réduplication, expliquée ci dessous au chapitre 2.2.7.

2.2.4 L'ordre strict des mots de la phrase chinoise

La phrase chinoise obéit assez strictement à l'ordre Sujet-Verbe-Objet. Cependant, il peut arriver que la phrase commence par un thème et que le sujet soit alors déplacé ou supprimé si le contexte le permet. Les compléments circonstanciels se placent généralement avant le verbe.

Contrairement au français, l'ordre déterminant + déterminé est strict, on trouvera notamment toujours l'adjectif avant le nom qu'il qualifie.

2.2.5 Les noms propres sont difficiles à repérer

Les sinogrammes n'ont pas de variante graphique correspondant aux majuscules / minuscules utilisées dans les langues européennes. Les noms propres ne sont donc pas marqués par un signe typographique distinctif. Un nom chinois complet se compose d'abord du nom de famille (en général un seul caractère) suivi du nom personnel (en général deux caractères). Les noms de famille chinois sont en nombre relativement réduit et peuvent avoir en dehors de leur contexte de nom de famille une signification qui les rend plus ou moins simples à repérer ; les prénoms sont laissés à l'appréciation des parents et trahissent parfois l'époque de la naissance.

Sinogramme	Transcription	Sens
赵	Zhao	Nom d'un état de la dynastie des Zhou (antiquité chinoise)
邓	Deng	Nom de famille
毛	Mao	Poil, impur, moisi, petit
黄	Huang	Jaune, pornographique
高	Gao	Grand, élevé
建民	Jianmin (prénom)	Édification du peuple
安安	Anan (prénom)	Serein, apaisé, en sécurité
一世	Yishi (prénom) – Isseï (Japon)	Toute la vie, génération, Moi, premier du nom

Table 3: Exemples de noms propres chinois

Les noms étrangers sont transcrits phonétiquement en caractères chinois et repérés par une convention typographique : leurs composants sont réunis par le signe ·, un point placé au milieu de la ligne d'écriture. Cependant, même s'il existe des règles pour les toponymes et des dictionnaires de correspondance entre les sons d'une langue étrangère et les caractères chinois à utiliser pour les représenter, (La Robertie 2005) souligne que "seul le contexte et la connaissance que le lecteur a de la langue lui permet de repérer un nom de personne". Comme de plus existe énormément de caractères homophones, plusieurs transcriptions d'un même nom étranger peuvent cohabiter avnt qu'une transcription unique s'impose.

Sinogrammes	Transcription(s)	Nom d'origine
维克多·雨果	Weikeduo·Yuguo	Victor Hugo
尼古拉·萨科奇 尼古拉·萨科齐	Nigula·Sakeqi	Nicolas Sarkozy
奥巴马	Aobama	Obama

Table 4: Exemples de transcription de noms propres étrangers

2.2.6 L'écriture des nombres – les classificateurs

Les textes chinois peuvent contenir des expressions numériques sous deux formes, en chiffres arabes ou en caractères chinois ; par exemple :

65 六十五 / *liu shi wu / six dix cinq*

2016 二零一六年 / *er ling yi liu nian / deux zéro un six année*

5 % 百分之五 / *bai fen zhi wu / cent diviser parmi cinq*

Il est important d'identifier les expressions numériques en tant que telles, sans les élémentariser.

A la différence du Français, le Chinois ne permet pas de faire dénombrer des éléments en écrivant simplement numéral + nom : il faut obligatoirement utiliser un classificateur entre le nombre et le nom ; certains classificateurs n'échappent pas à la polysémie et doivent être identifiés comme tels pour être segmentés correctement.

Exemple : 本 / ben

Classificateur : 三 | **本** | 书 / san ben shu / trois livres (ouvrages imprimés)

Adverbe : 我 |**本** | 不喜欢喝茶 / wo ben bu xi huan he cha /au début je n'aimais pas boire du thé

Nom : 目有 |**本** | / mu you ben / les arbres ont des racines

Formant d'un nom : 独到的 | 剧**本**| / du dao de ju ben / un scénario original

2.2.7 La formation des mots du chinois – la réduplication

Dans son ouvrage Parlons Chinois (Drocourt et Peyraube 2007) , Zhitang Yang-Drocourt présente en détails les formes et les procédés de construction des mots du chinois. En particulier, elle montre comment les mots chinois évoluent historiquement vers le dissyllabisme (soit graphiquement deux caractères) et comment l'allongement des mots permet d'en réduire la polysémie. Ainsi par exemple

Caractère	Transcription	Sens des formants	Sens du mot
生	Sheng		Naître, produire, cru, jeune homme
学生	Xuesheng	Étudier + jeune homme	Élève, écolier.ère, lycéen.ne, étudiant.e
大学生	Daxuesheng	Grand+élève	Étudiant.e
北大学生	Beidaxuesheng	Nord+Étudiant.e	Étudiant.e de l'université de Pékin

Table 5: Allongement des mots et réduction de la polysémie

On voit déjà dans l'exemple ci-dessus l'amorce d'un autre problème d'arbitrage pour la segmentation : chacun des 3 caractères qui forment le mot "étudiant.e" peut constituer aussi, en fonction de son contexte, un mot à part entière et figure dans les dictionnaires.

A l'inverse, un mot jugé trop long sera souvent abrégé, comme 超级市场 */ chao ji shi chang / supermarché* couramment abrégé en 超市 */ chao shi / supermarché.*

Les mots en chinois sont formés par de nombreux procédés, comme la composition, l'affixation, ou la réduplication, détaillée ci-après. Le point commun de ces procédés est de ne pas être marqué par des conjonctions ou autres repères syntaxiques ou typographiques. Ces mêmes procédés s'appliquent à plusieurs catégories syntaxiques : on peut ainsi former des noms, des adjectifs, des adverbes ou des verbes. Suivant les cas, on peut accoler des synonymes, des antonymes ou des mots appartenant au même champ sémantique, ou bien spécifier un morphème général. Le composé obtenu est stable, lexicalisé et ce n'est généralement pas en analysant séparément le sens des formants que l'on peut déduire le sens du composé. Ci-dessous, quelques exemples.

Catégorie	Caractères	Transcription	Sens des formants	Sens du disyllabe
NOM	朋友	peng + you	ami + amical	ami
AUX	应该	ying + gai	devoir+devoir	devoir, falloir
ADJ	沉重	chen + zhong	lourd + lourd	lourd
VERBE	买卖	mai + mai	acheter + vendre	faire des affaires
PRONOM	多少	duo + shao	beaucoup + peu	combien
NOM	黄牛	huang + niu	jaune + bœuf	revendeur clandestin de tickets
NOM	看法	kan + fa	voir + façon	point de vue
NOM	书店	shu + dian	livre + magasin	librairie

Tableau 6: Exemples de constructions dissyllabiques de mots chinois

Un procédé particulier : la réduplication

Certains verbes et certains adjectifs, mono ou dissyllabiques peuvent être redoublés pour exprimer des nuances différentes du même mot. Le redoublement du verbe peut par exemple exprimer qu'une action se fait aisément, de manière brève ou à titre d'essai (Darrobers 1998) . Le redoublement du verbe monosyllabique peut se faire directement ou en intercalant le caractère 一 / *yi / un* entre les deux, comme dans le quatrième exemple ci-dessous.

Catégorie	Caractères	Transcription / sens	Réduplication	Sens du dissyllabe
ADJ	干净	gan jing / propre	干干净净	bien propre
ADJ	红	hong / rouge	红红	tout rouge
VERBE	看	kan / regarder	看看	jeter un coup d'œil
VERBE	尝	chang / goûter	尝一尝	goûter un petit peu
VERBE	休息	xiuxi / se reposer	休息休息	prendre un peu de repos

Tableau 7: Réduplication des adjectifs et des verbes en Chinois

2.2.8 Les expressions proverbiales – formules quadrisyllabiques

Les chinois utilisent couramment des expressions figées issues de la tradition historique et littéraires, parfois appelées avec un peu d'ironie "proverbes chinois", ce sont les 成语 */ chengyu / expressions toutes faites.* Dans son ouvrage "Parlons Chinois", Zhitang Drocourt propose de les nommer "formules quadrisyllabiques"(Drocourt et Peyraube 2007) car elles sont en majorité formées de quatre syllabes. Le dictionnaire de Chengyu destiné aux écoliers (于明善, 杨东, et 说词解字辞书研究中心 2012) contient 5500 entrées, le dictionnaire en ligne de Chine Nouvelle[2] revendique 30000 entrées. Ces expressions sont entièrement lexicalisées et ne peuvent pas être divisées ; elles peuvent servir de sujet, de déterminant du nom ou de complément verbal. Leur sens n'est pas toujours déductible de leurs formants. Ci dessous quelques exemples :

Expression	Transcription	Littéralement	Sens
走马看花	zou ma kan hua	Aller à cheval regarder fleur	Prendre une décision à la va vite
离经叛道	li jing pan dao	S'éloigner de la règle suivre la rébellion	Déviance

Tableau 8: Formules quadrisyllabiques - exemples

2 http://www.chine-nouvelle.com/chinois/chengyu/dictionnaire

Ces expressions ne sont pas réservées à une élite intellectuelle : tous les Chinois les utilisent abondamment et elles portent une sémantique puissante.

2.3 La norme ISO – un idéal à viser

Pour la segmentation des textes écrits en Chinois, Japonais et Coréen, il existe une recommandation ISO particulière, côtée ISO / CD 24614-1 et commentée dans un article de (Choi et al. 2009) . Les règles de la segmentation du chinois y sont décrites pour chacun des 13 types de mots suivants :

> nom, verbe, adjectif, pronom, numéral, mot de mesure, adverbe, préposition, conjonction, auxiliaire, particules modales, interjections, onomatopées.

Les règles proposées sont généralement simples et cohérentes avec les habitudes grammaticales chinoises. Nous allons voir en détail quelques unes de ces règles, afin d'être en mesure de vérifier comment les segmenteurs les implémentent.

2.3.1 Les noms

Parmi ceux qui ne doivent pas être séparés, on trouve des unités lexicales comme des noms composés de deux caractères (ex : 牛肉 / *niu rou / boeuf viande* => bœuf) présents comme unité dans la plupart des dictionnaires, ainsi que des composition à base de préfixes (阿哥 / *a ge / grand frère*, 非对称 / *fei dui chen / asymétrique*) et à base des suffixes nominaux 子 / *zi*, 儿 / *er*, 头 / *tou*, 家 / *jia*, 手 / *shou*, 性 / *xing* par exemple. La liste des préfixes et des suffixes est suffisamment stable en chinois pour permettre le repérage.

Les dates ou expressions temporelles ne doivent pas être séparées, par ex 三月 / *san yue / littéralement 3+mois / mars* ou bien 6 时 45 分 / *6 shi 45 fen / six heure quarante cinq*.

2.3.2 Les verbes

Pour les verbes, les règles sont plus variées car les constructions aussi peuvent varier.

On considère comme une seule unité lexicale les formes verbales stables de type Verbe + Objet ou verbe + complément comme 见面 / *jian mian / rencontrer* ou bien 抽烟 / *chou yan / fumer ou encore* 吃饭 / chi fan / manger (littéralement manger riz). Par contre, il faut détacher 吃 | 鱼 /chi yu / manger du poisson en deux unités lexicales. En effet, ce type de verbe ne peut se construire seul, il prend soit un objet générique, soit l'objet réel de l'action.

Toutes les **formes verbales négatives** sont segmentées en négation | verbe, même dans la forme particulière verbe + négation +verbe identique, qui donne 3 unités lexicales.

Les auxiliaires, les particules d'aspects, les résultatifs (particules qui indiquent le résultat d'une action) doivent être séparées du verbe et constituent des unités autonomes.

Les règles plus complexes à appliquer sont l'identification des formes répétées lexicalisées à ne pas segmenter : 看看 / *kan kan / jeter un coup d'oeil* ou bien 来来往往 / *lai lai wang wang / aller et venir.* Mais d'autres formes propres au chinois doivent être divisées en plusieurs unités lors de la segmentation, comme expliqué dans le tableau ci-dessous.

Construction verbale	Exemple	Sens	Segmentation recommandée
juxtaposition directe verbe monosyllabique	想想 / xiang xiang	réfléchir un peu	想 | 想 2 unités lexicales
juxtaposition directe verbe dissyllabique	休息休息 / xiuxi xiuxi	se reposer un peu	休息 | 休息 2 unités lexicales
insertion du caractère 一 yi	等一等 / deng yi deng	attendre un moment	等 | 一 | 等 3 unités lexicales
insertion du caractère 了 le introduit l'aspect accompli	看了看 / kan le kan	avoir jeté un coup d'œil	看 | 了 | 看 3 unités lexicales

Table 9: Norme ISO - exemples de formes verbales à segmenter

2.3.3 Les numéraux

Les nombres exprimés en caractères chinois ne doivent pas être séparés : 七万五千二白三十四 */ qi wan wu qian er bai san shi si / 75234*, de même les expressions des pourcentage ou des fractions, comme 三分之一 */ san fen zhi yi / un tiers*.

Plus compliqué, les expressions d'approximation qui sont constituées de deux nombres juxtaposés, par exemple 五六公里 */ wu liu gong li / 5 ou 6 km*.

Enfin, les numéraux doivent être séparés des classificateurs, de la particule 第 */ di* indiquant un ordinal et des adverbes placés après un ordinal pour indiquer une quantité approximative.

2.3.4 Les autres mots

D'une manière générale, les prépositions, les conjonctions, les adverbes, les particules modales et les onomatopées doivent être isolées. De même que les mots outils, et les particules verbales aspectuelles doivent être segmentées séparément.

2.4 Les méthodes non supervisées

Les méthodes de segmentation supervisées s'appuient sur des dictionnaires et des ressources annotées pour entraîner des systèmes de segmentation en mots : c'est efficace mais se révèle insuffisant et même une source d'erreurs dès qu'on traite des textes peu normés et producteurs de néologismes comme les forums internet ou les textes de spécialités. De plus, compte tenu de l'absence de flexions et de la fluidité des catégories morpho syntaxiques du Chinois, il est primordial de ne pas négliger l'analyse des séquences de caractères dans leur ensemble pour en proposer une segmentation. C'est pourquoi de nombreux travaux explorent des approches non supervisées, sans entraînement préalable sur des corpus annotés.

Un lecteur humain opère une segmentation sur un texte chinois en utilisant bien plus que ses connaissances lexicales : c'est ce processus cognitif que Wu (Wu 2011) analyse dans "A Cognitive Model of Chinese Word Segmentation for Machine Translation" et souligne que "la stratégie de segmentation en mots la plus fréquente est de trouver des informations sémantiques et contextuelles sans se restreindre au contexte immédiat". Sa proposition méthodologique est donc d'imiter le processus cognitif de segmentation.

(Huang et al. 2007) expérimentent une méthode qui se passe entièrement de ressources lexicales. S'appuyant sur une classification des caractères en début-de-mot, fin-de-mot, milieu-de-mot, ils calculent la probabilité pour chaque limite de caractère d'être aussi une limite de mot. Leurs premiers résultats semblent prometteurs mais l'implémentation n'est pas disponible.

Plus radicalement, (Magistry 2012) présente un système non spécifique au mandarin, ni même à la séparation en mots, mais adaptable à d'autres langues ou d'autres unités. Suivant l'hypothèse de (Harris 1955) , il décide de "reformuler le problème de la segmentation d'une phrase comme la recherche du découpage qui maximise l'autonomie des mots qu'il délimite". L'autonomie de chaque n-gramme est donc calculée à gauche et à droite, puis à l'intérieur d'une phrase ou d'une séquence de caractères, il recherche la maximisation des autonomies. On voit que cette approche, en incluant du contexte, tend justement à imiter le processus cognitif. Développant ces travaux, (Chen, Chang, et Pei 2014) essayent, toujours avec une classification des caractères en termes de limites de mots, de combiner les probabilités d'un modèle de Markov Caché avec un modèle de Dirichlet hiérarchique, qui permet de calculer la probabilité d'une distribution.

2.5 Il s'agit de comparer via une analyse linguistique

L'objectif de mon mémoire n'est pas d'évaluer chaque outil en terme de performances : inutile de chercher ici des calculs de précision, rappel et F-mesure pour évaluer les outils par rapport à … à quoi au fait ? Comme on l'a vu ci-dessus, les "standards" de segmentation du chinois sont controversés et le resteront longtemps. Dans les conférences internationales, comme SIGHAN[3] qui est spécialisée pour le traitement du Chinois, un gold-standard est proposé par les organisateurs et il sert d'étalon à tous les systèmes présentés. Me situant en dehors de ce contexte particulier, je ne parle pas d'évaluation, mais de comparaison. Comme me l'a fait justement remarquer un collègue "Comparer trois objets différents, même pour le cerveau humain c'est plutôt difficile à concevoir comme démarche. Alors concrètement, tu vas faire comment ?"

Il avait tout à fait raison ce collègue et après quelques tâtonnements avec des outils basiques comme l'instruction bash "diff", MK Align[4] et LFAlign[5] (qui prennent tous exclusivement et logiquement 2 fichiers en entrée et pas 3), afin de bien constater par moi même qu'on ne peut pas aligner trois textes, j'ai adopté une approche différente. Au lieu de rechercher une exhaustivité inatteignable, j'ai choisi de repérer et d'analyser dans le corpus des différences significatives de segmentation, et les conséquences qu'elles auront sur les tâches de TAL qui peuvent suivre la segmentation : étiquetage morpho syntaxique, traduction, reconnaissance des entités nommées, analyse syntaxique, voire sur des outils comme la sélection au double click. De plus, si je n'ai pas de corpus annoté de référence, j'ai en revanche à ma disposition la norme ISO 24614-1 qui donne des règles précises pour chaque

3 http://sighan.org/
4 http://www.tal.univ-paris3.fr/mkAlign/
5 https://sourceforge.net/projects/aligner/

catégorie de mots : là encore, j'ai recherché comment cette règle est, ou non, implémentée par chacun des outils testés.

3 Corpus et outils

3.1 Description du Corpus

3.1.1 Origine du corpus

Le corpus est constitué de 1000 lignes de posts en chinois provenant du forum http://bbs.tnbz.com/forum récupérés le 17 août 2016 par un script Perl. Il s'agit d'un site de Chine Populaire, les participants s'expriment en chinois mandarin et écrivent en caractères simplifiés. D'après Alexa[6], qui propose un classement des sites par pays et par fréquentation, ce site est fréquenté à près de 98 % par des internautes de Chine Populaire. Le forum choisi est consacré au diabète, les participants échangent donc sur leur état de santé, leurs traitements et leur hygiène de vie.

La seule information récupérée est le texte brut de la contribution : ni la date, ni le pseudo ni le fil de discussion ne sont utilisés pour travailler purement sur la segmentation.

Le corpus contient 147134 caractères, il est encodé en UTF-8.

Il s'agit d'un corpus de texte brut, sans annotation ni analyse déjà réalisée. En effet, l'objectif est de faire une analyse linguistique comparative des résultats des trois différents outils.

3.1.2 Particularités linguistiques du corpus choisi

Des forums de discussion

Les règles de politesse des forums requièrent l'utilisation de formules de salutation, de politesse et de remerciement, telles que "bonjour" rendu par deux caractères en chinois : 你好 */ ni hao / littéralement "toi, tu"+"bon, bien"* ou encore "merci", dissyllabique aussi 谢谢*/ xie xie / merci merci*. Les textes de forums contiennent aussi des éléments dialogiques tels que les formes d'adresse "tu" : 你 */ ni* , "vous de politesse" : 您 */ nin*. Enfin, les intervenant utilisent des onomatopées comme 哈 */ ha / [bruit de rire]* 呵呵 */ he he / [bruit de rire]* ou des particules exclamatives comme 啦 */ la / [exclamation, question ou accord]*, des interjections comme 嗯 */ ng / [assentiment, accord]* qui appartiennent habituellement au registre oral.

Un lexique spécifique aux textes traitant du diabète...

Le sujet du diabète possède un vocabulaire spécialisé, aussi abondamment présent dans le corpus étudié qu'il sera absent d'un corpus généraliste ou relevant d'une autre spécialité. Les mots de spécialité médicale écrits en chinois sont soit formés de signifiants, soit formés d'éléments phonétiques. Par exemple, on peut combiner des morphèmes chinois signifiants pour former Hyperglycémie => Élevé + Sang + sucre => 高血糖 / *gao xue tang*.

On peut aussi combiner des morphèmes chinois pour leur valeur phonétique pour former Marathon => *ma la song* => 马拉松

6 http://www.alexa.com/

… et même une syntaxe particulière au diabète !

Dans certaines phrases, les contributeurs utilisent entre eux des raccourcis pour parler de leur quotidien, qui donnent lieu à des impossibilités syntaxiques, comme par exemple de trouver un déterminant placé après son déterminé ; d'où la phrase 餐一和餐三跟没法比 par exemple, analysée en détails au chapitre 4.2.3

3.2 Trois outils de segmentation à comparer

3.2.1 Critères de choix

L'offre de segmenteurs pour le Chinois est pléthorique : depuis les outils en ligne (exemple : Hanzidico[7]) permettant de copier / coller son texte jusqu'aux librairies Java et modules Perl, Ruby ou autres, la difficulté à choisir est réelle. Grâce à mon stage de M2 dans le laboratoire Er-tim[8], j'ai pu consacrer du temps à rechercher, installer et essayer un bon nombre d'entre eux, exclusivement non commerciaux Après quelques semaines de cette activité, j'ai déterminé les critères ci-dessous pour choisir trois outils.

Robustesse et pertinence de premier niveau

Il s'agit d'un critère négatif, qui m'a fait éliminer les outils "capricieux" qui ne savent pas traiter le mélange avec les caractères latins, les caractères chinois traditionnels ou simplifiés ; de même j'ai éliminé les outils rendant un résultat ostensiblement fantaisiste, par exemple celui qui m'a retourné une liste de candidats-mots contenant des chaînes de 8 caractères constituées en réalité de phrases entières avec locatif, sujet, verbe, et objet. Enfin j'ai éliminé les outils n'offrant qu'un service en ligne , qui ne permettent pas de traiter aisément de gros volumes et d'automatiser le traitement de nombreux fichiers, et privilégié les outils fonctionnels sur un système Linux.

Simplicité d'installation et d'utilisation

En fonction de la précision de la documentation, l'installation puis la mise en œuvre d'un outil de TAL peut ressembler au parcours du combattant. Les outils décrits ci-dessous ont la grande qualité de s'installer facilement pourvu qu'on suive la séquence des instructions données, et d'être utilisables très rapidement à partir des exemples données dans la documentation. Ils peuvent notamment être utilisés directement sans obliger l'utilisateur à un paramétrage complexe ou à un entraînement préalable.

Popularité

Pour trouver des outils à recenser dans le projet Multital, j'ai interrogé les talistes de mon entourage, lancé des requêtes sur des moteurs de recherche internet, épluché la littérature scientifique sur le sujet et parcouru les forums et autres blogs de talistes sinophones. Sans avoir fait de statistiques sur

7 http://www.hanzidico.com/dictionnaire-chinois-francais-anglais/segmentation-de-textes-chinois/
8 http://www.er-tim.fr/

la fréquence des recommandations de tel ou tel outil, j'ai vite vu que certains noms revenaient plus souvent dans les commentaires positifs.

Finalement, les trois outils dont j'ai choisi de comparer les performances sont décrits ci-dessous. Jieba et FNLP ont été développés par des chinois, ils sont spécifiques au traitement du chinois. Le troisième, Polyglot, comme son nom l'indique n'est pas spécifique à une langue, mais il possède des modèles tout prêts pour le chinois. Tous les trois sont des systèmes de segmentation par apprentissage supervisé.

3.2.2 Jieba : Chinois seulement, écrit en Python

Le logiciel est disponible sous licence la licence libre MIT[9], en téléchargement sur Github[10]. Son développeur, Sun Jinyi, se présente comme un doctorant de l'Académie chinoise des Sciences[11], habitant Pékin et travaillant chez Baidu[12], le premier site chinois en terme de trafic au classement Alexa. Jieba est un outil développé en Python, par un chinois, et souvent cité positivement dans les forums ou les blogs de discussion chinois spécialisés.

Dans son document Readme[13] rédigé en chinois et en anglais, l'auteur décrit ainsi la méthode de TAL utilisée pour calculer la segmentation : à partir d'une structure de dictionnaire de préfixes, il réalise une analyse efficace du graphe de mots. Il construit un graphe orienté acyclique pour trouver toutes les combinaisons possibles de mots. Puis utilise la programmation dynamique pour trouver la combinaison la plus probable de mots en fonction de la fréquence des mots. Pour la détection des mots inconnus, un modèle de Markov-Caché est utilisé avec un algorithme de Viterbi.

Le dictionnaire de mots avec fréquence fourni par le logiciel comprend 1047138 entrées, qui sont de la forme Mot fréquence du mot étiquette morpho syntaxique. Exemples :

Entrée			Sens	Commentaire
大学生	3879	n	Étudiant	
台大学生	3	n	Étudiant de l'université nationale de Taiwan	台大 : abréviation de 台湾大学 (université de Taiwan) + 学生 (élève)
世界大学生运动会	12	nz	Jeux mondiaux universitaires	Litt : monde / mondial étudiant sport/sportif réunion
三月初	28	t	Début mars	Litt : trois mois début
一盘散沙	52	n	désordre	Litt : un tas de sable en vrac

Table 10: Exemple du lexique fourni avec Jieba

9 https://opensource.org/licenses/MIT
10 https://github.com/fxsjy/jieba
11 http://english.cas.cn/
12 http://www.baidu.com/
13 https://github.com/fxsjy/jieba/blob/master/README.md

Il n'est pas fourni d'information sur l'origine de ce lexique, à partir de quelles sources il a été constitué.

3.2.3 FNLP : Chinois seulement, écrit en Java

FNLP a été développé à l'université FUDAN[14] de Shanghai ; le logiciel est disponible sous licence libre GPLv3[15] , en téléchargement sur Github[16]. Les trois auteurs, Xuanjing Huang, Xipeng Qiu, Qi Zhang, ont rédigé un article en anglais qui présente l'outil FNLP (Qiu, Zhang, et Huang 2013) dans son intégralité et pas seulement la segmentation en mots. Les auteurs expliquent que l'outil est d'abord construit sur des méthodes statistiques, complétées de méthodes à base de règles, et entraîné sur un corpus dont "la qualité est cruciale pour notre outil". Puis, après avoir écarté les corpus connus et accessibles, ils disent construire un corpus d'entraînement "en accord avec la compréhension commune de la grammaire chinoise" ; malheureusement, il n'y a pas de références précises permettant d'identifier le corpus ou les règles utilisées. Ils détaillent ensuite les 3 couches de leur système : pré-traitement des données, module central d'apprentissage automatique et enfin modules de traitement du langage naturel.

Pour réaliser la segmentation en mots, les auteurs utilisent d'abord un étiquetage de séquence de caractères (CRF) pour trouver les limites des mots. Ensuite, ils ajoutent une partie de sémantique extraite de la base de connaissance chinoise HowNet[17] (Dong et Dong 2006) , qui d'après eux améliore grandement la performance. Enfin, l'utilisation d'un algorithme de Viterbi est implémenté pour permettre aux utilisateurs d'ajouter leur propre dictionnaire.

3.2.4 Polyglot : "toutes" les langues, écrit en Python

Contrairement à FNLP et Jieba qui sont spécialisés pour le chinois, Polyglot est un outil de traitement des langues délibérément multilingue et utilisant l'apprentissage non supervisé. Les modules proposés incluent la détection de langue, la segmentation en mots et en phrases, l'étiquetage morpho syntaxique et la reconnaissance des entités nommées. Son auteur, Rami Al-Rfou[18] a développé Polyglot dans le cadre de sa thèse et l'a mis à disposition sous licence GPLv3+ comme un package Python. Le principe, décrit dans un article détaillé (Al-Rfou, Perozzi, et Skiena 2013) est d'entraîner le système pour l'étiquetage morpho-syntaxique en utilisant les pages Wikipedia comme corpus pour 117 langues ; des vecteurs de mots sont générés à partir des corpus Wikipedia volontairement peu normalisés. L'approche décrite présente pas mal d'aspects créatifs et j'attendais avec impatience la description du processus de traitement du chinois. Mais en y regardant de plus près, quelle déception en voyant que la tokenisation préalable était réalisée par Apache OpenNLP si disponible (ce qui n'est pas le cas pour le chinois) ou sinon par l'algorithme Lucene de segmentation suivant les recommandations Unicode. La consultation de la recommandation Unicode[19] est assez stimulante, jusqu'à la phrase "For example, reliable detection of word boundaries in languages such as Thai, Lao, Chinese, or Japanese requires the use of dictionary lookup" qui explique que tous les mécanismes subtils décrits jusque là ne s'appliquent pas au

14 http://www.fudan.edu.cn/en/
15 http://www.gnu.org/licenses/quick-guide-gplv3.fr.html
16 https://github.com/FudanNLP/fnlp
17 http://keenage.com/html/e_index.html
18 https://sites.google.com/site/rmyeid/
19 http://www.unicode.org/reports/tr29/

Chinois, qui sera traité à l'aide d'un dictionnaire. Et donc finalement, Polyglot comme les autres systèmes repose pour la détection des mots en Chinois sur un vocabulaire des 100000 tokens les plus fréquents extraits de Wikipedia avec vérification dans le dictionnaire du Chinois contemporain (Yu ming shan 2013) ; la version utilisée pour Polyglot d'après l'article contient 65 000 mots.

3.3 Protocole expérimental

3.3.1 Constitution du corpus

A l'aide du module Perl Web::Scraper[20], j'ai récupéré une liste d'URL des forums du site http://bbs.tnbz.com/, puis dans chaque post, j'ai récupéré uniquement le texte de la contribution. Je n'ai pas conservé les noms des contributeurs ou la date des contributions, mon objectif étant d'obtenir seulement du texte brut. Ensuite j'ai conservé seulement 1000 lignes de texte, car je n'avais pas besoin de gros volumes.

Je tiens à remercier Zhai Yuming[21] pour le code Perl que j'ai utilisé, directement issu de notre travail d'analyse des forums du deuxième semestre de M2.

3.3.2 Dé-ponctuation / re-ponctuation

De façon assez inattendue, les segmenteurs Jieba et FNLP ont planté brutalement sur certains caractères de ponctuation présents dans le texte. Même avec un fichier converti et reconnu comme UTF-8, certains caractères généraient un message d'erreur ou arrêtaient le processus de segmentation.

Caractère	Description
?	Point d'interrogation suivi d'un espace insécable
.	Point utilisé comme séparateur numérique ou séparateur d'adresse internet
!	Point d'exclamation suivi d'un espace insécable
。	Point final asiatique suivi d'un espace insécable
?	Point d'interrogation de petite taille (plus petit que les caractères et plus petit que celui qui est suivi d'un espace insécable)
.	Point final occidental suivi d'un espace insécable

Tableau 11: Ponctuations problématiques pour les segmenteurs

Peut être que c'est uniquement le caractère "espace insécable" qui génère des erreurs dans les traitements : comme les erreurs n'étaient homogènes entre les systèmes, cette hypothèse n'est apparue qu'après la fin du travail d'analyse et n'a pas été testée. Pour obtenir des corpus segmentés complets, une étape de normalisation minimale a été réalisée, consistant à remplacer les caractères ci-dessus avant la segmentation, puis à les restaurer après pour conserver la sémantique. Une instruction *sed* en ligne de commande avant et après chaque segmentation a suffit.

20https://metacpan.org/pod/Web::Scraper
21https://fr.linkedin.com/in/yumingzhaital/fr

3.3.3 Segmentation

Pour rester cohérente avec mon deuxième critère de choix d'outils (simplicité d'installation et d'utilisation), j'ai choisi d'utiliser les segmenteurs de la manière la plus dépouillée possible, c'est à dire en utilisant exclusivement les ressources fournies par les développeurs, sans ajouter de lexique et en utilisant si possible les options par défaut, en séparant les unités lexicales par des espaces.

En effet mon objectif n'est pas d'offrir un manuel détaillé de chacun des outils, mais de vérifier ce qui se passe pour un utilisateur n'ayant pas un profil technique très pointu, qui voudrait pouvoir segmenter des textes sans "mettre les mains sous le capot". Dans ce cadre, je me suis d'ailleurs interrogée sur l'opportunité de la phase de normalisation de la ponctuation, qui peut déjà paraître complexe à certains utilisateurs. Je l'ai cependant maintenue, car sans cela, d'une part il n'y avait aucune comparaison possible des résultats, d'autre part, c'est un problème commun à tous les segmenteurs que je connais.

Au final, j'ai donc lancé une seule commande par segmenteur.

Outil	Commande	Paramétrage
Jieba	python -m jieba *1000lignes_nopunct.txt* **--delimiter** > *jieba_seg.txt*	Le délimiteur est un espace
FNLP	java -Xmx1024m -Dfile.encoding=UTF-8 -classpath "fnlp-core/target/fnlp-core-2.1-SNAPSHOT.jar:libs/trove4j-3.0.3.jar:libs/commons-cli-1.2.jar" org.fnlp.nlp.cn.tag.CWSTagger -f **models/seg.m** *1000lignes_nopunct.txt fnlp_seg.txt*	Limitation à la tâche de segmentation
Polyglot	polyglot **--lang zh tokenize** --input *1000lignes_nopunct.txt > polyglot_seg.txt*	Indication de la langue et appel de la tâche de segmentation

Tableau 12: Instructions d'appel des segmenteurs

Le paramétrage des instructions (indiqué en **gras**) est minimal : la commande est le plus proche possible des exemples fournis dans les documentations des outils.

4 Analyse des résultats

4.1 Différences de surfaces

Très simplement à l'aide de WordCount, on obtient un décompte des mots par fichier.

Outil	Nombre de mots après segmentation	Part du total
Fnlp	29558	33,55 %
Jieba	29124	32,98 %
Polyglot	29630	33,47 %
Total pour les trois fichiers	88312	

Tableau 13: Nombre de mots en fonction du segmenteur

En volume, les différences semblent minimes. Par contre, en poursuivant l'analyse par quelques simples commandes bash, on peut repérer quelques différences plus marquées. Ci dessous, une comparaison pour chaque segmentation, du nombre de monosyllabes, dissyllabes, trisyllabes et quadrisyllabes identifiés

Répartition par la taille des mots, en nombre de caractères				
	Monosyllabes	Dissyllabes	Trisyllabes	Quadrisyllabes
FNLP	909	934	461	116
Jieba	911	958	436	149
Polyglot	909	897	300	56

Tableau 14: Répartition des mots par nombre de syllabes après segmentation

Comme on l'a vu ci-dessus en 2.2.7, les mots du chinois contemporain sont de moins en moins monosyllabiques et le lexique tend vers le dissyllabisme et dans une moindre mesure trisyllabisme. Mais les mots plus longs existent aussi. Les quadrisyllabes ont une place à part dans le lexique chinois : ce sont très souvent des expressions toutes faites, les 成语 / *chengyu* / *expressions idiomatiques* qui sont constitués de 4 caractères.

Dans un corpus comme le nôtre, on va également trouver des mots spécialisés : issus de langues étrangères ou désignant un symptôme, un médicament... ils ne respectent pas toujours les règles de formation traditionnelle des mots en chinois.

Le décompte des mots de 5 caractères et plus n'a pas de sens ici, car les résultats obtenus se révèlent être majoritairement des mots en caractères latins : parties d'URL vers des documents externes (excel, health, family...), composés chimiques (Peptide, Glucose).

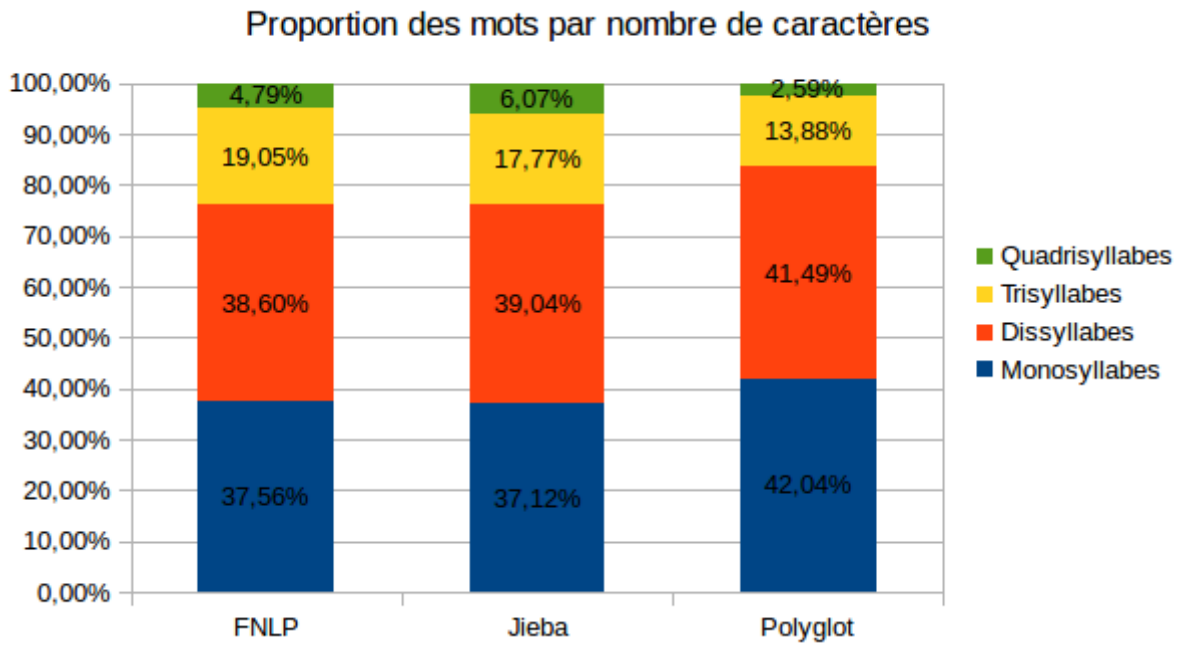

Illustration 1: Répartition des mots par nombre de caractères

Le graphique ci-dessus montre au premier coup d'œil que les choix de segmentation des 3 outils diffèrent : si la proportion des mono et dissyllabiques est très semblable entre FNLP et Jieba, elle est beaucoup plus élevée pour Polyglot. Il va donc nécessairement y avoir des différences d'interprétation, voire des erreurs de segmentation, soit d'un côté, soit de l'autre.

4.2 Analyse des spécificités de vocabulaire

4.2.1 Pourquoi et comment calculer des spécificités

Quand on réalise une analyse textométrique d'un corpus, il est d'usage de le diviser en plusieurs sous-parties, par exemple suivant des périodes de temps, ou des auteurs, puis de calculer les fréquences relatives de chaque forme de vocabulaire dans chaque sous-partie du corpus. Ce calcul permet de dégager d'une part les formes relevant de la généralité ou de la banalité, ce sont celles qui sont répartie uniformément dans chaque sous-partie du corpus, et d'autre part les formes relevant de la spécificité, ce sont celles qui sont réparties inégalement entre les sous-parties. Sur une forme donnée, plus ses spécificités entre les sous-parties sont contrastées, plus la forme est spécifique à l'une ou l'autre sous-partie. (Lafon 1980) expose en détails les fondements théoriques et donne une application en exemple de ce mode d'analyse.

Dans le cadre de ce travail, le corpus à étudier est constitué par les trois fichiers issus de la segmentation d'un même texte de départ ; chaque sous-partie du corpus correspond à l'un des résultats de segmentation. Si les segmentations étaient identiques, toutes les formes de vocabulaire seraient identiques et les spécificités entre sous-parties seraient nulles. On a vu au chapitre précédent que ce ne sera pas le cas ; dans le chapitre qui vient, l'examen des spécificités de vocabulaire sur les formes les plus contrastées va donner des indices des divergences à rechercher.

4.2.2 Caractère 一/ yi / numéral un

Le calcul des spécificités entre les trois résultats de segmentation sur le caractère 一 */ yi / numéral un* met en évidence des disparités marquées. Or, ce caractère, le deuxième plus fréquent dans la langue chinoise d'après le Wiktionnaire[22] a des usages très variés qui peuvent donner lieu à des choix de segmentations particuliers à chaque usage :

- Si le caractère est utilisé pour dénombrer, il doit être isolé.

- s'il entre dans la composition d'un nombre, par exemple 二零零一 */ er ling ling yi / 2001*, il ne doit pas être isolé

- s'il fait partie d'une expression lexicalisée comme 一定 */ yi ding /sûrement* ou bien 一般 / *yi ban / généralement*, il ne doit pas être isolé.

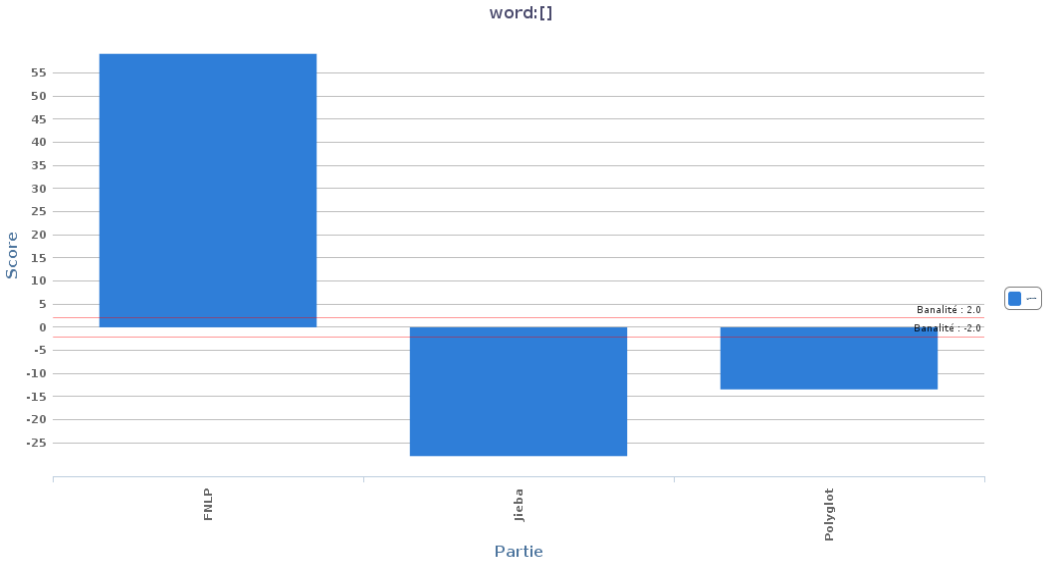

Illustration 2: Spécificité du caractère "Un"

Or, le diagramme de spécificité indique d'emblée que si un des segmenteurs a la bonne interprétation, les autres ont fait des erreurs. Si par exemple Jieba repère très peu de formes isolées de 一 */ yi / numéral un* c'est peut-être qu'il a correctement conservé ensembles les nombres et les adverbes, alors que FNLP les a isolés à tort. A l'inverse, il se peut que FNLP ait bien interprété les formes verbales de type Verbe 一 Verbe qui doivent être séparées en trois unités lexicales tandis que Jieba aura, à tort, gardé ensemble le groupe de trois caractères.

4.2.3 餐后 / can hou / postprandial

Ce dissyllabe a une importance particulière : dans un forum sur le diabète, la mesure de la glycémie postprandiale (deux heures après le début du repas) est cruciale pour l'auto-surveillance de la maladie. L'expression est donc beaucoup plus fréquente dans le corpus analysé que dans des textes généralistes, mais d'autre part, elle est constitué de deux caractères 餐 */ can / repas* et 后 */ hou / après* qui peuvent aussi être interprétés comme des unités isolées.

22 https://fr.wiktionary.org/wiki/Wiktionnaire:1000_hanzi

Il est frappant de constater que ce dissyllabe n'est pas présent une seule fois après les segmentations réalisées par Polyglot, qui a systématiquement séparé le nom de la préposition.

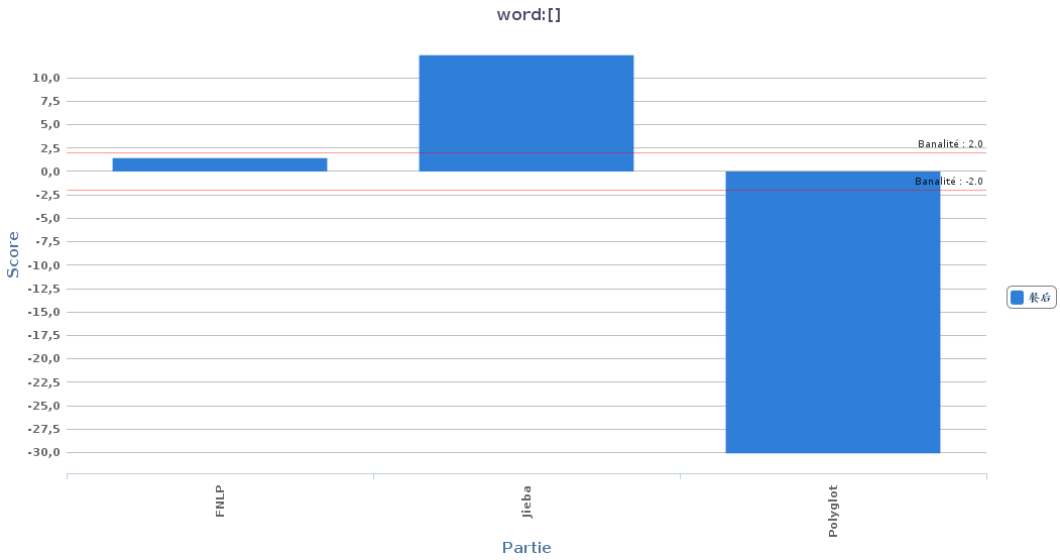

Illustration 3: Diagramme des spécificités du dissyllabe 餐后 / canhou / postprandial

4.2.4 Spécificités 糖 / tang / sucre

Le caractère 糖 / *tang* / *sucre* est nécessairement très présent dans les discussions portant sur le diabète. Bien que pouvant être utilisé seul, il sert aussi à former des composés très fréquents dans le corpus comme 糖尿病 / *tang niao bing* / diabète, 血糖 / *xue tang* /glycémie *ou* 葡萄糖 / *pu tao tang* / *glucose*. Or, le diagramme des spécificités montre les divergences d'interprétation des trois systèmes.

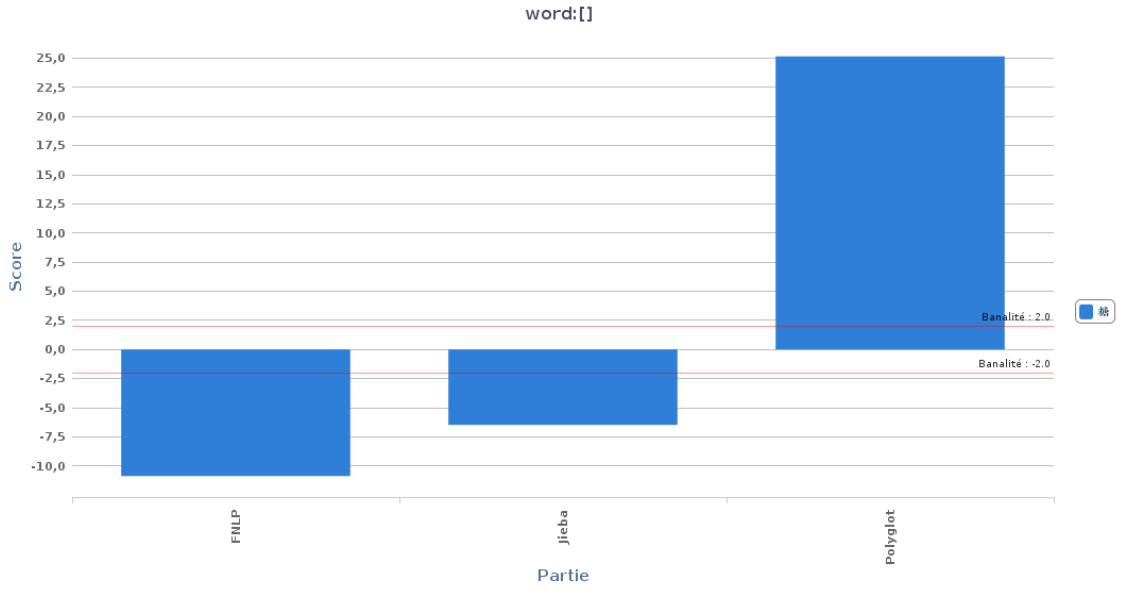

Illustration 4: Spécificités du mot "sucre"

On voit que Polyglot sur-segmente 糖 / *tang* / *sucre* comme il avait sous-segmenté 餐后 / *canhou* / *postprandial*. La tendance de Polyglot à segmenter en unités plus petites, observée en 4.1, se retrouve logiquement sur des formes diverses du corpus.

4.2.5 Négation 不 / bu， Particule 了 / le,liao, et classificateur 个 / ge

Ces trois caractères peuvent chacun constituer un mot isolé. Ils sont aussi parmi les quinze caractères les plus fréquents de la langue chinoise et entrent en composition de très nombreux mots ; le fait de les séparer comme des unités unités lexicales autonomes n'est donc pas sans conséquence.

Par exemple, la négation 不 / *bu* peut former l'adverbe 不够 / *bugou* / *insuffisant* : dans le corpus, le mot dissyllabique est correctement identifié 5 fois par Jieba et Polyglot, mais seulement 2 fois par FNLP.

La particule 了 / *le,liao* peut être autonome, par exemple pour indiquer l'aspect accompli d'une action ; mais elle entre aussi dans la composition de nombreux mots. Par exemple, le verbe 了解 / *liaojie* / *comprendre* est présent 8 fois dans le corpus et identifié correctement 8 fois par Jieba et Polyglot ; par contre, FNLP dans un des cas isole 了 / *le-liao* et fabrique un 解么 / *jie me* qui n'a pas de sens.

Enfin le classificateur 个 / *ge*, s'il est le plus souvent une unité autonome et doit être séparé du numéral qu'il suit, peut aussi entrer en composition pour former des mots plus longs. Dans le corpus d'origine, on trouve ainsi 8 occurrences de 半个 / *ban ge* / *la moitié* , le même nombre dans le corpus segmenté par Polyglot, mais plus aucun dans le corpus segmenté par Jieba ou FNLP.

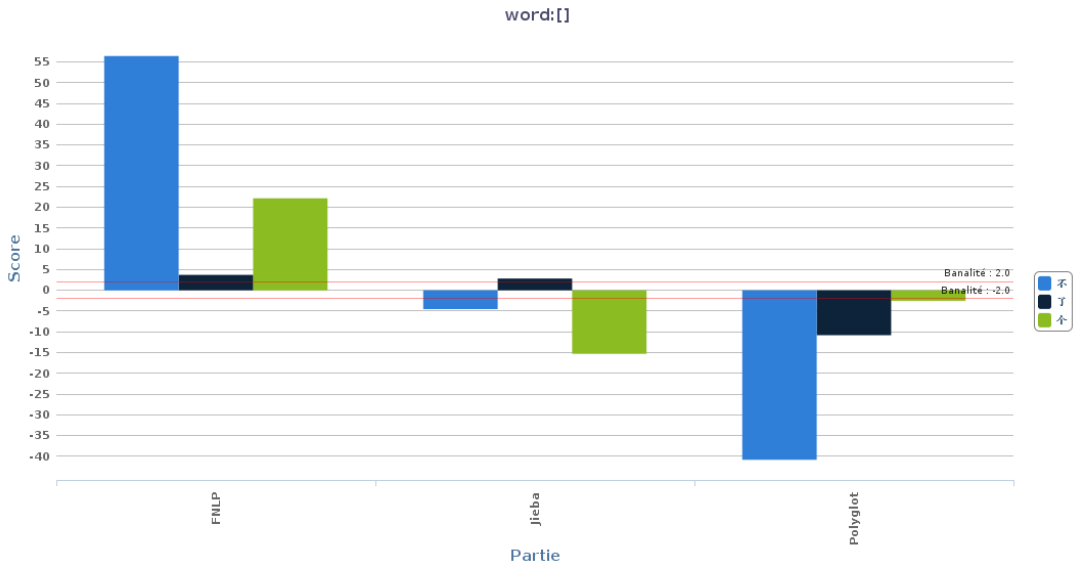

Illustration 5: Spécificités 不 / bu 了 / le 个 / ge

4.3 Exemples de divergences de segmentation

4.3.1 Méthodologie et représentation

Repérage

Pour parvenir à une analyse plus approfondie, j'ai cherché parmi les différences de traitement des segmenteurs des exemples permettant d'illustrer de façon la plus parlante possible les problèmes qu'une segmentation erronée peut causer. Pour rester cohérente avec ma démarche de "vulgarisation" du TAL, j'ai eu recours à deux outils très simples pour repérer des différences dans les résultats de traitement :

- des instructions bash pour repérer les "mots" de 4 caractères et plus et leur présence ou non dans chaque résultat de segmentation
- le logiciel de textométrie TXM pour repérer les spécificités saillantes forme par forme.

Ensuite, je suis chaque fois revenue au texte d'origine pour extraire un contexte signifiant autour de la partie repérée, phrase ou préposition et vérifier la pertinence ou non des segmentations proposées. Enfin, j'ai choisi 10 exemples pour illustrer huit particularités de traitement de la langue chinoise.

Présentation

Dans les paragraphes ci-dessous, je présente les exemples analysés en utilisant d'abord une représentation visuelle des résultats obtenus, ce sont les **silhouettes** de la phrase en fonction des longueurs des mots obtenus après les segmentations.Cette représentation en bâtons permet de faire abstraction de l'écriture chinoise (qui peut gêner les non-sinophones) et même du sens pour mieux mettre en évidence la diversité des résultats obtenus.

Je donne ensuite les quatre segmentations réalisées : humaine, avec FNLP, avec Jieba et avec Polyglot. Là encore, je propose une représentation graphique à base de cases de taille proportionnelle au nombre de caractères, pour permettre une lecture graphique détachée du sens.

Enfin, je commente les résultats obtenus en terme d'usage, de sens ou d'écart à la grammaire chinoise ou à la norme ISO, exemple par exemple.

4.3.2 成语 / cheng yu / proverbes à quatre caractères

Premier exemple

Séquence à segmenter : 要循序渐进不要一步到位

Signification: Il faut progresser peu à peu, il ne faut pas tout faire d'un coup.

Difficultés de segmentation : La phrase ci-dessus est très typique de la manière de s'exprimer dans la langue chinoise : elle oppose deux 成语 / *Cheng yu* / *proverbes à quatre caractères* ayant des signification contraires, en utilisant le même verbe une fois en positif 要 / *yao* / *falloir* et une fois en négatif 不要 / *bu yao* / *négation falloir*. La difficulté ici pour les segmenteurs est bien entendu d'identifier les proverbes en tant qu'unités lexicales.

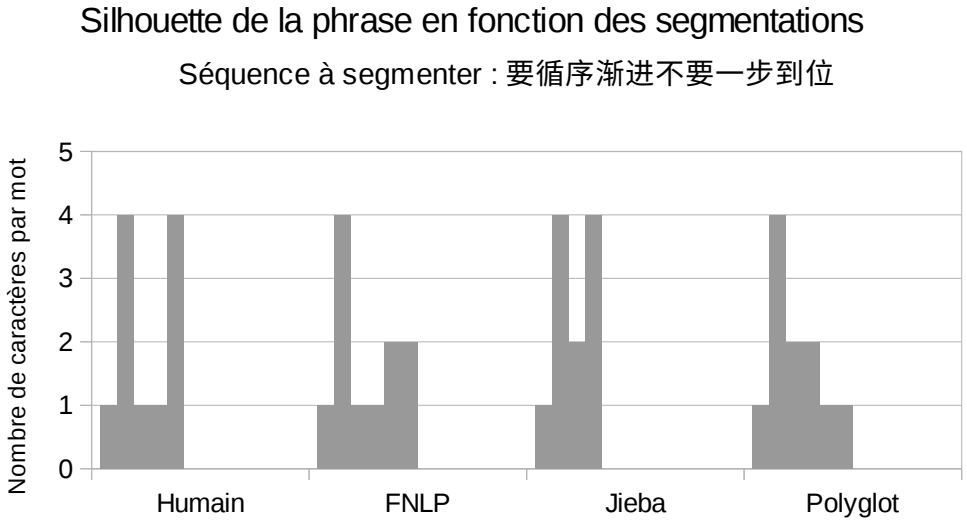

Illustration 6: Silhouettes : Deux expressions quadrisyllabiques

Segmentations obtenues

Humaine	要 循 序 渐 进 不 要 一 步 到 位
FNLP	要 循 序 渐 进 不 要 一 步 到 位
Jieba	要 循 序 渐 进 不 要 一 步 到 位
Polyglot	要 循 序 渐 进 不 要 一 步 到 位

Tableau 15: Divergences de segmentation - Formules quadrisyllabiques

Observations

Bien que s'appuyant sur le dictionnaire du chinois contemporain(Yu ming shan 2013) , dans lequel figurent les deux expressions à quatre caractères 循序渐进 / *xun xu jian jin / procéder pas à pas* et 一步到位 / yi bu dao wei / *régler le problème d'un seul coup,* Polyglot n'en identifie qu'une sur les deux. FNLP échoue également à identifier les deux expressions, mais on ne sait pas précisément ce que contient son corpus d'apprentissage.

Deuxième exemple
Séquence à segmenter : 最后再画蛇添足多说几句

Signification: Pour finir, encore quelques phrases superflues

Difficultés de segmentation : L'exemple ci-dessous analyse le traitement d'une expression à quatre caractères lexicalisée très courante en Chinois. L'expression 画蛇添足 */ hua she tian zu / superflu* se trouve telle quelle dans le dictionnaire du Chinois contemporain en tant qu'unité lexicale. Elle signifie littéralement "dessiner le serpent ajouter des pattes", elle est donc de la forme Verbe-Objet-

Verbe-Objet mais ne doit en aucun cas s'analyser en éléments unitaires. Sur les trois segmenteurs, seul FNLP ne parvient pas à l'identifier correctement.

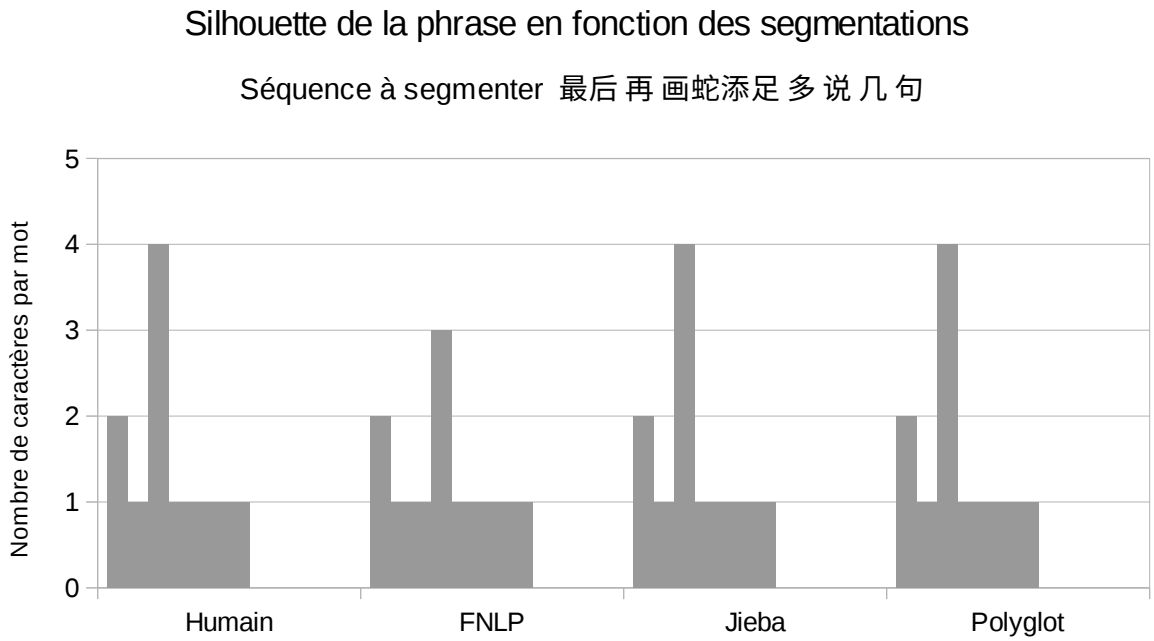

Illustration 7: Silhouettes - Proverbe "Dessiner serpent ajouter pattes"

Segmentations obtenues

Humaine	最 后 | 再 | 画 蛇 添 足 | 多 | 说 | 几 | 句
FNLP	最 后 | 再 | 画 | 蛇 添 足 | 多 | 说 | 几 | 句
Jieba	最 后 | 再 | 画 蛇 添 足 | 多 | 说 | 几 | 句
Polyglot	最 后 | 再 | 画 蛇 添 足 | 多 | 说 | 几 | 句

Tableau 16: Divergences de segmentation - Proverbe "Dessiner serpent ajouter pattes"

Observations

Les segmentations de Jieba et Polyglot sont parfaites ; pour Polyglot, on peut l'expliquer par la présence de l'expression 画蛇添足 */ hua she tian zu / superflu* dans le dictionnaire sur lequel il s'appuie ; de même pour Jieba, dont le lexique contient l'expression exacte avec la fréquence assez haute de 31. Par contre, FNLP détache le premier terme comme verbe et propose ensuite un groupe de trois caractères qui n'a pas de sens.

4.3.3 Formes verbales et déterminant + déterminé

Séquence à segmenter : 水货不是国产的而已，就是不交税的进口产品

Signification: Les produits du marché parallèle ne sont pas simplement des produits de production nationale, ce sont des produits importés qui n'ont pas payé de droits de douane.

Difficultés de segmentation :

Dans la première partie de la phrase, la négation 不 / *bu* doit être séparée du verbe qu'elle précède et non confondue avec l'expression 不是 / *bu shi* / *non*. Dans la deuxième partie, on a un syntagme Verbe + Objet, à séparer, ainsi que qu'une expression déterminant+déterminé simplement juxtaposés, sans conjonction. Ce sont des formes extrêmement courantes en Chinois.

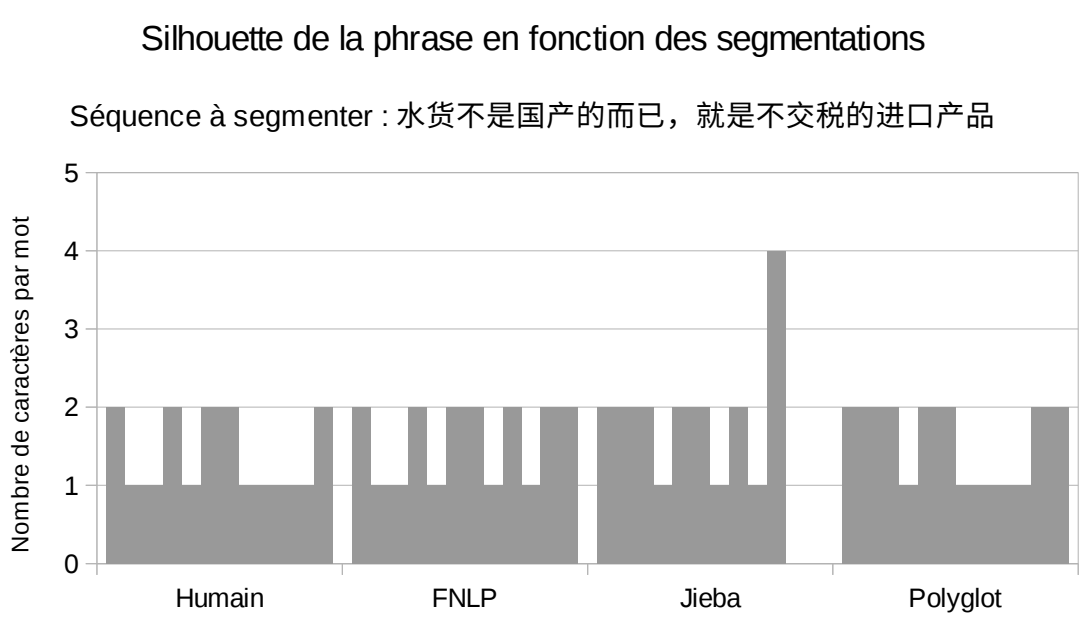

Illustration 8: Silhouettes - Formes verbales et Déterminant + Déterminé

Segmentations obtenues

Humaine	水 货 不 是 国 产 的 而 已 ， 就 是 不 交 税 的 进 口 产 品
FNLP	水 货 不 是 国 产 的 而 已 ， 就 是 不 交 税 的 进 口 产 品
Jieba	水 货 不 是 国 产 的 而 已 ， 就 是 不 交 税 的 进 口 产 品
Polyglot	水 货 不 是 国 产 的 而 已 ， 就 是 不 交 税 的 进 口 产 品

Tableau 17: Divergences de segmentation - Formes verbales et Déterminant + Déterminé

Observations

Dans la première partie de la phrase, Jieba et Polyglot n'ont pas séparé correctement négation et verbe. Jieba n'a pas non plus détaché 进口 / *jin kou* / *importé* de 产品 / *chan pin* / *produit*, créant ainsi un bloc difficile à analyser. Enfin, le groupe 交税 / *jiao shui* / payer des taxes qui est un verbe+objet n'a été séparé ni par Jieba ni par FNLP, alors qu'il ne s'agit pas d'un composé stable.

4.3.4 Adjectif réduplqué AB => AABB

Séquence à segmenter : 医院不会把结果随随便便发给

Signification:L'hôpital ne peut pas tout simplement envoyer les résultats à [une clinique spécialisée]

Difficultés de segmentation : La séquence contient l'adjectif 随便 */ sui bian / sans façon* dans sa forme réduplqué 随随便便 */ sui sui bian bian / tout simplement* . Or, les adjectifs réduplqués sont bien lexicalisés mais ils ne sont pas notés dans les dictionnaires sous cette forme. Par contre, ils sont en nombre limité, leurs règles de redoublement sont connues et leur présence apporte des indications stylistiques et sémantiques. Ils constituent donc une particularité linguistique de la langue chinoise à côté de laquelle on ne peut passer.

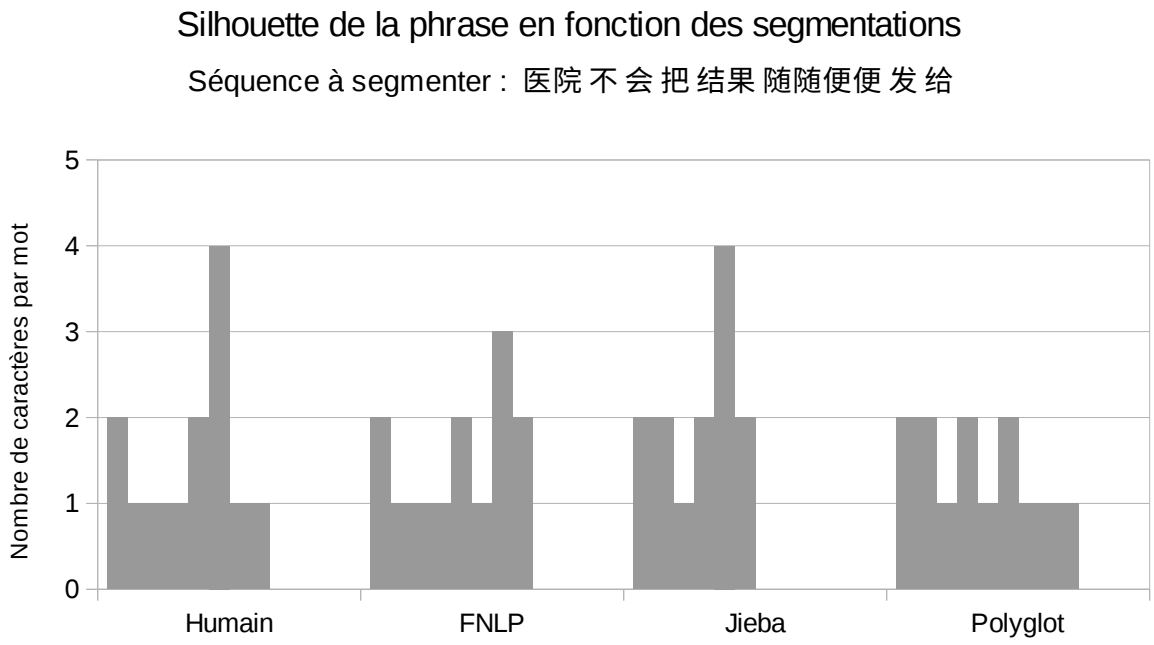

Illustration 9: Silhouettes - Adjectif réduplqué

Segmentations obtenues

Humaine	医 院 不 会 把 结 果 随 随 便 便 发 给
FNLP	医 院 不 会 把 结 果 随 随 便 便 发 给
Jieba	医 院 不 会 把 结 果 随 随 便 便 发 给
Polyglot	医 院 不 会 把 结 果 随 随 便 便 发 给

Tableau 18: Divergences de segmentation - Adjectif réduplqué

Observations

Dans l'exemple ci-dessous, seul Jieba a correctement découpé l'adjectif 随便 / *sui bian* redoublé en 随 随 便 便 / *sui sui bian bian* ; par contre, Jieba a mal séparé les verbes de la négation (comme Polyglot) et de la préposition (comme FNLP).

4.3.5 Verbes redupliqués, A donne A | A et AAB donne A | AB

Séquence à segmenter : 爬楼 4 趟， 晚餐后出去跑跑步 、 玩玩单杠

Signification: [Je] monte 4 étages, après le dîner je sors faire du jogging, m'amuser à la barre fixe.

Difficultés de segmentation : La phrase contient deux verbes, un monosyllabique 玩 / *wan* / *s'amuser* et un dissyllabique 跑步 / *pao bu* / *faire du jogging*. Ces deux verbes sont utilisés dans une forme de réduplication propre à chacun. Le monosyllabique est simplement doublé : on a 玩玩 / *wan wan* au lieu de 玩 / *wan*, le dissyllabique voit seulement son premier élément redoublé, on a donc 跑 跑 步 / *pao pao bu* au lieu de 跑 步 / *pao bu* . Comme expliqué dans (*Le Chinois Pour Tous Bescherelle Edition.* 2014), "la duplication indique que l'action est exécutée brièvement ou à titre d'essai" ; ici le contributeur exprime qu'il pratique une activité physique réelle mais modérée, sans trop se forcer. La difficulté est non seulement d'identifier ces unités, mais aussi de les segmenter correctement.

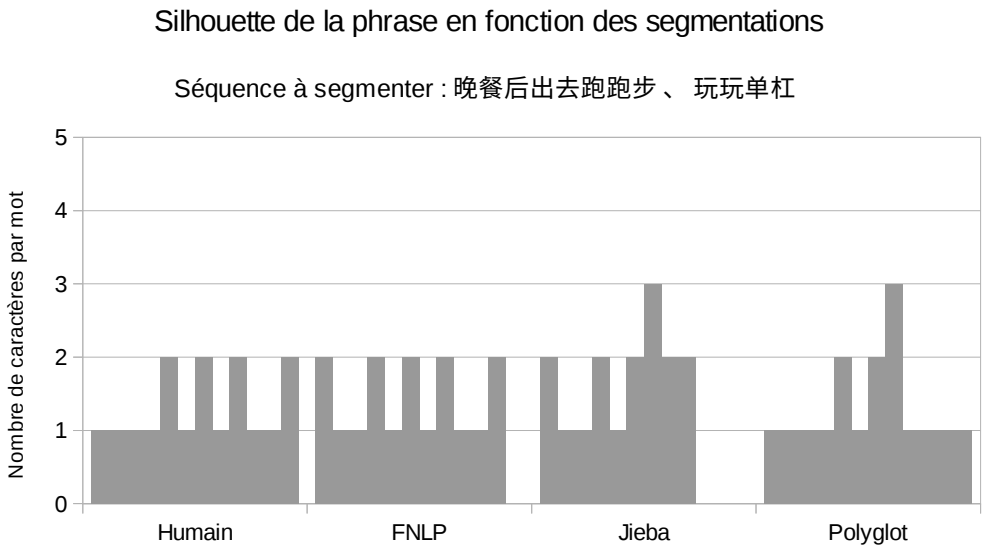

Illustration 10: Silhouettes – Verbes réduplicatifs

Segmentations obtenues

Humaine	爬 楼 4 趟 , 晚 餐 后 出 去 跑 跑 步 、玩 玩 单 杠
FNLP	爬 楼 4 趟 , 晚 餐 后 出 去 跑 跑 步 、玩 玩 单 杠
Jieba	爬 楼 4 趟 , 晚 餐 后 出 去 跑 跑 步 、玩 玩 单 杠
Polyglot	爬 楼 4 趟 , 晚 餐 后 出 去 跑 跑 步 、玩 玩 单 杠

Tableau 19: Divergences de segmentations– Verbes réduplicatifs

Observations

Si on suit la norme ISO, on considère comme un seul mot :

- la forme reduplicative lexicalisée d'un verbe monosyllabique (ex : 看 / *kan / voir* donne 看看 / *kan kan / jeter un coup d'oeil*) ; mais ce n'est pas le cas ici avec 玩 */ wan / s'amuser,* qui est bien la réduplication du verbe 玩 */ wan / s'amuser* pour lequel aucune forme redoublée stable n'existe.

- la forme reduplicative stable d'un verbe dissyllabique formé sur le modèle Verbe + Verbe (ex : 来来往往 / *lai lai wang wang / aller et venir*) ; ce cas ne se trouve pas dans l'exemple présent.

Ici, le verbe 跑步 */ pao bu / faire un footing* est de la forme Verbe (跑 / *pao* /*courir*) + Objet (步 / *bu / un pas*) ; le verbe (跑 / *pao* /*courir*) seul conserve tout son sens, à côté de la forme lexicale

d'origine 跑步 */ pao bu / faire un footing* qui est une forme stable, d'usage courant et doit donc rester telle quelle. Mais Jieba et Polyglot, en construisant comme unité lexicale la forme 跑跑步 / *pao pao bu* génèrent une unité inexploitable. La segmentation proposée par FNLP sur les verbes réduplicatifs rend évidemment un meilleur service que les autres. Cependant dans cette séquence, on note aussi la présence du Verbe + Objet 爬 */ pa / grimper* + 楼 */ lou / étage* qui n'est pas une forme stable lexicalisée et doit être séparée en deux unités : FNLP a échoué à les isoler.

4.3.6 Transcription d'un mot étranger

Séquence à segmenter : 等你能够跑马拉松了

Signification: Du moment que tu es capable de courir un Marathon

Difficultés de segmentation : la phrase contient un nom propre d'origine étrangère "Marathon" transcrit phonétiquement à l'aide de caractères chinois possédant leur propre sens mais choisis uniquement pour leur son : 马拉松 / *ma la song / cheval tirer relâcher*.

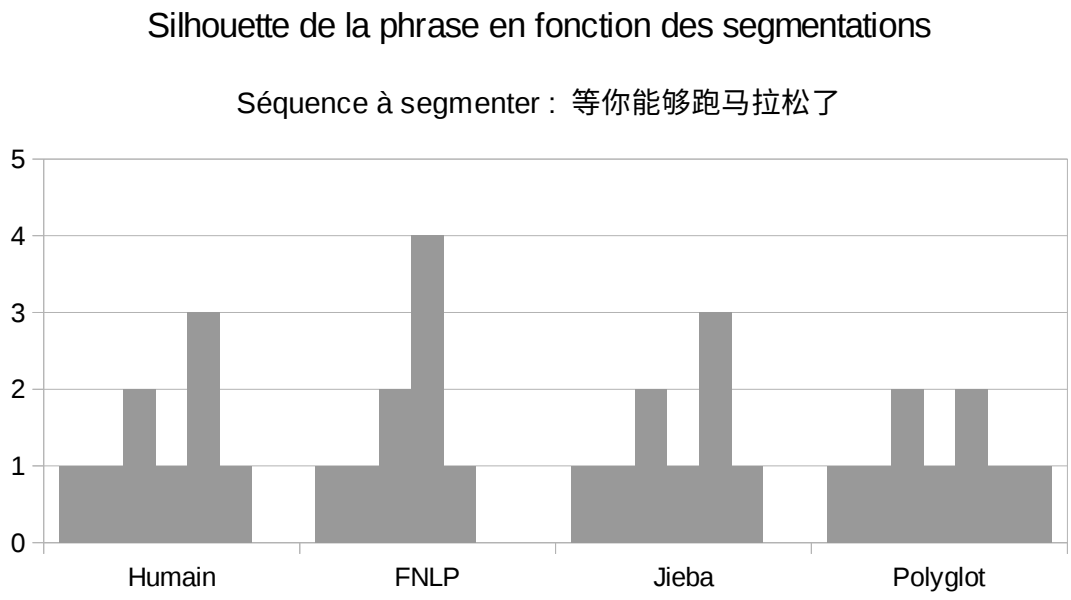

Illustration 11: Silhouettes – Transcription d'un nom étranger

Segmentations obtenues

Humaine	等 你 能 够 跑 马 拉 松 了
FNLP	等 你 能 够 跑 马 拉 松 了
Jieba	等 你 能 够 跑 马 拉 松 了
Polyglot	等 你 能 够 跑 马 拉 松 了

Tableau 20: Divergences de segmentation – Transcription d'un nom étranger

Observations

Seul jieba a correctement identifié le nom propre et détaché le verbe ; FNLP a créé le verbe "courir le marathon", tandis que Polyglot a détaché 马拉 / *ma la* et 松 / *song*, perdant complètement le sens de la phrase.

4.3.7 Numéraux

Séquence à segmenter : 糖人的百分之三衰退

Signification: chez les diabétiques, la baisse est de 3 %

Difficultés de segmentation : Quand les nombres sont exprimés à l'aide de caractères chinois et non de chiffres arabes, les segmenteurs peuvent commettre des erreurs d'interprétation. En effet, l'expression 3 % se transcrit en caractères chinois 百分之三 / *bai fen zhi san / 100 diviser parmi 3* et chacun des caractères peut s'utiliser seul en fonction du contexte.

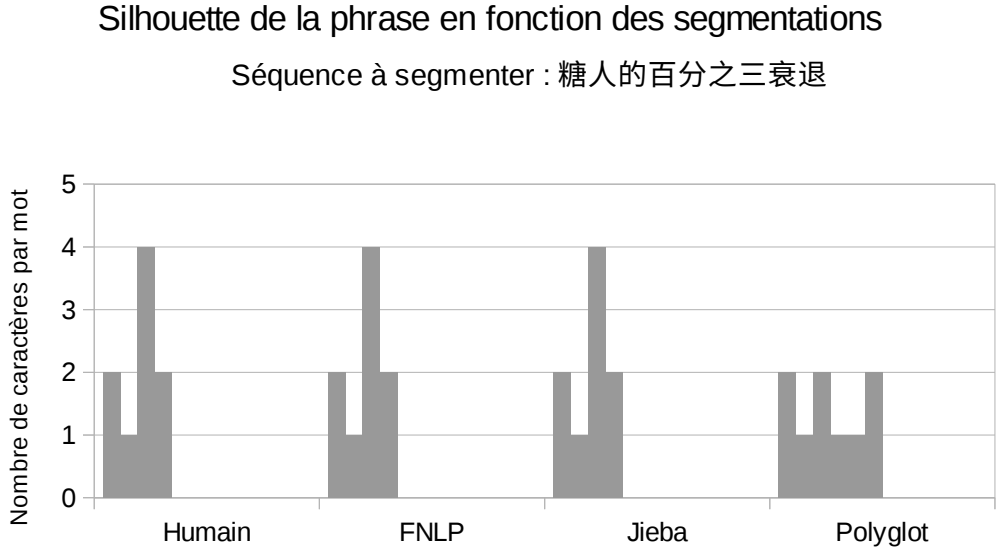

Illustration 12: Silhouettes - Expression d'un pourcentage

Segmentations obtenues

Humaine	糖 人 的 百 分 之 三 衰 退
FNLP	糖 人 的 百 分 之 三 衰 退
Jieba	糖 人 的 百 分 之 三 衰 退
Polyglot	糖 人 的 百 分 之 三 衰 退

Tableau 21: Divergences de segmentation - Pourcentage

Observations

La norme ISO recommande de considérer comme un seul mot les expressions numériques et particulièrement les expressions de pourcentages. Ici, la segmentation de Polyglot entraîne la perte de l'unité lexicale.

4.3.8 Numéraux – Jargon spécialisé

Séquence à segmenter : 餐一和餐三跟没法比

Signification: on ne peut pas comparer une postprandiale h+1 avec une postprandiale h+3

Difficultés de segmentation : La phrase à analyser est complexe car elle fait référence aux conversations précédentes, où les contributeurs évoquent leurs glycémies mesurées une, deux ou trois heures après le repas (postprandial h+1, h+2 et h+3). Au fil des échanges, les expressions sont raccourcies jusqu'à contenir seulement une expression de type [repas] [nombre]. Le fait d'avoir un numéral derrière un nom est inhabituel en chinois, où l'ordre déterminant+déterminé est strict ; de plus comme il n'y a ni ordinal ni classificateur devant le numéral, la présence du numéral dans cette phrase ne correspond à aucune règle grammaticale. Prise en dehors de son contexte, cette phrase est difficile à comprendre même pour une personne de langue maternelle chinoise. Mais si on suit le standard ISO, il suffit de séparer chaque numéral, ainsi que les conjonctions et les deux verbes.

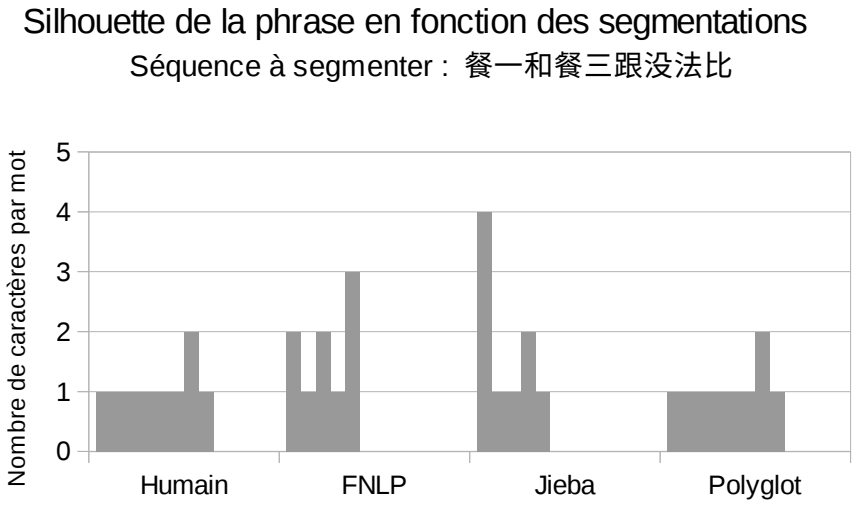

Illustration 13: Silhouettes - Jargon de spécialiste

Segmentations obtenues

Humaine	餐 一 和 餐 三 跟 没法 比
FNLP	餐一 和 餐三 跟 没法比
Jieba	餐一和餐 三 跟 没法 比
Polyglot	餐 一 和 餐 三 跟 没法 比

Tableau 22: Divergences de segmentation - Jargon de spécialiste

Observations

FNLP a créé des unités lexicales à partir d'un nom+suivi d'un numéral, ce qui n'existe dans aucun dictionnaire et sera assez difficile à analyser dans la suite d'une étude TAL ; il s'est probablement appuyé sur la présence des conjonction 和 */ he / et* et 跟 */ gen / avec* pour en déduire la structure des mots précédents et suivants . De plus, il n'a pas séparé les deux verbes de la fin de la phrase.

La proposition de Jieba est inexplicable, il a isolé un numéral et pas l'autre et n'a pas repéré les conjonctions.

Polyglot, qui a tendance à segmenter en plus petites unités, a réalisé une segmentation parfaite.

4.3.9 Composés chimiques : lexique spécialisé

Exemple 1

Séquence à segmenter : 可以升高内源性胰高血糖素样肽-1（Glucagon-like Peptide-1，GLP-1)

Signification: pourrait augmenter le Glucagon-like Peptide-1 endogène.

Difficultés de segmentation :

Cette phrase contient le nom d'un composé chimique transcrit en chinois avec des caractères porteurs de sémantique : pancréas – élevé – sang - sucre – élément – type – peptide. Le composé 胰高血糖素样肽 */ yi gao xue tang su yang tai / Glucagon-semblable à peptide* comporte 7 caractères et un chiffre, ce qui est d'une longueur inhabituelle pour un mot chinois. Il pourrait être isolé grâce à la présence avant lui d'un adjectif (内源性 */ nei yuan xing / endogène*) et au tiret après lui qui signe forcément la fin d'un mot. Le mot est absent du dictionnaire du chinois contemporain, mais présent tel quel dans le dictionnaire en ligne non spécialisé LINEDict[23] .

23 http://ce.linedict.com/dict.html#/cnen/entry/a58d831f1fe9460b942e2476f85eaacd

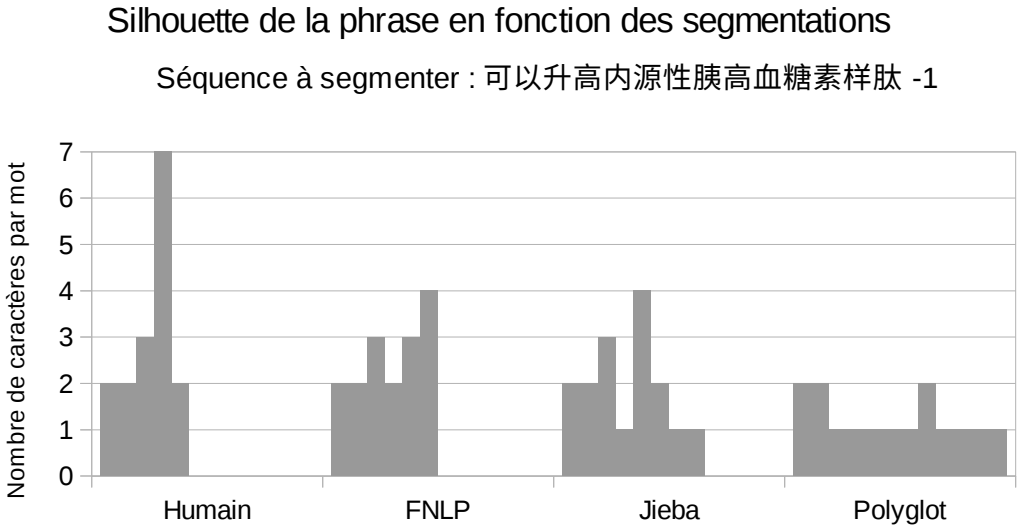

Illustration 14: Silhouettes – Lexique spécialisé, composé chimique

Segmentations obtenues

Humaine	可 以 | 升 高 | 内 源 性 | 胰 高 血 糖 素 样 肽 | - 1
FNLP	可 以 | 升 高 | 内 源 性 | 胰 高 | 血 糖 素 | 样 肽 - 1
Jieba	可 以 | 升 高 | 内 源 性 | 胰 | 高 血 糖 素 | 样 肽 | - | 1
Polyglot	可 以 | 升 高 | 内 | 源 | 性 | 胰 | 高 | 血 糖 | 素 | 样 | 肽 | - | 1

Tableau 23: Divergences de segmentation – Lexique spécialisé, composé chimique

Observations

Aucun des trois segmenteurs ne l'a identifié comme une seule unité lexicale et de plus, les trois proposent une segmentation différente. Cependant, FNLP et Jieba ont identifié correctement l'adjectif 内源性 */ nei yuan xing / endogène*.

Exemple 2

Séquence à segmenter : 今日由拜糖平换了米格列醇

Signification: Aujourd'hui j'ai remplacé le Miglitol par l'Acarbose.

Difficultés de segmentation :

L'exemple ci-dessous est une phrase très courte contenant deux noms de médicaments de 3 et 4 caractères respectivement, donc une longueur de mot habituelle pour le chinois contemporain.

Chaque nom est délimité pour l'un par une préposition et un verbe, pour l'autre par une particule et une virgule. Le premier nom 拜糖平 */ bai tang ping / Acarbose* est un nom commercial fabriqué pour sonner plus avantageusement que la transcription phonétique précédente du même produit 阿卡波糖 */ A ka bo tang / Acarbose*. Contrairement à la transcription phonétique, les caractères de 拜糖平 */ bai tang ping / Acarbose* sont porteurs de sémantique : ils signifient plus ou moins "rendre grâce - niveau de sucre" et sont plus facilement identifiés pour leur sens individuel que comme unité trisyllabique. Le deuxième nom 米格列醇 */ mi ge lie chun / miglitol* est la transcription phonétique d'un mot étranger et ne transporte pas de sémantique.

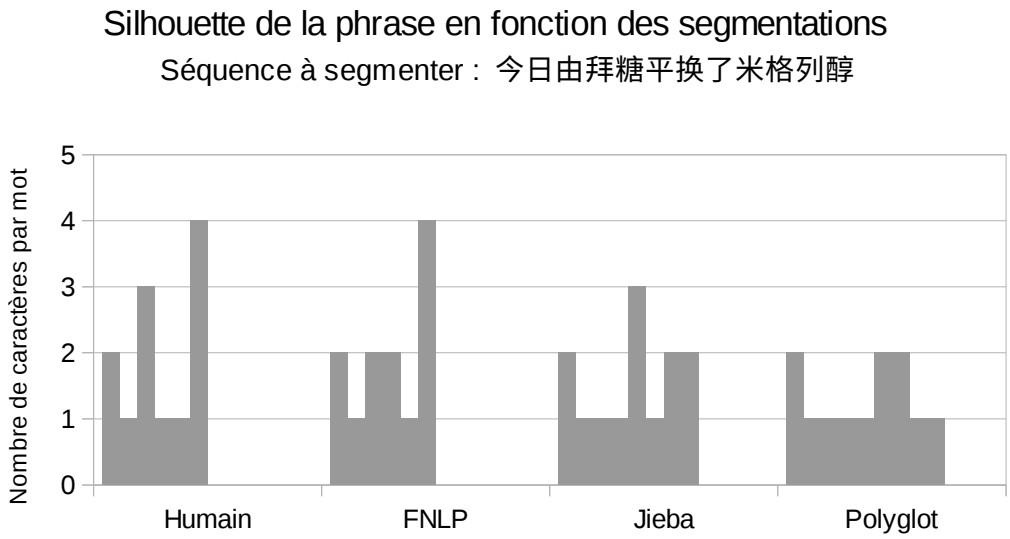

Illustration 15: Silhouettes - Noms de médicaments

Segmentations obtenues

Humaine	今 日 由 拜 糖 平 换 了 米 格 列 醇
FNLP	今 日 由 拜 糖 平 换 了 米 格 列 醇
Jieba	今 日 由 拜 糖 平 换 了 米 格 列 醇
Polyglot	今 日 由 拜 糖 平 换 了 米 格 列 醇

Tableau 24: Divergences de segmentations - Noms de médicaments

Observations

Aucun des trois segmenteurs n'a identifié les deux noms de médicaments comme des mots, seul Fudan a isolé le 米格列醇 */ mi ge lie chun / miglitol*. FNLP et Jieba ont détaché le dernier caractère du premier nom pour le rattacher au verbe (monosyllabique) et extraire de cette phrase l'expression moderne et non pertinente quant au contexte 平换 */ ping huan / échange d'appartement*.

4.4 Synthèse des résultats

Le tableau ci-dessous récapitule pour chacun des exemples analysés et sur chaque erreur rencontrée, comment les trois segmenteurs ont traité le problème : Ok signifie que l'objet linguistique (Numéral, mot étranger...) a été correctement identifié.

	Objet linguistique	FNLP	Jieba	Polyglot
Proverbes à quatre caractères – Exemple 1	Chengyu 1	Ok	Ok	Ok
	Chengyu 2	-	Ok	-
	négation	Ok	-	-
Proverbes à quatre caractères – Exemple 2	Chengyu	-	Ok	Ok
Formes verbales et Déterminant + déterminé	Négation	Ok	-	-
	Verbe Objet	-	-	Ok
	Déterminant Déterminé	Ok	-	Ok
Adjectif redupliqué	ADJ AABB	-	Ok	-
	Négation	Ok	-	-
	Verbe+prép	-	-	Ok
Verbe Réduplíqué	Verbe Objet	-	-	Ok
	Verbe AAB	Ok	-	-
	Verbe AA	-	Ok	-
Marathon	Verbe Objet	-	Ok	Ok
	Mot étranger	-	Ok	-
Pourcentage	Numéraux	Ok	Ok	-
Numéraux et jargon spécialisé	Numéral 1	-	-	Ok
	Numéral 2	-	-	Ok
	Conjonction	Ok	-	Ok
	Verbe+prép	-	Ok	Ok
Composés chimiques - Glucagon	Adjectif trisyllabique	Ok	Ok	-
	Glucagon	-	-	-
Composés chimiques - deux médicaments	Nom 1	-	-	-
	Verbe	-	-	Ok
	Nom2	Ok	-	-

Tableau 25: Synthèse des résultats sur les erreurs analysées

Le tableau permet surtout de montrer que les trois outils ont tous des failles, ce qui n'est pas une surprise. Cependant, il permet aussi de mettre en évidence une autre lecture, par "objet linguistique", pour lire les performances des segmenteurs. C'est cette lecture qui va servir de base à une recommandation pour aider au choix d'un segmenteur.

Par contre, il n'est pas possible à partir de ces quelques exemples de dégager une tendance générale par segmenteur : pour la négation 不 / *bu* par exemple, on a vu dans les spécificités que FNLP en identifie plus fréquemment que Polyglot et de fait dans les exemples analysés, la segmentation de FNLP est correcte alors que celle de Polyglot est incorrecte. Comment utiliser ce repérage pour généraliser et automatiser la comparaison des performances des segmenteurs sur certains objets linguistiques et non en globalité ? C'est l'objet de la proposition de démarche faite au chapitre 5.

5 Segmenter parfois, segmenter comment

Comme on l'a vu dans les analyses et exemples ci-dessus, chacun des outils a commis des erreurs d'interprétation, susceptibles de démultiplier les erreurs des traitements suivants. Impossible donc d'en recommander un comme un système clé en main : ce serait un peu comme proposer une seule taille de vêtement pour tout le monde. Quelle que soit la tâche qu'on se propose de réaliser sur un texte chinois nécessitant une segmentation en mots préalable, il faut prévoir le choix du segmenteur comme une étape préalable et une tâche à part entière, comportant plusieurs étapes, avec des cycles d'essais et d'amélioration avant le choix définitif. Je propose ci-dessous quelques étapes très pragmatiques pour arriver à ce choix.

5.1 Synthèse - Conséquences sur les tâches post-segmentation

La segmentation en mots n'est pas une fin en soi ; elle n'est même pas indispensable pour toutes les tâches de traitement automatique des langues.

Pour la **catégorisation automatique** par exemple, (Ke et Zweigenbaum 2009) ne trouvent aucune différence de résultats avec une segmentation en mots ou une segmentation en caractères.

De même pour la **recherche d'informations,** de nombreux travaux montrent l'efficacité de travailler directement sur les séquences de caractères sans tenter de segmenter les textes en mots.

Par contre, il est clair que les tâches nécessitant de restituer de la sémantique peuvent pâtir d'une mauvaise segmentation en mots.

Pour **la traduction automatique** la tâche est cruciale. Comme le note (Wu 2011), "les erreurs de segmentation sont amplifiées par la traduction automatique", car une erreur sur la segmentation d'un mot peut causer une traduction erronée d'une phrase complète. Des travaux spécifiques sont menés par Systran (Yang, Senellart, et Zajac 2003), qui a développé successivement plusieurs méthodes de segmentation en mots, à base de dictionnaires, d'automates et de règles linguistiques. Le modèle présenté dans l'article utilise des lexiques bilingues généralistes et spécialisés par domaines, et des règles de calcul élaborées qui optimisent la segmentation au niveau de la phrase. On est donc loin des outils opensource accessibles à tout un chacun.

Les fonctionnalités d'**extraction de mots-clés**, qui reposent principalement sur des calculs de fréquence de mots sont bien évidemment inopérantes dès lors que les mots sont mal découpés. Les

simples graphiques de spécificités présentées en 4.2 , si on se souvient qu'elles concernent un micro-corpus, donnent une idée assez précise des dérives possibles en fonction des choix de segmentation. Les services de **résumé automatique de textes**, qui utilisent parfois l'extraction de phrases entières, tiennent compte aussi de la fréquence des mots.

Une abondante littérature traite de la **reconnaissance des Entités Nommées**, étroitement liée à la segmentation en mots et à la transcription des mots étrangers. En réalité, de nombreux travaux contournent la segmentation en mots, par exemple (Zong et al. 2010) dans "Named Entity Resolution in Chinese News Comments on the Web" utilisent les entités nommées repérées dans des articles pour chercher celles ci dans les commentaires : ils se ramènent donc à un problème de recherche d'information. Quant à (Ji et al. 2006) ils concluent que "l'extraction terminologique basée sur les caractères donne de bien meilleurs résultats que l'utilisation de la segmentation en mots comme pré-traitement".

L'analyse grammaticale de textes chinois, qui passerait par une segmentation en mots suivie d'un **étiquetage morpho syntaxique** est un sujet intéressant : l'expérience montre que le POS tagging est en lui même un apport à la segmentation en mots, puisque l'analyse de la probabilité d'une séquence syntaxique donne de bons indices sur la validité de la segmentation en mots. Cependant, la fluidité des catégories morpho syntaxiques du Chinois, et la faculté de supprimer certains éléments dans une phrase si le contexte est clair pour un lecteur humain, rendent ces tâches peu robustes.

5.2 Proposition de démarche pour le choix d'un outil de segmentation

5.2.1 Faire une short-list des outils à tester

Soit en utilisant des critères comme ceux que j'ai décrits plus haut, soit en les déterminant soi-même, sélectionner deux à quatre outils à tester sur ses propres données. L'utilisation de la plate forme MultiTal[24] peut faire gagner beaucoup de temps à cette étape, puisque les outils y sont déjà repérés, essayés et décrits.

5.2.2 Définir ses priorités de traitement

Suivant ses objectifs, il faut définir les objets linguistiques dont on veut privilégier le bon traitement : par exemple, peut-on sacrifier le traitement des négations si celui des numéraux est optimal ? Est-il indispensable de repérer un maximum d'expressions proverbiales ? Quel traitement souhaité pour les néologismes ?

5.2.3 Repérer les tendances de chaque outil

Échantillonner le corpus

Extraire du corpus brut une portion réduite mais significative pour tester dessus les segmenteurs ciblés : suivant la technique d'échantillonnage classique, on prendra soin d'extraire au hasard des séquences de texte dans chacun des types de texte qu'on s'apprête à traiter, notamment s'il s'agit de textes hétérogènes en terme d'origine, de style, de graphies et de lexiques.

24 http://helium.lab.parisdescartes.fr:2230/browse

Segmenter et annoter manuellement l'échantillon

Cette étape s'avère finalement indispensable pour généraliser les observations faites par une première approche comme celle qui est décrite au chapitre 4. En effet, sans référence, impossible d'estimer la distance de chaque résultat au résultat attendu. Pour les expressions quadrisyllabiques par exemple, on a vu ci-dessus que sur les trois analysées, Jieba en a identifié trois, Polyglot deux et FNLP une seule : mais cela ne nous dit rien du taux de réussite sur le total des expressions présentes dans le corpus. Il est donc indispensable de combiner une approche qualitative comme décrite ci-dessus avec une approche quantitative par objet d'étude.

Faire passer les outils et analyser les résultats

Sur l'échantillon brut, éventuellement normalisé (ponctuations, jeux de caractères), faire passer les outils choisis de façon à obtenir un résultat homogène. D'un point de vue pratique, il s'agit de paramétrer les outils pour faciliter les traitements suivants. Ce n'est pas toujours trivial, car parmi les outils testés pour MultiTal, les formats de sortie sont très hétéroclites ; certains sont pensés comme des chaînes de traitement et laissent difficilement à l'utilisateur le choix des tâches à exécuter.

En fonction des priorités de traitement et des objets linguistiques repérés dans l'échantillon, il s'agit de comparer avec le corpus manuellement segmenté et annoté de façon à calculer la distance de chaque segmenteur au résultat attendu. Par cette méthode, on obtient un résultat généralisé à l'ensemble de l'échantillon et non des exemples isolés. Pour une lisibilité immédiate, on peut construire un diagramme à plusieurs axes, correspondant aux objets linguistiques annotés, et positionner les résultats de chaque outil.

5.2.4 Améliorer les résultats

Suivant les résultats obtenus et les possibilités des segmenteurs testés, on peut tenter d'améliorer les résultats. Chaque outil offre des options de paramétrage, pas toujours homogènes entre elles ; les plus courantes sont :

- ajout de lexiques : pour les textes traitant d'un domaine spécifique, l'ajout de ressources lexicales ad hoc améliore radicalement les performances par exemple.

- réglage de la taille maximale des unités lexicales

- sens de la segmentation : parcourt des séquences de caractères de gauche à droite ou de droite à gauche

- taille des séquences analysées pour les calculs de probabilité de plusieurs segmentations séquence par séquence

Les étapes d'analyse et d'amélioration sont bien entendu à répéter autant que nécessaire (ou autant que possible) avant d'arrêter son choix.

6 Conclusion

6.1 Il n'y a pas de meilleur segmenteur du Chinois

Cette rapide comparaison de trois outils de segmentation habituellement regardés comme fiables, appuyée par la consultation de quelques articles du domaine, amène à la conclusion qu'il n'existe pas de "meilleur" outil de segmentation clés en mains.

De plus, il est primordial de bien comprendre que le choix d'un outil dépend avant tout du besoin de traitement à réaliser. Ensuite, il faut être conscient de la nécessité d'adapter les outils à ses objectifs, par des tests, des paramétrages et la fournitures de ressources lexicales ciblées.

Mais le plus important, c'est peut-être de commencer par remettre en question la nécessité et la pertinence même de la segmentation en mots du Chinois.

6.2 Et si le chinois était autre chose que de l'anglais écrit serré ?

L'intérêt principal de segmenter le Chinois en mots est d'utiliser pour son traitement les méthodes et les outils déjà mis au point pour traiter les langues alphabétiques occidentales, notamment l'anglais qui est la mieux dotée, en faisant abstraction du système d'écriture particulier des caractères. C'est une approche efficace pour certaines tâches, mais limitée voire handicapante pour d'autres travaux de traitement automatique. En effet en ignorant les particularités du Chinois dans son traitement, on ignore aussi une grande part de l'information contenue dans les textes.

Cette approche partant des particularités de la langue à traiter au lieu de reproduire les mécanismes de traitement d'une autre langue est très bien présentée entre l'Anglais et le Thai par (Aroonmanakun et others 2007) ; le Thai utilise une écriture alphabétqiue mais comme en Chinois, il n'y a pas d'espace entre les mots en Thai, et le problème des mots composés et de leur segmentation s'y pose également ; de plus, même les frontières des phrases sont floues. Aroonmanakun propose donc une méthode de détermination des unités lexicales propre au Thai et ne reposant pas sur des règles orthographiques utilisées pour l'Anglais ou le Français.

Une des possibilités d'exploiter davantage d'information dans un texte Chinois est analysée par (Meishan Zhang et al. 2013) ; l'article montre comment les caractères chinois interagissent à l'intérieur des mots comme les mots à l'intérieur d'une phrase. Les auteurs prennent aussi en compte la charge sémantique portée par chaque caractère pour proposer un système d'étiquetage morpho syntaxique propre aux caractéristiques linguistiques du Chinois et qui se révélerait plus puissant que les systèmes à base de mots.

Sans renier l'importance des travaux déjà réalisés, ni l'utilité des systèmes disponibles, il semble certain que les systèmes de traitement automatique conçus en envisageant les particularités du Chinois comme des forces et non comme des inconvénients sont une piste à suivre pour concevoir des systèmes nettement plus performants.

7 Annexes

7.1 Tableaux

Index des tableaux

7.2 Graphiques

Index des illustrations

7.3 Extraits de corpus et de code

7.3.1 Extrait de corpus

Le corpus brut

吃的东西有关

好多天了，吃什么基本都是这样

米，面条，馒头都是这样，郁闷

医生给我开的药其中有瑞格列奈片和盐酸二甲双胍肠溶片，盐酸吡格列酮片。我在论坛里看见有帖子上说吃促分泌药，吃久了会没有效果是真的吗?

无根据

谢谢

Le corpus segmenté par FNLP

吃 的 东西 有关

好多 天 了 ， 吃 什么 基本 都 是 这样

米 ， 面条 ， 馒头 都 是 这样 ， 郁闷

医生 给 我 开 的 药 其中 有 瑞格列 奈片 和 盐酸二甲 双胍肠 溶片 ， 盐酸吡 格列酮 片 。 我 在 论坛 里 看见 有 帖子 上 说 吃促 分泌药 ， 吃久 了 会 没有 效果 是 真的 吗 ？

无 根据

谢谢

Le corpus segmenté par Jieba

吃 的 东西 有关

好多天 了 ， 吃 什么 基本 都 是 这样

米 ， 面条 ， 馒头 都 是 这样 ， 郁闷

医生 给 我 开 的 药 其中 有 瑞格 列奈片 和 盐酸 二甲 双 胍 肠溶片 ， 盐酸 吡 格 列 酮 片 。 我 在 论坛 里 看见 有 帖子 上 说 吃 促 分泌 药 ， 吃久 了 会 没有 效果 是 真的 吗 ？

无 根据

谢谢

Le corpus segmenté par Polyglot

吃 的 东西 有关

好多 天 了 ， 吃 什么 基本 都是 这样

米 ， 面条 ， 馒头 都是 这样 ， 郁闷

医生 给 我 开 的 药 其中 有 瑞 格 列 奈 片 和 盐酸 二甲 双 胍 肠 溶 片 ， 盐酸 吡 格 列 酮 片 。 我在 论坛 里 看见 有 帖子 上 说 吃 促 分泌 药 ， 吃 久 了 会 没有 效果 是 真的 吗 ？

无 根据

谢谢

7.3.2 Extraits de code

Récupération du texte des contribution à partir d'une liste d'URL

```
use Web::Scraper;
use Encode;

my $c = 1 ;

open (IN, "<:encoding(UTF-8)","threadUrl.txt") || die $! ;
open (OUT, ">>listeposts.txt") ;

while (my $ligne=<IN>)
{

        chomp $ligne ;
        my $url="$ligne";
        # réinitialiser à vide avant de crawler chaque URL :
        my $tout ;
        my $resultat = scraper
        {
        process '//div[@id="postlist"]/div[starts-with(@id,"post_")]', 'entree[]' => scraper{

                process '//td[starts-with(@id,"postmessage")]//text()[not(ancestor::i or ancestor::blockquote)]', 'contenu[]' => 'TEXT' ;
                }
        };
```

Instructions de remplacement de ponctuation

```
cat 1000lignes.txt | sed
"s/? /inter/g;s/\./num/g;s/! /excla/g;s/。/ptfinal/g;s/?/inter/g;s/\
. /num/g;/^$/d" > 1000lignes_nopunct.txt
```

Instruction de recherche de quadrisyllabiques

```
egrep -o '\s[^[:punct:][:space:][:digit:]]{4}\s|\s[^[:punct:]
[:space:][:digit:]]{4}$' ok_*.txt
```

7.4 Bibliographie

Al-Rfou, Rami, Bryan Perozzi, et Steven Skiena. 2013. « Polyglot: Distributed word representations for multilingual nlp ». *arXiv preprint arXiv:1307.1662*. http://arxiv.org/abs/1307.1662.

Aroonmanakun, Wirote, et others. 2007. « Thoughts on word and sentence segmentation in Thai ». In *Proceedings of the Seventh Symposium on Natural language Processing, Pattaya, Thailand, December 13–15*, 85–90.

Chen, Miaohong, Baobao Chang, et Wenzhe Pei. 2014. « A joint model for unsupervised Chinese word segmentation ». http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.684.694.

Chih-Hao, Tsai. 2000. *MMSEG: a word identification system for mandarin chinese text based on two variants of the maximum matching algorithm*.

Choi, Key-Sun, Hitoshi Isahara, Kyoko Kanzaki, Hansaem Kim, Seok Mun Pak, et Maosong Sun. 2009. « Word segmentation standard in Chinese, Japanese and Korean ». In *Proceedings of*

the 7th Workshop on Asian Language Resources, 179–186. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1690325.
Darrobers, Robert. 1998. *Eléments fondamentaux de la phrase chinoise*. Paris: Éditions You-Feng.
Dong, Zhendong, et Qiang Dong. 2006. *HowNet and the Computation of Meaning*. Hackensack, NJ: World Scientific.
Drocourt, Zhitang, et Alain Peyraube. 2007. *Parlons chinois*. Paris, France, Pays multiples.
Harris, Zellig S. 1955. *From Phoneme to Morpheme*. Indianapolis, Ind.: Bobbs-Merrill.
Huang, Chu-Ren, Petr Šimon, Shu-Kai Hsieh, et Laurent Prévot. 2007. « Rethinking Chinese word segmentation: tokenization, character classification, or wordbreak identification ». In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 69–72. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1557791.
Ji, Luning, Qin Lu, Wenjie Li, et YiRong Chen. 2006. « A Comparative Study of the Effect of Word Segmentation On Chinese Terminology Extraction ». In *Proceedings of The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC 2006), pp101-108*. https://www.researchgate.net/profile/Qin_Lu3/publication/33017613_A_Comparative_Study_of_the_Effect_of_Word_Segmentation_On_Chinese_Terminology_Extraction/links/02e7e5269e68d87b15000000.pdf.
Ke, Guiyao, et Pierre Zweigenbaum. 2009. « Catégorisation automatique de pages web chinoises - documents spécialisés vs grand public sur le tabagisme ». In *COnférence en Recherche d'Infomations et Applications - CORIA 2009, 6th French Information Retrieval Conference, Presqu'île de Giens, France, May 5-7, 2009. Proceedings*, 203–218. LSIS-USTV. http://asso-aria.org/coria/2009/203.pdf.
La Robertie, Pierre de. 2005. « Le nom propre en chinois. Essai de morphosyntaxe ». *Corela. Cognition, représentation, langage*, nº HS-2. http://corela.revues.org/1187.
Lafon, Pierre. 1980. « Sur la variabilité de la fréquence des formes dans un corpus ». *Mots* 1 (1): 127-65. doi:10.3406/mots.1980.1008.
Le Chinois Pour Tous Bescherelle Edition. 2014. Educa Books.
Magistry, Pierre. 2012. « Segmentation non supervisée: le cas du mandarin ». *JEP-TALN-RECITAL 2012*, 1. http://www.anthology.aclweb.org/F/F12/F12-3001.pdf.
Meishan Zhang, Yue Zhang, Wanxiang Che, et Ting Liu. 2013. « Chinese Parsing Exploiting Characters ». doi:10.13140/2.1.3839.8727.
Organisation internationale de normalisation, Terminologie et autres ressources langagières et ressources de contenu Comité technique ISO/TC 37, et Gestion des ressources linguistiques Sous-comité SC 4. 2010. *Language Resource Management: Word Segmentation of Written Texts = Gestion Des Ressources Langagières : Segmentation Des Mots Dans Les Textes Écrits*. Geneva: International Standard Organization.
Qiu, Xipeng, Qi Zhang, et Xuanjing Huang. 2013. « FudanNLP: A Toolkit for Chinese Natural Language Processing ». In *ResearchGate*, 49-54. https://www.researchgate.net/publication/270878350_FudanNLP_A_Toolkit_for_Chinese_Natural_Language_Processing.
Rabut, Isabelle, Wu Yongyi, et Liu Hong. 2012. *Méthode de chinois: premier niveau*. Paris: L'Asiathèque.
« resume-multital.pdf ». 2016. Consulté le octobre 25. http://www.er-tim.fr/sites/default/files/resume-multital.pdf.
Wu, Zhijie. 2011. « A Cognitive Model of Chinese Word Segmentation for Machine Translation ». *Meta: Journal des traducteurs* 56 (3): 631. doi:10.7202/1008337ar.
Yang, Jin, Jean Senellart, et Remi Zajac. 2003. « Systran's Chinese word segmentation ». In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*,

180–183. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1119279.

Yu ming shan, éd. 2013. 现代汉语词典. 北京, Chine: 华语教学出版社, 2013Bei jing : Hua yu jiao xue chu ban she.

Yushi, Yao, et Huang Zheng. 2015. « Combine CRF and MMSEG to Boost Chinese Word Segmentation in Social Media ». *arXiv preprint arXiv:1510.07099*. http://arxiv.org/abs/1510.07099.

Zong, Liang, Xiaojun Wan, Lihong Zhao, Jianwu Yang, et Yuqian Wu. 2010. « Named Entity Resolution in Chinese News Comments on the Web ». In , 307-13. IEEE. doi:10.1109/APWeb.2010.20.

于明善, 杨东, et 说词解字辞书研究中心. 2012. 插图本小学生成语学习词典. 北京: 华语教学出版社.

刘源, 谭强, et 沈旭昆. 1994. 信息处理用现代汉语分词规范及自动分词方法. 北京: 清华大学出版社.