TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Segmentation et induction de lexique non-supervisées du mandarin

Pierre Magistry Benoît Sagot
Alpage, INRIA Paris–Rocquencourt & Université Paris 7,
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
{pierre.magistry,benoit.sagot}@inria.fr

Résumé. Pour la plupart des langues utilisant l'alphabet latin, le découpage d'un texte selon les espaces et les symboles de ponctuation est une bonne approximation d'un découpage en unités lexicales. Bien que cette approximation cache de nombreuses difficultés, elles sont sans comparaison avec celles que l'on rencontre lorsque l'on veut traiter des langues qui, comme le chinois mandarin, n'utilisent pas l'espace. Un grand nombre de systèmes de segmentation ont été proposés parmi lesquels certains adoptent une approche non-supervisée motivée linguistiquement. Cependant les méthodes d'évaluation communément utilisées ne rendent pas compte de toutes les propriétés de tels systèmes. Dans cet article, nous montrons qu'un modèle simple qui repose sur une reformulation en termes d'entropie d'une hypothèse indépendante de la langue énoncée par Harris (1955), permet de segmenter un corpus et d'en extraire un lexique. Testé sur le corpus de l'Academia Sinica, notre système permet l'induction d'une segmentation et d'un lexique qui ont de bonnes propriétés intrinsèques et dont les caractéristiques sont similaires à celles du lexique sous-jacent au corpus segmenté manuellement. De plus, on constate une certaine corrélation entre les résultats du modèle de segmentation et les structures syntaxiques fournies par une sous-partie arborée corpus.

Abstract. For most languages using the Latin alphabet, tokenizing a text on spaces and punctuation marks is a good approximation of a segmentation into lexical units. Although this approximation hides many difficulties, they do not compare with those arising when dealing with languages that do not use spaces, such as Mandarin Chinese. Many segmentation systems have been proposed, some of them use linguistitically motivated unsupervized algorithms. However, standard evaluation practices fail to account for some properties of such systems. In this paper, we show that a simple model, based on an entropy-based reformulation of a language-independent hypothesis put forward by Harris (1955), allows for segmenting a corpus and extracting a lexicon from the results. Tested on the Academia Sinica Corpus, our system allows for inducing a segmentation and a lexicon with good intrinsic properties and whose characteristics are similar to those of the lexicon underlying the manually-segmented corpus. Moreover, the results of the segmentation model correlate with the syntactic structures provided by the syntactically annotated subpart of the corpus.

Mots-clés : Segmentation non-supervisée, entropie, induction de lexique, unité lexicale, chinois mandarin.

Keywords: Non-supervized segmentation, entropy, lexicon induction, Mandarin Chinese.

1 Introduction

La segmentation d'un texte en formes [1] est la première étape de presque tout traitement automatique de données textuelles. Pour la plupart des langues utilisant l'alphabet latin, dont le français ou l'anglais, un découpage selon les espaces et les symboles de ponctuation est une bonne approximation d'une segmentation en unités lexicales. À l'inverse, dans le cas des systèmes d'écriture utilisés par exemple pour écrire le chinois, le japonais, le thai, le khmer ou le vietnamien, la typographie n'est pas utilisée pour indiquer des frontières entre les mêmes unités linguistiques : en vietnamien, qui utilise une variante de l'alphabet latin, l'espace sépare des unités sous-lexicales. En chinois ou japonais, seuls les signes de ponctuation indiquent des frontières entre unités lexicales ; ailleurs, les caractères, qui représentent aussi des unités sous-lexicales, sont directement juxtaposés. L'étape de segmentation en unités lexicales est donc un problème délicat pour ces langues dites *non-segmentées*, et donne lieu à une littérature

1. Dans cet article, une *forme* est un segment continu de texte venant occuper de façon autonome une position syntaxique. Travaillant sur le mandarin, nous pouvons faire l'approximation qu'il y a identité entre la notion de forme et celle d'*unité lexicale*, Pour une discussion plus détaillées de l'unité lexicale en mandarin, se reporter à Packard (2000) ou en français à Nguyen (2006).

abondante (Zhao & Liu, 2010), y compris dans la communauté francophone (Seng *et al.*, 2009; Wu, 2010). Mais de tels travaux peuvent aussi être utiles pour les langues *segmentées*, en raison des cas de non-correspondance entre séparateurs et frontières d'unités lexicales, lesquelles restent difficiles à définir et à repérer quelle que soit la langue (Zhikov *et al.*, 2010).

Parmi les méthodes de segmentation, nous nous intéressons en particulier aux méthodes non supervisées qui cherchent une définition implicite du mot en faisant émerger la segmentation à partir des propriétés non-aléatoires de la distribution des formes en corpus. Ces méthodes sont difficiles à évaluer car elles ne s'adaptent pas à un standard donné. En contrepartie, elles présentent un plus grand potentiel d'adaptation à la dynamique d'une langue (changement de domaine, variantes géographiques, évolution diachronique, traitement des néologismes), et peuvent être utilisées pour la segmentation de langues peu ou pas dotées.

Nous décrivons ici une série d'expériences de segmentation en unités lexicales réalisées sur le chinois mandarin au moyen d'un système non-supervisé qui repose sur une hypothèse motivée linguistiquement formulée par (Harris, 1955), en adaptant sa modélisation présentée par (Tanaka-Ishii, 2005) dans le même but (Jin & Tanaka-Ishii, 2006). Nous insistons en particulier sur l'évaluation des résultats obtenus, tâche rendue délicate par la nature non-supervisée de l'approche et par la variété des conventions de segmentation qui existent pour le chinois mandarin.

Dans la section suivante nous présentons la tâche de segmentation et les problèmes que posent la méthode traditionnellement utilisée pour son évaluation. Les sections 3 et 4 présentent les systèmes dont nous nous inspirons et celui que nous avons développé. Nous utilisons ensuite notre système de segmentation pour extraire un lexique, dont nous proposons une évaluation (section 5). Enfin nous cherchons à corréler la sortie du système de segmentation étudié avec des informations syntaxiques extraites d'un corpus arboré.

2 La segmentation du chinois

La segmentation est la première étape de tout système d'analyse automatique du chinois écrit. En français et dans la majorité des langues utilisant l'alphabet latin, un découpage sur les espaces (et autour des signes de ponctuation), souvent appelé *tokenisation* et dont la sortie est un flux de *tokens*, constitue une première étape raisonnable, que l'on peut ensuite affiner pour identifier les cas de non-alignement entre *tokens* et *formes* (qui peuvent par exemple être des formes composées). À l'inverse, l'écriture chinoise ne comporte pas de séparateur typographique comme l'espace. Un découpage effectué uniquement autour des caractères de ponctuation produirait des *tokens* bien plus longs que des formes. À l'inverse, un découpage isolant chaque caractère chinois (ci-après *sinogramme*) ressemblerait plutôt à une segmentation en morph(ème)es qu'en formes. Il faut donc considérer un texte en chinois comme un flux de sinogrammes, la tâche de segmentation consistant à identifier entre quels sinogrammes il faut segmenter le texte afin de délimiter les formes, que l'on peut, en mandarin, assimiler à des *unités lexicales*.

2.1 État de l'art, enjeux actuels

Un grand nombre de méthodes ont été proposées pour effectuer une segmentation automatique. Certaines reposent sur des règles et des lexiques, d'autres utilisent des méthodes d'apprentissage automatique supervisé ou non-supervisé. Cinq campagnes du « *Chinese Word Segmentation Bakeoff* » ont été organisés par l'ACL, dont la dernière s'est tenue à l'été 2010. Zhao & Liu (2010) donnent un résumé des performances obtenues par les systèmes en compétition. Ils soulignent que si la précision peut sembler satisfaisante, la tolérance au changement de domaine et la reconnaissance des mots inconnus restent les limitations majeures.

Notons que lors de cette campagne, le système de base (*baseline*) et le meilleur système (*topline*) sont obtenus avec le même algorithme, un simple *maximum-matching* (*minimisation du nombre de mots*) reposant sur un inventaire d'unités lexicales, et ne se distinguent que par le lexique utilisé : la *baseline* utilise un lexique extrait à partir du corpus d'entraînement, tandis que la *topline* utilise un lexique extrait à partir de la totalité du corpus et connaît donc tous les formes attendues. Xue (2003), qui présentait un système d'apprentissage supervisé reposant sur une classification IOB (*Inside, Outside, Begin)* des sinogrammes, commente les résultats d'une autre heuristique simple qui repose sur un lexique, celle dite du *longest-match* gauche-droite (*plus longue chaîne d'abord*) : cette heuristique fournit de très bons résultats (f-mesure 0,952) si le lexique est exhaustif mais se dégrade très rapidement lorsque le corpus de test contient des mots inconnus (f-mesure de 0,898). Le *maximum-matching* utilisé lors du *bakeoff* obtient quant à lui des scores (f-mesure) supérieurs à 0,98 sur différents corpus (la *topline*) avec un lexique exhaustif,

et des scores de 0,72 à 0,88 selon les domaines dans la configuration *baseline*. Les 18 systèmes présentés lors du *segmentation bakeoff* ont tous obtenu des résultats intermédiaires entre ces deux niveaux. Il faut donc souligner l'importance pour cette tâche des ressources lexicales.

Parmi ces systèmes, et en général parmi les systèmes de segmentation du chinois mandarin, les deux principaux paradigmes d'apprentissage automatique ont été utilisés. Chacun présente des avantages et des inconvénients.

Les méthodes supervisées nécessitent un corpus d'entraînement constitué d'un ensemble de textes déjà segmentés (la réponse attendue, considérée comme « bonne »). À partir de ce jeu d'exemples, l'algorithme effectue une généralisation qui lui permet ensuite d'imiter la prise de décision effectuée par l'humain lors de la segmentation manuelle du corpus. De nombreuses méthodes d'apprentissage supervisé existent et les systèmes de segmentation actuels tendent à les combiner (cf. par exemple celui décrit par (Wu *et al.*, 2010), très bien classé au dernier *segmentation bakeoff*, qui repose sur un « *conditional support Markov model* »). Les méthodes supervisées sont celles qui obtiennent les meilleurs résultats, mais elles nécessitent l'utilisation d'un corpus d'entraînement dont la construction est longue et coûteuse. Ce corpus influe sur le comportement des systèmes qui dépendent de choix linguistiques particuliers, ainsi que de la nature du corpus (l'état de la langue à une époque donnée et pour un domaine donné). L'adaptation à d'autres domaines est un enjeu de recherche pour ce type de système.

Les méthodes non-supervisées n'utilisent pas de corpus pré-annoté mais se contentent d'une grande quantité de données brutes non-segmentées. L'hypothèse sous-jacente est que les données ne sont pas distribuées aléatoirement mais possèdent une certaine structure que l'on cherche à faire émerger par l'analyse de leur distribution. Parmi les méthodes utilisées pour la segmentation du chinois, on peut citer des approches utilisant l'information mutuelle, comme dans les travaux pionniers de Sproat *et al.* (1996), puis plus récemment des méthodes reposant sur l'algorithme *Expectation Maximization* (Peng & Schuurmans, 2001), ou sur la *Minimum Description Length* (Hua, 2000). La *complexité contextuelle*, inspirée des hypothèses de Harris (Harris, 1955) et utilisée dans les travaux de (Tanaka-Ishii & Jin, 2006; Jin & Tanaka-Ishii, 2006) est présentée plus en détail à la section 3. Les méthodes non supervisées présentent l'avantage de s'adapter à un corpus brut peu coûteux et plus facile à obtenir que des corpus segmentés manuellement. Mais elles sont difficiles à évaluer : il n'existe pas à priori de raison pour que la sortie d'un tel système corresponde à un guide de segmentation plutôt qu'à un autre.

2.2 Méthodes d'évaluation des systèmes de segmentation

Les différents systèmes de segmentation sont entraînés et évalués sur des parties de corpus dits « de référence » en utilisant les mesures classiques en apprentissage automatique (rappel, précision, f-mesure sur les formes ou sur les frontières), mais ce mode d'évaluation sous-estime une réalité linguistique complexe. Les différents corpus disponibles segmentés manuellement ne suivent pas les mêmes guides d'annotation. Ainsi le corpus de l'Université de Pékin suit le guide de Yu (1999) tandis que le corpus équilibré de l'*Academia Sinica* suit Huang *et al.* (1996) et que le Chinese Treebank respecte les conventions de Xia (2000).

Il a été plusieurs fois observé que le taux d'accord entre locuteurs natifs non linguistes à qui il était demandé de segmenter un texte est assez faible ((Sproat *et al.*, 1996) rapportent 76% de moyenne entre rappel et précision sur les mots, Jin (2007) rapporte une f-mesure de 0,839) . Ceci peut s'expliquer en partie par le fait que la tâche de segmentation recouvre différents problèmes qui ne sont spécifiques ni au mandarin ni à l'écriture chinoise : la définition des unités lexicales n'est triviale pour aucune langue et Packard (2000) propose 8 définitions différentes du « mot » : il fait remarquer que les critères phonologiques, syntaxiques, sémantiques, sociologiques, et autres ne coïncident pas toujours. La question de la segmentation soulève en effet des problèmes relatifs aux expressions multi-mots, au traitement des entités nommées et aux phénomènes de figement et de collocation (des exemples sont donnés à la section 5.3). Certains désaccords sur la segmentation relèvent d'une différence d'analyse morpho-syntaxique systématique et sont explicables, motivés et le plus souvent homogénéisables (Xia, 2000). C'est le cas par exemple du traitement de la marque du pluriel sur les nom humains (們 men) analysée en tant que suffixe ([N 們] = une unité) ou en tant que postposition (N + 們 = deux unités). Il en va de même pour les marques aspectuelles et résultatives sur les verbes. Les désaccords autour de figements lexicaux sont eux bien plus difficile à trancher (exemple : 全球暖化 quánqiúnuǎnhuà terre-entière-chaleur-devenir, *réchauffement planétaire*)

Enfin, dans un contexte applicatif, Wu (2003) note que différentes applications de TAL nécessitent différents critères de segmentation en amont. Dans notre cas, notre objectif premier est la construction de ressources lexicales à des fins de linguistique expérimentale sur corpus. Les contraintes et besoins que cela implique diffèrent donc légèrement des besoins posés par la conception d'applications TAL à visée plus industrielle.

2.3 Contexte et motivations

L'objectif de notre travail est notamment l'induction de lexiques et le pré-traitement de corpus à des fins d'études linguistiques. Ces corpus sont susceptibles de manifester une importante variation liée à trois facteurs au moins :
– l'espace : au travers des différentes variantes du mandarin pratiquées à Pékin, Hong Kong, Singapour ou Taïwan ;
– le temps : la publication récente des n-grammes de GoogleBooks ouvre de nouvelles possibilités pour une étude du lexique à la fois diachronique et quantitative, mais pose le problème de la segmentation sous un angle différent ; il est en effet exclu d'utiliser un système entraîné sur un corpus de la fin du XXe siècle pour segmenter des textes bien plus anciens [2] ;
– le domaine : des corpus de natures différentes, voire des corpus de spécialité, utiliseront des lexiques différents et en partie spécifiques, significativement différents de ce que l'on peut trouver dans les corpus d'apprentissage utilisés par les systèmes supervisés.

Les méthodes classiques de segmentation et d'évaluation exploitant des lexiques pré-existants ou des corpus segmentés manuellement semblent donc peu appropriées pour nos recherches où les mots inconnus et la tolérance à la variation nous intéressent particulièrement. D'un autre côté, les analyses proposées dans des travaux de linguistiques portant sur la définition de l'unité lexicale (Nguyen, 2006; Magistry, 2008) sont difficiles à automatiser.

Cette motivation à la fois linguistique et quantitative a nourri notre intérêt pour les méthodes non supervisées et particulièrement celles qui reposent sur l'hypothèse, motivée linguistiquement, de Harris (1955). Un exemple en est notamment les expérimentations menées sur corpus par Jin & Tanaka-Ishii (2006).

3 L'hypothèse harrissienne et sa reformulation entropique

Dans son article « *From phoneme to morpheme* », Harris (1955) formule l'hypothèse de l'existence d'un lien entre les frontières de morphèmes ou de mots et le nombre de successeurs possibles à une suite de phonèmes dans la chaîne parlée. Il effectue ensuite différentes expériences visant à confirmer cette hypothèse et à préciser la procédure de segmentation.

À l'époque, ces expériences ne pouvaient pas tirer parti de grands volumes de textes ou d'enregistrements et furent réalisées sous forme d'enquêtes. Plus récemment, cette idée a été déclinée pour réaliser différentes tâches telles que la détection de collocations (Frantzi & Ananiadou, 1996) ou la segmentation du chinois (Jin & Tanaka-Ishii, 2006). Les expériences sur le chinois reposent sur la reformulation de l'hypothèse de Harris dans le cadre de la théorie de l'information proposée par Tanaka-Ishii (2005), qui repose sur la notion d'*entropie* (notée H) : la distribution des successeurs possibles d'une suite de tokens, modélisée ici par un n-gramme x_n (de phonèmes ou de sinogrammes), permet de définir et de calculer une entropie $h(x_n)$, dite *entropie de branchement*, comme suit :

$$h(x_n) = H(\chi|x_n) = -\sum_{x \in \chi} P(x|x_n).\log P(x|x_n),$$

où χ est l'ensemble de tous les sinogrammes connus, mais aussi des lettres latines (en raison des expressions ou noms étrangers) et des chiffres arabes, et $P(x|x_n)$ est la probabilité conditionnelle de trouver le caractère x à la suite du n-gramme x_n. Cette reformulation a ainsi servi à tester l'hypothèse de Harris en corpus (Tanaka-Ishii & Jin, 2006).

Le modèle de Jin & Tanaka-Ishii (2006), qui est plus proche de l'article de Harris que notre système (décrit ci-dessous), est aussi beaucoup plus complexe. Il repose sur cinq modèles de langue (reposant sur des 1 à 5-grammes de sinogrammes) utilisés conjointement. Ceci permet de calculer une entropie de branchement après une séquence de longueur variable. Cependant à chaque intervalle entre deux sinogrammes, son système applique une série de critères qui lui permettent de décider de façon binaire si il faut segmenter ou non. La condition principale correspondant à l'hypothèse de Harris est qu'une frontière d'unité linguistique correspond à un point où l'entropie branchante atteint un maximum local. L'écriture chinoise produisant de nombreuses occurrences de mot d'un seul sinogramme, Jin et Tanaka-Ishii considèrent qu'il existe une frontière à chaque point où l'entropie est croissante. Leur système est ensuite évalué de façon classique par rappel/précision/f-mesure sur un extrait du corpus segmenté manuellement (celui de l'Université de Pékin suivant Yu (1999)).

2. Les n-grammes de Google ont toutefois été extraits après segmentation des textes, mais aucune information n'est donnée sur la méthode utilisée, ce qui pose problème pour l'exploitation de ces données.

Notre travail est motivé par l'idée que ce type de résultat binaire et ce mode d'évaluation ne révèle pas tout le potentiel de l'hypothèse et des modèles sous-jacents. Le système est évalué à chaque point du corpus alors que des généralisations pertinentes peuvent être obtenues à partir de l'ensemble des mesures de variation d'entropie effectuées, même bruités. Par ailleurs, les mesures de variations d'entropie ont des valeurs continues qui semblent à même de rendre compte plus finement du problème linguistique que les modes d'évaluation standard réduisent à une tâche de classification binaire.

Dans la section suivante nous présentons un système d'analyse qui conserve l'information sur les variations d'entropie afin de pouvoir utiliser celle-ci pour induire un lexique (section 5) ou corréler la variation d'entropie à la syntaxe sans la discrétiser (section 6).

4 Architecture de notre système de segmentation

Le modèle présenté ci-dessus peut être décliné en divers systèmes ayant en commun l'utilisation d'une mesure de « surprise » pour détecter les frontières. Contrairement aux travaux de Jin et Tanaka-Ishii, notre objectif n'est pas la segmentation en elle même mais l'induction de lexiques. Il nous est donc possible de ne pas prendre une décision binaire sur la segmentation à chaque intervalle entre deux sinogrammes mais de propager la mesure de « surprise » à un système qui prend une décision sur l'intégration ou non une suite de sinogrammes donnée à notre lexique.

Afin d'obtenir des résultats plus lisibles, nous avons choisi dans un premier temps d'utiliser un système simplifié ne reposant que sur un seul modèle de langue (4-grammes) qui calcule l'entropie branchante $h(x_3)$ à chaque inter-sinogramme et propage celle-ci pour en observer la variation.

Pour la séquence 台北市政府昨日開會決議… Táiběishìzhèngfǔ zuórì kāihuì juéyì… (*La municipalité de Taipei à décidé hier en réunion…*), notre chaîne de traitement produit la sortie suivante :

台 -4,98 北 1,11 市 1,53 政 -4,51 府 4,77 昨 -4,26 日 1,55 開 -0,06 會 -0,77 決 -0,92 議

Segmenter lorsque l'entropie est croissante produit donc le découpage : 台北 *(Taipei)* 市 *(ville)* 政府 *(gouvernement)* 昨日 *(hier)* 開會決議 (au lieu de 開會/決議 *tenir une réunion / décider*).

5 Induction de lexique : méthodologie et évaluation comparative

Dans cette section nous cherchons à induire un lexique à partir des informations sur la variations d'entropie et à définir une mesure de confiance dans les unités lexicales induites.

Une chaîne $w = c_1c_2...c_n$ est une unité lexicale candidate s'il en existe au moins une occurrence w_i dont les variations d'entropie inter-sinogrammes sont notées $e_{w_i,0}, \ldots, e_{w_in}$ avec $e_{w_i,k}$ la variation d'entropie après le k-ème sinogramme de la i-ème occurrence de w qui vérifie $e_{w_i,0} > 0,\ e_{w_i,n} > 0$ (l'entropie est croissante avant et après la chaîne) et $\forall k \in [1, n-1]\,,\ e_{w_ik} \leq 0$ (l'entropie est décroissante ou constante à l'intérieur de la chaîne).

Nous avons appliqué notre segmenteur au corpus de l'Academia Sinica (Chen *et al.*, 1996), qui contient environ $7,7$ millions d'occurrences de 198 236 unités lexicales (segmentées manuellement) pour environ 12 millions de sinogrammes. Nous en avons ainsi extrait 193 714 unités lexicales candidates. Pour chaque unité lexicale candidate, nous disposons donc d'un ensemble d'occurrences auxquelles sont associées des variations d'entropie aux frontières et internes. Le corpus utilisé, qui compte 12 millions de sinogrammes, produit 193 714 unités lexicales candidates.

5.1 Métriques de confiance

Nous avons défini puis comparé différentes métriques pour filtrer le bruit parmi les unités lexicales candidates, qui combinent de façons différentes leur fréquence et une mesure de confiance, définie ci-dessous. La fréquence d'une unité lexicale candidate w est directement estimée à partir du nombre d'occurrence de celle-ci dans le corpus, noté $Nocc(w)$. Pour tirer parti de l'information sur la variation d'entropie, nous définissons pour chaque unité lexicale candidate une mesure de confiance définie à partir des variations d'entropie comme suit :

– pour chaque occurrence w_i de w, on définit une confiance locale $c_{w_i} = \min(e_{w_i,o},\ e_{w_i,n},\ -e_{w_i,1}, \ldots,\ -e_{w_i,n-1})$;

– on associe à la chaîne candidate w la confiance (globale) $c_w = \max_{i=1}^{Nocc(w)}(c_{w_i})$.

En d'autres termes, la confiance accordée à une occurrence est définie par valeur de variation d'entropie la moins fiable parmi celles ayant abouti à cette segmentation, et la confiance accordée à une unité lexicale est égale à la confiance de l'occurrence à laquelle on fait le plus confiance.

Différentes combinaisons de l'indice de confiance et du nombre d'occurrence sont alors possibles. Nous avons retenu les quatre combinaisons suivantes :

fréquence seule : $Nocc(w)$. Cette métrique simple présente l'avantage de cibler les unités lexicales couvrant le plus grand nombre d'occurrences en corpus, mais n'utilise pas l'information sur la variation d'entropie que nous conservons et se comporte ainsi comme les systèmes de segmentation binaires.

produit : $c_w \times Nocc(w)$. Cette métrique introduit la mesure de confiance, en lui conférant une importance identique à celle de la fréquence.

log : $c_w \times \log(Nocc(w))$. Cette métrique diminue l'impact des hautes fréquences.

confiance seule : c_w. Cette métrique ignore la mesure de fréquence et n'utilise que les informations extraites du modèle d'entropie.

Ces métriques sont choisies arbitrairement mais de manière à donner une importance croissante à la mesure de confiance afin d'évaluer sa pertinence.

5.2 Filtrage et évaluation du lexique de façon semi-supervisée.

Chacune des métriques présentées à la section précédente permet de définir un ordre de confiance sur les entrées lexicales, et donc de trier le lexique induit. On peut alors choisir par exemple de ne conserver que les n premières entrées. Pour un même n, chaque métrique conduit donc à un lexique filtré distinct. Il reste à choisir la meilleure métrique, puis un seuil sur cette métrique en dessous duquel filtrer le lexique, afin d'obtenir le meilleur compromis entre couverture et bruit.

Ces choix sont délicats à effectuer *a priori*, de même que l'évaluation de la qualité du lexique obtenu. Nous avons donc commencé par utiliser notre système dans une configuration semi-supervisée afin de pouvoir le comparer à une référence établie manuellement, et comprendre notamment quelle métrique semble se comporte le mieux.

Afin d'avoir une idée de la qualité d'un lexique L_i induit (filtré ou non), nous le comparons au lexique L_m extrait à partir de la segmentation effectuée manuellement à l'Academia Sinica. Rappelons qu'il ne s'agit pas d'un standard mais d'une analyse possible des données, motivée linguistiquement, mais qui n'est pas la seule. Pour comparer les deux lexiques nous avons utilisé l'*indice de Jaccard* et la *f-mesure* (qui donnent des résultats similaires). En l'absence de « bonne » ou de « mauvaise » réponse, ces indicateurs donnent tout de même une idée de la cohérence entre le résultat de notre système non-supervisé et une analyse effectuée manuellement. Ces deux mesures sont définies classiquement comme suit.

$$f(L_i, L_r) = \frac{2pr}{p+r}, \text{ avec } p(L_i, L_m) = \frac{|L_i \cap L_m|}{|L_i|} \text{ et } r(L_i, L_r) = \frac{|L_i \cap L_m|}{|L_m|} \; ; \quad jaccard(L_i, L_r) = \frac{|L_i \cap L_m|}{|L_i \cup L_m|}$$

Le lexique L_m obtenu à partir de la segmentation manuelle est trié par nombre d'occurrences (on considère que l'on a une confiance égale en toutes les entrées de ce lexique et l'on cherche à détecter en priorité les formes les plus fréquentes). Les différentes métriques décrites ci-dessus, qui combinent de diverses façons le nombre d'occurrences $Nocc$ et l'indice de confiance c permettent de moduler l'importance respective de ces deux quantités, depuis l'utilisation du nombre d'occurrence seul jusqu'à l'utilisation du seul indice de confiance.

Pour chacune des quatre métriques retenues, nous avons calculé le Jaccard et la f-mesure entre lexique induit filtré à différentes valeurs de la métrique et lexique manuel filtré à différents niveaux de fréquence. La figure 1 montre les résultats obtenus avec la f-mesure, en en indiquant les lignes de niveaux. Ceci nous permet de choisir un seuil en fonction d'un taux d'accord avec la référence. On observe que la qualité du tri par la fréquence se dégrade plus rapidement que les autres. À l'inverse, l'utilisation du seul indice de confiance ne semble pas accorder suffisamment d'importance aux formes fréquentes (le sommet est plus éloigné de l'origine du graphique).

Afin de nous faire une idée plus précise du contenu des lexiques induits par rapport à celui sous-jacent au corpus de l'Academia Sinica, nous avons choisi de les filtrer de la façon suivante : nous avons choisi le seuil de façon à maximiser la taille du lexique filtré tout en préservant une f-mesure d'au moins 0.6 par comparaison avec le lexique

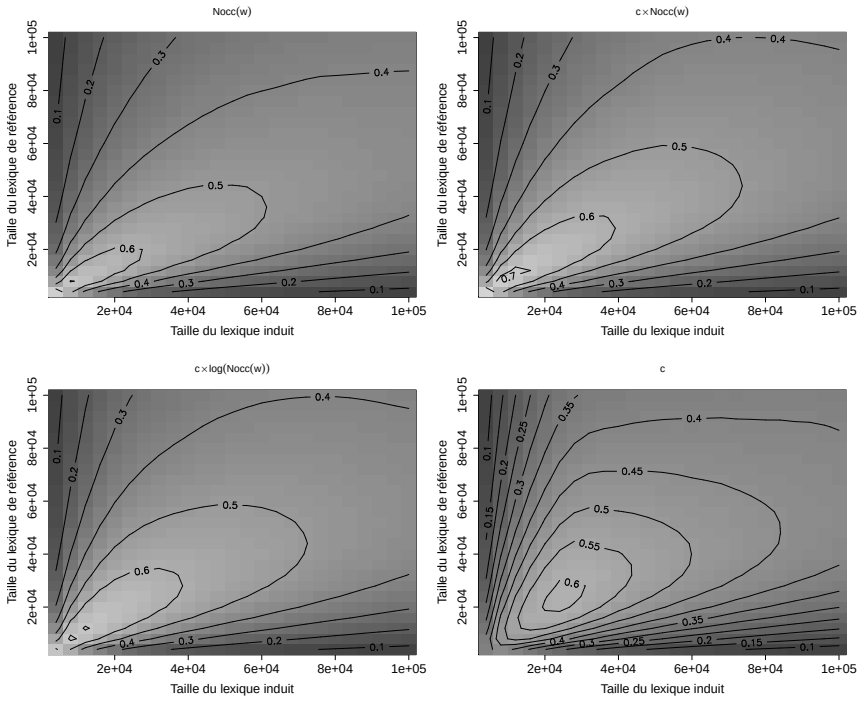


FIGURE 1 – comparaison des lexiques induits avec différentes mesures de confiance avec le lexique obtenu après segmentation manuelle (f-mesure)

mesure de confiance	taille $\lvert L\rvert$	formes validées $\lvert L \cap L_m\rvert$	occurrences couvertes	couverture	nombres de caractères
lexique manuel L_m	*116 844*	*116 844*	*7 584 040*	*100 %*	*5 861*
$Nocc(w)$	27 500	16 929	6 719 083	89 %	3 421
$c \times Nocc(w)$	38 000	24 045	6 901 992	91 %	4 016
$c \times \log(Nocc(w))$	38 000	24 571	6 892 594	91 %	4 070
c	31 000	23 620	6 695 416	88 %	4 097

TABLE 1 – Comparaison des lexiques induits L_i. Pour chaque mesure de confiance, le lexique est de taille maximale parmi ceux ayant une f-mesure de 0.6 par rapport au lexique L_m extractible du corpus de l'Academia Sinica.

manuel L_m (sur les graphiques de la figure 1, le seuil correspond donc à l'abscisse du point le plus à droite de la ligne à f-mesure de 0.6). Pour chacun des quatre lexiques ainsi extraits, nous avons effectué les mesures suivantes :
– nombre d'unités lexicales ($|L_i|$) ;
– nombre d'unités lexicales présentes dans la référence manuelle ($|L_i \cap L_m|$)
– nombre d'occurrences dans le corpus des unités lexicales communes (celles de $L_i \cap L_m$) ;
– couverture de $L_i \cap L_m$, c'est-à-dire proportion du corpus couverte par les unités lexicales communes ;
– nombre de sinogrammes distincts utilisés dans le lexique.
Les résultats sont données dans le tableau 1, où nous donnons également les valeurs correspondantes pour le lexique manuel L_m. On constate que l'utilisation de notre indice de confiance améliore bien la proportion de formes valides tandis que la fréquence reste une valeur intéressante pour optimiser la couverture du corpus.

Remarquons que les formes valides capturées par nos lexiques diffèrent sensiblement. Le tableau 2 donne les indices de Jaccard entre nos lexiques calculées 2 à 2 et confirme l'intérêt de combiner les deux informations de confiance et de fréquence.

	$Nocc(w)$		
$c \times Nocc(w)$	0,64	$c \times Nocc(w)$	
$c \times \log(Nocc(w))$	0,59	0,92	$c \times \log(Nocc(w))$
c	0,46	0,71	0,76

TABLE 2 – Indices de Jaccard entre les quatre lexiques induits filtrés, en se restreignant aux unités lexicales également présentes dans le lexique manuel.

type d'erreur	quantité	type d'erreur	quantité
suffixation	37	écriture non-chinoise	14
dates et nombres	35	conjonction	14
verbes	27	adverbes	10
expressions figée	21	entité nommée	4
translittération	2	autre	36

TABLE 3 – Répartition des faux négatifs

5.3 Analyse d'erreur

Dans cette section, nous présentons les résultats d'un analyse d'erreur, ou de divergence, entre le lexique de référence et le lexique construit à la section précédente au moyen de la mesure $c_w \times \log(Nocc(w))$. Nous avons concentré notre analyse sur deux axes : les unités lexicales de haute fréquence absentes de notre lexique induit mais présentes dans le lexique de référence (faux négatifs) et les chaînes considérées comme des unités lexicales avec un haut niveau de confiance mais absentes du lexique de référence (faux positifs). Pour chaque groupe, nous avons considérés les 200 premiers cas.

Le système mis en œuvre pour cet article est volontairement simpliste pour établir un système de base auquel se comparer. Les erreurs observées dans cette section suggèrent différentes pistes d'amélioration en amont (modification sur le modèle de langue utilisé) et en aval (ajout de règles linguistiques basées sur les catégories fermées).

5.3.1 Analyse des faux négatifs

Nous avons classé les faux négatifs suivant leur morphologie lorsque leur construction interne était transparente, dans le cas contraire nous avons les avons classé en parties du discours, mais le mandarin étant très ambigu sur ce point (le phénomène de conversion est fréquent), de nombreux cas sont restés non classés. Les résultats de cette analyse sont donnés dans le tableau 3 dont nous détaillons la moitié gauche ci-dessous.

Le type d'erreur le plus représenté concerne des unités lexicales, essentiellement nominales, construite sur le modèle *base+suffixe*. C'est là un phénomène de morphologie constructionnelle très productif en mandarin moderne dont on peut donner l'exemple 法務部長 fǎwùbùzhǎng *ministre de la justice* construit sur la base 法務部 fǎwùbù *ministère de la justice* à laquelle on ajoute le suffixe 長 zhǎng pour tête/chef (部 bù étant lui même un suffixe indiquant un ministère) . Dans (Magistry, 2008), des méthodes quantitatives ont permis d'estimer la productivité de ce procédé et ont ainsi montré que les règles très productives correspondent aux cas où le statut morphologique ou syntaxique de la composition est le plus discutable. C'est aussi un des points sur lesquels les guides de segmentation manuelle peuvent diverger. La grande proportion de ce type d'erreur n'est donc pas étonnante mais un soin particulier devra être apporté au traitement de ce phénomène dans des travaux futurs.

Les dates et nombres sont aussi une erreur attendue, la distribution des tokens qui les composent étant particulière.

Le cas des verbes est moins clair. cependant la présence de marques d'aspect (formant une petite classe fermée) directement à la suite du verbe peuvent induire notre système en erreur. Il coupera après le marqueur d'aspect. Il en va de même (mais dans une moindre mesure) des constructions *verbe+résultatif* (ex : 吃完 chīwán manger-finir, *avoir fini de manger*) qui bien que sécables (ex : 吃不完 chībuwán manger-négation-finir, *ne pas pouvoir finir de manger*) n'ont pas toujours un sens compositionnel. Certaines combinaisons sont lexicalisées et peuvent donner lieu à débat concernant leur bonne segmentation.

Parmi les 200 premiers nous avons compté 21 cas qui nous semblent être des figements à différents degrés, comme 高爾夫-球場 gāo'ěrfū-qiúchǎng golf-terrain, *terrain de golf*, l'unité pour *golf* étant autonome et celle pour *terrain* pouvant concerner tout type de terrain sport utilisant une balle. Mais ceci inclut aussi des « expressions en quatre

Type d'erreur	Qnt	Type d'erreur	Qnt
nom	17	verbe+DE	8
adverbe+verbe	15	adverbe+copule	8
suffixe+DE	14	adverbe+adverbe	8
nombre+classificateur	14	adverbe+avoir	7
verbe+aspect	13	adverbe+auxiliaire	7
démonstratif+classificateur	9	pronom+DE	6

TABLE 4 – Répartition des faux positifs

caractères » dont la concision et la structure interne relèvent d'un état antérieur de la langue (ex : 前所未有 avant/ce que/pas encore/avoir, *sans précédent*). Ce type d'erreurs regroupe ainsi des expressions figées dont certaines sont idiomatiques et d'autres compositionnelles (en quantités équivalentes).

5.3.2 Analyse des faux positifs

Nous avons ensuite analysé les faux positifs en observant leur composition interne. Une première remarque est que sur les 200 premiers, 198 sont des bigrammes et 2 sont des unigrammes (cette préférence disparaît à mesure que la confiance diminue). Nous avons donc classifié les erreurs en fonction des deux sinogrammes qui les composent. La dispersion est plus grande, nous ne donnons donc dans le tableau 4 que les types les plus importants.

Les 17 noms sont des unités lexicales dont le statut est discutable. 13 comportent en seconde position un élément appartenant à une classe très fermée de « mots » indiquant un lieu ou une direction (上,下,内,裡,中 shàng, xià, nèi, lǐ, zhōng *sur, sous, dans, à, au milieu*) et en première position un nom monosyllabique.

Les *adverbe+verbe* sont des combinaisons d'adverbe monosyllabique (*très, le plus, trop, aussi, relativement*) et de verbes d'état (*rapide, grand, petit, suffisant, nombreux, difficile).*

Suffixe+DE désigne un groupe composé en première position d'un suffixe nominal très productif (voir plus haut) et en seconde de 的 DE qui marque une relative ou la possession. Il s'agit ici d'une double erreur de segmentation dûe au fait que de nombreuses bases peuvent commuter devant ces suffixes et qu'il viennent terminer une large classe de nom et peuvent donc fréquemment se trouver devant le DE.

Les séquences *nombre+classificateur* et *démonstratif+classificateur* sont liées à la structure des groupes nominaux (non nus) en mandarin dans lesquels un classificateur est requis et doit obligatoirement être précédé d'un élément dénotant une quantité ou d'un démonstratif. Il est donc peu étonnant que notre système tende à les regrouper.

Remarquons aussi les 6 séquences pronom+DE qui se traduiraient par des possessifs *mon/le mien, ton/le tien, son/le sien…* auxquels s'ajoutent deux variantes avec un possesseur non-humain. On ne compte que 6 erreurs de ce type, mais c'est là une liste exhaustive des pronoms singuliers + DE.

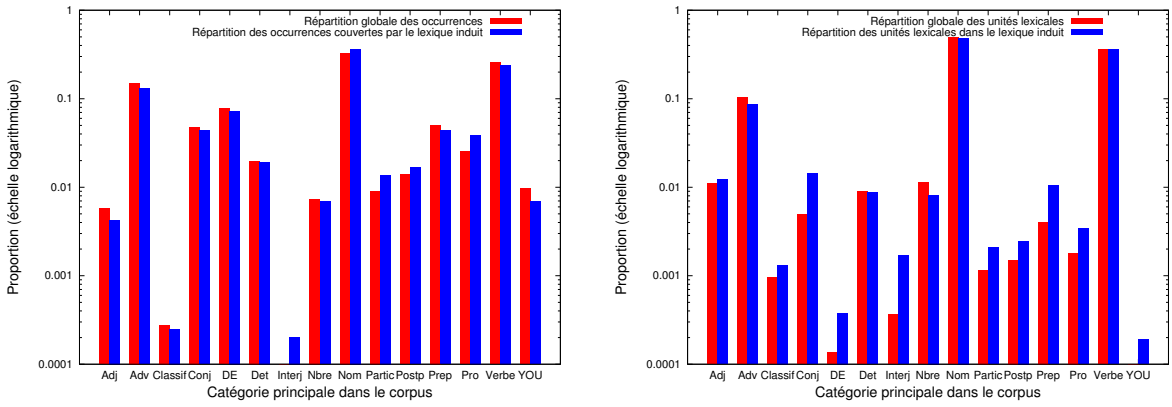
Une analyse de ces erreurs réalisée sur les caractères et non sur couples de caractères erronés montre que toutes ces erreurs incluent une unité lexicale qui est un unigramme très fréquent ou appartenant à une classe très fermée.

5.4 Caractéristiques générales du lexique

Chen *et al.* (1993) fournissent des informations sur la distribution globale d'un lexique du chinois mandarin et des occurrences en corpus. On peut ainsi vérifier si notre lexique induit possède bien les même propriétés que le lexique obtenu manuellement.

Tout d'abord, en observant le nombre d'occurrence des unités lexicale ordonnées par fréquence décroissante, on obtient une courbe zipfienne qui se superpose bien avec celle obtenue en utilisant le lexique manuel.

Les observations plus spécifiques au mandarin concernent la répartition des unités lexicales et de leurs occurrences en fonction de leur longueur en nombre de sinogrammes. Sur les 38 000 unités lexicales « de confiance » (ou les plus fréquentes pour le lexique manuel), le lexique induit est bien constitué principalement de bigrammes et de trigrammes (respectivement 65,8% et 27,3%) tandis que les unigrames constituent 6,5% du lexique. Le lexique manuel compte lui 7% d'unigrammes, 67,7% de bigrammes et 19% de trigrammes. Concernant le nombre d'occurrence observées dans le corpus, on obtient 37% d'unigramme, 54% de bigrammes et 7% de trigrammes (pour le corpus segmenté manuellement, on obtient respectivement 45%, 49% et 4%)



(a) (b)

FIGURE 2 – Comparaison des répartitions en catégories entre les occurrences de formes segmentées à l'identique par notre système et par le corpus lui-même (a), et entre les unités lexicales correspondantes (b).

6 Évaluation de la segmentation par comparaison avec les informations morphosyntaxiques et syntaxiques

Après avoir évalué notre modèle de segmentation au travers du lexique qu'elle permet d'extraire du corpus de l'Academia Sinica, nous avons poursuivi nos expériences d'évaluation en cherchant à tirer parti des annotations morphosyntaxiques et syntaxiques que fournit le Sinica Treebank (Chen *et al.*, 1996), qui en couvre une partie.

6.1 Répartition en catégories

L'étude de la distribution du lexique induit L_i en fonction de sa fréquence, bien qu'importante, ne suffit pas à montrer que ses propriétés sont cohérentes avec celles du lexique motivé linguistiquement L_m qui est sous-jacent au corpus de l'Academia Sinica. Nous avons donc cherché à comparer la répartition en catégories de ces deux lexiques. Nous avons donc identifié dans le treebank de l'Academia Sinica les occurrences de formes communes entre la segmentation du corpus et celle de notre système. Nous avons alors comparé la répartition de ces occurrences en parties du discours par rapport à celle de l'ensemble des occurrences de formes dans le corpus de l'Academia Sinica (figure 2a). Nous avons effectué la même chose avec unités lexicales correspondantes (figure 2b). Les catégories que nous avons utilisées sont les suivantes : Adj (adjectif), Adv (adverbe), Classif (classifieur), Conj (conjonction), DE (particules 的,之,得 et 地), Det (déterminant), Interj (interjection), Nbre, Nom, Partic (particule), Postp (postposition), Prep (préposition), Pro (pronom), Verbe et YOU (verbe 有 , *avoir*).

Nous discuterons de façon plus détaillée ces figures à la section suivante, mais on constate globalement une bonne corrélation entre les deux répartitions, tant pour les catégories les plus fréquentes que pour celles qui le sont moins.

6.2 Corrélation entre variation d'entropie et structure syntaxique

Notre modèle de segmentation reposant sur les variations d'entropie, il ne produit pas simplement pour chaque paire de sinogrammes adjacents une décision binaire (segmenter ou non), mais bien une mesure quantitative de la séparabilité des deux sinogrammes concernés. Nous avons cherché à confronter ce *degré de séparation linéaire* S_l avec les informations syntaxiques (arbres en constituants) fournies par le Sinica Treebank. L'intuition sous-jacente est que l'on pourrait constater une corrélation entre la séparabilité linéaire produite par notre modèle de segmentation et un *degré de séparation syntaxique*, mesure qui serait d'autant plus élevée que les deux sinogrammes étudiés appartiennent à des unités lexicales éloignées l'une de l'autre au sein de la structure en constituants.

Pour effectuer cette expérience, nous avons défini le degré de séparation syntaxique S_s entre deux unités lexicales adjacentes comme étant la longueur (en arcs) du plus court chemin permettant de les relier entre elles dans l'arbre de constituance. Il résulte de cette définition que S_s est nécessairement au moins égal à 2, ce qui est le cas lorsque

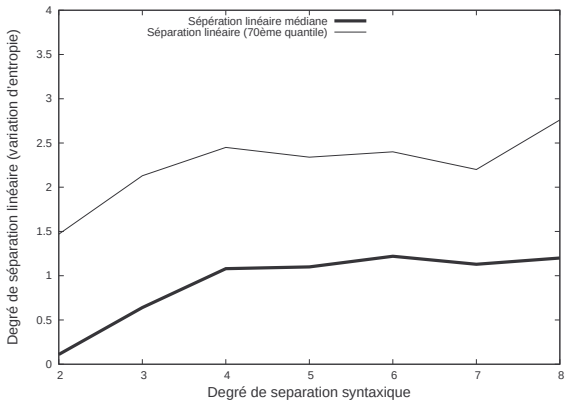


FIGURE 3 – Évolution de la séparation linéaire S_l (variation d'entropie) en fonction de la séparation syntaxique S_s sur les frontières communes au corpus de l'Academia Sinica et à la segmentation induite par notre système.

les unités lexicales ont un même nœud père.

La figure 3 montre pour chaque valeur de la séparation syntaxique S_s quelle est la valeur médiane de la séparation linéaire S_l, ainsi que son soixante-dixième quantile. On constate deux choses. Tout d'abord, lorsque le degré de séparation syntaxique est de 4 ou moins, il y a une nette corrélation entre S_l et S_s. Autrement dit, le modèle de segmentation utilisé réussit à capturer une partie des informations syntaxiques locales, de niveau terme voire *chunk*. En revanche, au-delà d'une séparation syntaxique de 4, la séparation linéaire médiane n'évolue quasiment plus. Deux hypothèses, non-exclusives l'une de l'autre, viennent à l'esprit. Tout d'abord, le modèle utilisé est 4-gramme, et il est difficile de capturer des frontières entre longs constituants avec un modèle local de ce type. Par ailleurs, le modèle simple que nous utilisons, inspiré de Harris, est un modèle très surfacique qui n'a aucune raison de pouvoir capturer des informations sur la macro-structure de l'arbre syntaxique d'une phrase.

Ce résultat, bien qu'obtenu à l'échelle de tout le corpus au moyen d'un calcul de médiane, est néanmoins prometteur : il ne semble pas exclu de pouvoir utiliser notre modèle de segmentation non-supervisé, tel quel ou sous une forme raffinée, non seulement pour induire une segmentation en unités lexicales et un lexique associé mais également pour identifier des collocations, termes, locutions et autres unités lexicales complexes, et de tenter de leur associer une structure interne. On peut ainsi espérer avoir accès à un moyen objectif, qui n'utilise pas de connaissance *a priori* et qui est donc indépendant de la langue, pour mettre en évidence le continuum qui relie les unités lexicales les plus classiques aux expressions semi-compositionnelles ou collocationnelles.

7 Conclusion et perspectives

Dans cet article, nous montrons sur le chinois mandarin qu'un modèle simple utilisant une hypothèse linguistiquement motivée mais indépendante de toute connaissance *a priori* sur une langue particulière donne des résultats prometteurs pour la segmentation non supervisée de textes et l'induction d'unités lexicales cohérentes avec des annotations de niveau syntaxique. De plus, certains résultats pouvant apparaître comme des erreurs de segmentation sont susceptible de questionner de façon constructive des analyses linguistiques traditionnelles parfois influencées par les états antérieurs de la langue.

Certaines erreurs résultent toutefois des limites de notre système dans son état actuel. En particulier, un traitement plus fin des catégories fermées (démonstratifs, DE,...) pourrait nettement améliorer les résultats tout en demandant une quantité d'analyse bornée par la taille de ces catégories. Mais d'autres améliorations destinées à rendre le modèle plus proche de considérations linguistiques pourront également être testées. Le modèle lui-même peut également faire l'objet de raffinements, par exemple pour comprendre si la prise en compte du mot situé à *droite* de l'intersinogramme considéré est de nature a améliorer les résultats.

Nous prévoyons de tester notre système sur des corpus relevant de différentes variétés du chinois mandarin, pour en étudier notamment les variations des distributions lexicales, mais également de le tester sur d'autres langues non-segmentées, pour valider l'approche sur un échantillon plus large de langues. Nous souhaitons également me-

ner des expérimentations sur diverses langues, y compris le français, pour segmenter non seulement des flux de sinogrammes mais ainsi, par exemple, des flux de phonèmes (en vue d'une segmentation en morphèmes) ou de *tokens* (en vue de l'identification d'unités lexicales multi-mots et de termes).

Références

CHEN C. Y., TSENG S. F., HUANG C. R. & CHEN K. J. (1993). Some distributional properties of Mandarin Chinese — A study based on the academia sinica corpus. In *Proceedings of Pacific Asia Conference on Formal and Computational Linguistics I*, p. 81–95.

CHEN K. J., HUANG C. R., CHANG L. P. & HSU H. L. (1996). Sinica corpus : Design methodology for balanced corpora. In *Proceedings of PACLIC 11th Conference*, p. 167--176.

FRANTZI K. T. & ANANIADOU S. (1996). Extracting nested collocations. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, p. 41–46.

HARRIS Z. S. (1955). From phoneme to morpheme. *Language*, **31**(2), 190–222.

HUA Y. (2000). Unsupervised word induction using MDL criterion. In *Proceedings of ISCSL*.

HUANG C. R., CHEN K. J. & CHANG L. L. (1996). Segmentation standard for chinese natural language processing. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, p. 1045–1048.

JIN Z. (2007). *A Study on Unsupervised Segmentation of Text Using Contextual Complexity*. PhD thesis, University of Tokyo, Graduate School of Information Science and Technology, Tokyo, Japon.

JIN Z. & TANAKA-ISHII K. (2006). Unsupervised segmentation of chinese text by use of branching entropy. In *Proceedings of the COLING/ACL on Main conference poster sessions*, p. 428–435.

MAGISTRY P. (2008). Productivité morphologique : Étude sur le chinois mandarin. Master's thesis, Université Paris Diderot, UFR de Linguistique, Paris, France.

NGUYEN . (2006). *Unité lexicale et morphologie en chinois mandarin*. PhD thesis, Université de Montréal, Montréal.

PACKARD J. L. (2000). *The morphology of Chinese : A linguistic and cognitive approach*. Cambridge Univ Pr.

PENG F. & SCHUURMANS D. (2001). Self-supervised chinese word segmentation. *Advances in Intelligent Data Analysis*, p. 238–247.

SENG S., BIGI B., BESACIER L. & CASTELLI E. (2009). Segmentation multiple d'un flux de données textuelles pour la modélisation statistique du langage. In *Actes de la conférence TALN 2009*, Senlis, France.

SPROAT R., GALE W., SHIH C. & CHANG N. (1996). A stochastic finite-state word-segmentation algorithm for chinese. *Computational linguistics*, **22**(3), 377–404.

TANAKA-ISHII K. (2005). Entropy as an indicator of context boundaries : An experiment using a web search engine. *Natural Language Processing–IJCNLP 2005*, p. 93–105.

TANAKA-ISHII K. & JIN Z. (2006). From phoneme to morpheme : Another verification using a corpus. *Computer Processing of Oriental Languages. Beyond the Orient : The Research Challenges Ahead*, p. 234–244.

WU A. (2003). Customizable segmentation of morphologically derived words in chinese. *International Journal of Computational Linguistics and Chinese Language Processing*, **8**(1), 1–27.

WU L.-C. (2010). Outils de segmentation du chinois et textométrie. In *Actes de la conférence TALN 2010*, Montréal, Canada.

WU Y. C., YANG J. C. & LEE Y. S. (2010). Chinese word segmentation with conditional support vector inspired markov models. In *Proceedings of the Joint Conference on Chinese Language Processing*.

XIA F. (2000). The segmentation guidelines for the penn chinese treebank (3.0). *IRCS Technical Reports Series*.

XUE N. (2003). Chinese word segmentation as character tagging. *International Jourrnal of Computational Linguistics and Chinese Language Processing*.

YU S. (1999). Guidelines for the annotation of contemporary chinese texts : word segmentation and POS-tagging. *Institute of Computational Linguistics, Beijing University, Beijing*.

H. ZHAO & Q. LIU, Eds. (2010). *The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff*.

ZHIKOV V., TAKAMURA H. & OKUMURA M. (2010). An efficient algorithm for unsupervised word segmentation with branching entropy and MDL. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, p. 832–842.