RECITAL 2009, Senlis, 24–26 juin 2009

Extraction de lexique dans un corpus spécialisé en chinois contemporain

Gaël Patin[1, 2]
(1) Er-Tim, Inalco, 75343 Paris
(2) Arisem, Thales, 91300 Massy
gael.patin@inalco.fr

Résumé. La constitution de ressources lexicales est une tâche cruciale pour l'amélioration des performances des systèmes de recherche d'information. Cet article présente une méthode d'extraction d'unités lexicales en chinois contemporain dans un corpus spécialisé non-annoté et non-segmenté. Cette méthode se base sur une construction incrémentale de l'unité lexicale orientée par une mesure d'association. Elle se distingue des travaux précédents par une approche linguistique non-supervisée assistée par les statistiques. Les résultats de l'extraction, évalués sur un échantillon aléatoire du corpus de travail, sont honorables avec des scores de précision et de rappel respectivement de 52,6 % et 53,7 %.

Abstract. Building lexical resources is a vital task in improving the efficiency of information retrieval systems. This article introduces a Chinese lexical unit extraction method for untagged specialized corpora. This method is based on an incremental process driven by an association score. This work features an unsupervised statistically aided linguistic approach. The extraction results — evaluated on a random sample of the working corpus — show decent precision and recall which amount respectively to 52.6% and 53.7%.

Mots-clés : corpus spécialisé, unité lexicale, lexie, extraction de lexique, chinois.

Keywords: specialized corpus, lexical unit, lexicon extraction, Chinese.

1 Introduction

Les *lexiques* sont des ressources indispensables aux systèmes de recherche d'information. Ils permettent d'améliorer notablement les résultats des procédés automatiques d'analyse linguistique — étiquetage morpho-syntaxique, interprétation sémantique ou indexation — dans des domaines particuliers. Or la constitution de lexiques est confrontée à deux types de difficultés : les unes d'ordre pragmatique, telles que le coût de leur élaboration ou leur réutilisabilité, sont d'une grande importance pour la mise en œuvre industrielle ; les autres d'ordre théorique, comme la définition de l'unité lexicale dans différentes langues ou la caractérisation des particularités lexicales d'un corpus spécialisé, sont primordiales pour la pertinence et la validité des résultats. Cette confrontation entre intérêt économique et qualitatif est une problématique récurrente dans le milieu de l'entreprise. La recherche scientifique appliquée doit être à même de proposer des solutions pour répondre à cette double exigence. Cette étude propose un élément de réponse au problème de l'identification de lexique dans un corpus spécialisé en chinois contemporain via un système de classement de lexies (unités lexicales) candidates. Cette étude

s'intéresse en particulier au cas du chinois contemporain, langue pour laquelle nous ne disposons que de peu de ressources lexicales.

Dans la suite de cet article nous présentons en premier lieu les travaux similaires à notre méthode (section 2), puis nous introduisons une analyse linguistique de la notion de lexie en chinois contemporain écrit (section 3). Après une description de notre méthodologie (section 4), nous présentons nos résultats d'expérience (section 5). Enfin, nous concluons par une discussion et une comparaison des résultats (section 6).

2 Travaux similaires

L'extraction de lexique est un problème de première importance pour le traitement automatique du chinois contemporain. En effet, le processus de segmentation en mots est généralement considéré comme étant l'étape initiale pour le traitement de textes en chinois. Or le vocabulaire inconnu est un facteur d'erreur significatif pour la segmentation en mots en chinois (Sproat & Emerson, 2003). De nombreuses approches s'appuient sur les résultats d'un processus de segmentation en mots (Wu & Jiang, 2000; Li *et al.*, 2004) ou d'annotations morpho-syntaxiques (Piao *et al.*, 2006) entraîné statistiquement pour extraire du lexique inconnu. Ces techniques obtiennent de bons résultats mais l'application de ces procédés sur des corpus de domaine et genre différents de celui du corpus d'entraînement est peu efficace. La constitution de corpus annotés étant très coûteuse, nous renonçons à ce type de méthode.

D'autres approches s'intègrent dans un système de segmentation en mots et visent à améliorer son efficacité. Dans leur article, Fung et Wu (1994) introduisent une méthode de sélection de segments via une mesure d'information mutuelle mais leur méthode, testée sur un corpus spécialisé, souffre d'un faible score de rappel. Chang et Su (1997) présentent un processus itératif non-supervisé d'extraction de lexique orienté par la qualité de la segmentation obtenue avec le lexique ainsi découvert. Leur approche bien qu'efficace impose une restriction arbitraire de la longueur des lexies candidates (quadrigrammes). Feng *et al.* (2004) introduisent une méthode simple basée sur une mesure de variation du contexte avec des résultats très convaincants. Certains travaux utilisent des techniques d'extraction originales, comme la recherche de répétitions caractéristiques dans la structure interne des lexies spécialisées (Nakagawa *et al.*, 2004) ou le calcul des segments répétés (Yang & Li, 2002), mais leurs évaluations ne font pas mention du score de rappel de leur méthode. Enfin Sun *et al.* (1998) proposent un système de segmentation en mots sans ressources lexicales basé sur le score d'association mutuelle. Notre approche s'inspire de cette méthode en proposant une construction incrémentale des lexies guidée par une mesure d'association sur un corpus spécialisé non-annoté.

Il est cependant difficile de comparer rigoureusement nos travaux à d'autres et ce pour plusieurs raisons : la majeure partie des expérimentations utilise des corpus constitués de dépêches d'agence de presse. Seuls (Fung & Wu, 1994) utilisent un corpus spécialisé (comptes rendus de conseil législatif). Par ailleurs, les critères de sélection linguistique des lexies valides ne sont pas décrits précisément. D'autres travaux restreignent arbitrairement la longueur des lexies recherchées, comme celle de Chang et Su (1997) et de Feng *et al.* (2004), ou fixent un seuil de fréquence pour la sélection des lexies (Fung & Wu, 1994). Enfin certains articles ne mentionnent pas de score de rappel (Yang & Li, 2002; Nakagawa *et al.*, 2004). Seules deux études ont des conditions d'expérimentation proches des nôtres : la méthode de (Fung & Wu, 1994) et la méthode de (Feng *et al.*, 2004).

3 Définition de la lexie en chinois contemporain

Notre étude met l'accent sur l'importance de bien comprendre et analyser les phénomènes linguistiques recherchés pour pouvoir appliquer de manière efficace les outils technologiques à notre disposition. Les théories lexicologiques sont diverses et variées, nous n'avons pas la prétention d'affirmer que notre orientation est la plus juste, nous souhaitons seulement nous doter d'un cadre théorique bien fondé afin de ne pas nous contenter de simples intuitions sur la langue pour produire un traitement automatique. Dans cet article, nous nous appuyons sur les théories de Mel'cuk (1993) et Polguère (2003) appliquées par Nguyen (2006) pour le cas du chinois contemporain, même si quelques approximations ou altérations dues aux particularités de la langue chinoise peuvent apparaître dans notre présentation. Notons que les définitions introduites ne sont valides que dans le cadre de l'étude synchronique du chinois contemporain écrit.

La *lexie* (ou *unité lexicale*) est l'unité de description du lexique. Une lexie peut être issue d'un processus de formation morphologique, on parle alors de *lexème*, ou syntaxique, on parle alors de *locution*. En chinois, certaines particularités morpho-syntaxiques et typologiques rendent difficile la détection automatique de lexies dans les textes. En effet, la morphologie compositionnelle riche, souvent similaire à la syntaxe, augmente combinatoirement le nombre de candidats potentiels ; alors que le nombre réduit d'indices d'identification des lexies — dû à la morphologie flexionnelle pauvre, la rareté des morphèmes grammaticaux et l'absence de mots outils dans la construction des lexies — limite la sélection des candidats. Enfin, à l'écrit, les textes sont dépourvus d'espaces et la plupart des graphies ont individuellement un signifié lexical (même si elles ne sont pas nécessairement autonomes), ce qui restreint également les indices d'autonomie des signes. La *lexie spécialisée* est une lexie qui acquiert un sens particulier à travers une pratique langagière au sein d'un groupe social dans un domaine spécifique. Il peut s'agir de formes communes dont le sens habituel est dérivé ou modifié, ou bien de lexies exclusivement utilisées par les membres d'un groupe. Cette définition rejoint les postulats énoncés par L'homme et Polguère (2008) dans leur description du terme .

Une *graphie* est l'unité orthographique autonome minimale en chinois écrit. La graphie correspond approximativement à la notion de « caractère chinois ». Les glyphes suivants sont des exemples de graphies : 知, 保, 险, 葡, 萄, 买. Un *morphe* (noté |M|) est un signe linguistique au signifié élémentaire dont le signifiant est représentable par une suite d'une ou plusieurs graphies. L'élémentarité du signifié exprime que le morphe ne peut être le résultat d'une combinaison d'autres morphes[1]. Les suites de graphies suivantes sont des morphes car associées à un signifié et non décomposables en morphes : |寿| *'vie'*, |葡萄| *'raisin'*, |阿司匹林| *'aspirine'*, |保| *'protéger'*, |险| *'danger'*, |买| *'acheter'*. En revanche, la graphie 萄 — n'ayant aucun signifié associé — n'est pas un morphe.

Un *mot-forme* (noté (M)) est une séquence autonome et insécable de morphes. L'autonomie implique que le mot-forme peut être énoncé isolément et prendre place dans un paradigme syntaxique. L'insécabilité exprime que la séparation des morphes entraîne la perte de leur relation lexicale. Le *lexème* (noté ⟪ L ⟫) est un ensemble de mots-formes fléchis au même signifié lexical. Un *syntagme* (noté [S]) est une combinaison syntaxique de mots-formes dont les composantes ont un certain degré de liberté. Une *locution* (notée ⟦ L ⟧) est un ensemble de syntagmes lexicalisés fléchis au même signifié lexical.

[1] La définition donnée par Polguère parle plutôt de non représentabilité en termes de signes, or ceci n'est valide qu'en chinois oral. En effet, à l'écrit, l'analyse grammatologique des sinogrammes explique comment le sens du sinogramme ⟨明↔'clair'⟩ est représenté en termes de composantes graphiques ⟨日↔'soleil'⟩ et ⟨月↔'lune'⟩.

Le terme *lexie* regroupe les deux concepts de lexème et locution, c'est l'unité d'étude principale de la lexicologie. Les exemples du tableau (1) ci-dessous présentent des lexèmes et locutions avec leurs mots-formes et syntagmes associés. Cette étude propose une méthode pour identifier mots-formes et syntagmes lexicalisés et, par extension, identifier lexèmes et locutions. Notre méthode étant appliqué sur un corpus spécialisé, nous nous attendons à extraire un ensemble de lexies[2] dont une part significative devraient être des lexies spécialisées.

Lexème	***mots-formes***	***Locution***	***syntagmes***
aspirine 《 阿司匹林 》	aspirine (|阿司匹林|)	tirer (à feu) 〚 开枪 〛	tirer (à feu) [(开)(枪)]
prendre 《 拿 》	prendre */accompli/* prendre */duratif/* prendre */exprérience/* (|拿|了|) (|拿|着|) (|拿|过|)	être jaloux 〚 吃醋 〛	être jaloux [(吃)(醋)]
assurance 《 保险 》	assurance (|保|险|)	compagnie d'assurance 〚 保险公司 〛	compagnie d'assurance [(保险)(公司)]

TAB. 1 – *Exemples de lexies.*

4 Méthodologie

Notre méthode s'appuie sur la description de la lexie en chinois contemporain pour extraire le lexique inconnu d'un corpus spécialisé. Nous ne traitons pas le cas des lexies mono-morphèmiques, ni des lexies déjà connues avec un statut particulier dans le corpus qui feront l'objet d'une étude ultérieure. Notre idée est de reproduire le processus morpho-syntaxique de formation des mots-formes et syntagmes en partant de la notion de morphe sans procéder au préalable à une segmentation en mots, puis de détecter les mots-formes et syntagmes lexicalisés grâce à une mesure d'association. Il s'agit concrètement de parcourir le corpus en associant incrémentalement morphes et mot-formes, puis de sélectionner les combinaisons stables, récurrentes et indépendantes sur l'ensemble du corpus. Ici les statistiques et ressources lexicales à notre disposition servent à décider quels sont les éléments à associer à chaque étape de notre processus. L'essence de notre méthode est le principe de construction incrémentale de la lexie, davantage que le score d'association utilisé.

Dans la section précédente, nous avons expliqué que l'unité minimale porteuse de sens est le morphe et que la construction lexicologique chinoise s'appuie sur l'association de composantes lexicales. C'est pourquoi nous initialisons notre processus d'identification par une représentation du texte en morphes (P_0). Nous avons également vu que la production des lexies en chinois s'appuie autant sur la syntaxe que sur la morphologie. Nous posons un postulat (P_n) pour décrire les phénomènes de production morphologique et syntaxique, sachant que nous nous intéressons uniquement aux syntagmes présents en séquence continue dans le corpus :

Postulat d'initialisation (P_0) : *Le morphe est l'unité de construction élémentaire des lexies en chinois contemporain.*

Postulat de récursion (P_n) : *Les mots-formes du chinois moderne sont une séquence de deux morphes ou mot-formes. Les syntagmes sont une combinaison syntaxique de mots-formes.*

Ces deux postulats permettent de poser les bases de notre méthode en définissant l'unité de recherche initiale et en décrivant le processus de construction des lexies.

[2]Dans la suite de l'article, le terme « lexie » sera utilisé pour désigner un « mot-forme ou syntagme lexicalisé représentant d'une lexie » lorsque nous parlerons des lexies découvertes dans le texte.

4.1 Description des corpus

Le corpus de référence contient 2 millions de graphies. Il est composé de deux corpus de chinois contemporain : *The Lancaster Corpus of Mandarin Chinese* (McEnery & Xiao, 2004) et *PFR China Daily corpus* (Beijing University ICL, 2001) tous deux annotés selon la norme de l'Université de Beijing[3]. Il nous a servi à recueillir toutes les informations lexicales nécessaires à notre méthode. En nous basant sur ces annotations, nous avons extrait un lexique de référence d'environ 75 000 lexies, une liste des mots-vides et une liste des bigrammes syntaxiques (voir 4.2.3). Le corpus de test est un corpus de documents commerciaux de compagnies d'assurance publiés sur leur site internet. Il est composé de dix millions de caractères. Il a été recueilli automatiquement à partir d'une liste d'adresses de sites de compagnies d'assurance chinoises. Les textes sur internet contenant de nombreuses répétitions (par exemple dans les menus), tous les paragraphes doublons ont été éliminés afin de ne pas perturber les calculs statistiques.

4.2 Algorithme

Les lexies candidates sont construites de manière incrémentale en associant progressivement des couples de grains. Un *grain* est constitué d'une séquence de morphes pouvant représenter un morphe, un mot-forme, un syntagme ou une partie de syntagme. Le morphe est le niveau de granularité initiale. Nous déduisons ensuite le niveau de granularité suivant en associant des couples de grains. Nous utilisons une mesure d'association pour nous orienter dans le choix des couples de grains à associer et la sélection des lexies candidates. Le score d'association entre les grains est recalculé à chaque changement de granularité afin d'intégrer les déductions obtenues lors des étapes précédentes. L'algorithme s'arrête lorsqu'un plafond de nombre d'associations maximal est atteint[4]. L'étape finale consiste à sélectionner et ordonner les lexies candidates en fonction de leur pertinence.

Algorithme d'extraction de lexies candidates

```
Sélectionner des grains initiaux (P_0)
Faire
| Calculer le score d'association des grains
| Sélectionner les grains de niveau supérieur (P_n)
Tant que le plafond d'association maximal n'est pas atteint
Sélection et ordonnancement des lexies candidates
```

4.2.1 Grains initiaux : morphes

Nous considérons trois types de morphes : les morphes unigrammes, les morphes bigrammes (蝴蝶 HÚDIÉ 'papillon') et les morphes translittérés (意大利 YÌDÀLÌ 'Italie'). Un *morphe unigramme* est une graphie à laquelle est associée un sens. Nous considérons que toutes les graphies sont potentiellement des morphes à l'exception de celle incluses dans un morphe bigramme ou translittéré. Un *morphe translittéré* est un calque phonétique d'un mot étranger utilisant la prononciation de graphies. Les graphies utilisées pour le calque phonétique sont pour la plupart récurrentes et il est possible de les détecter avec des outils statistiques[5]. Un morphe bigramme

[3] *http://icl.pku.edu.cn/icl_groups/corpus/coprus-annotation.htm*

[4] Estimé à 5 associations selon nos observations sur les lexies les plus grandes dans le corpus de référence.

[5] Nous avons utilisé pour cela l'implémentation CRF++ de la théorie des *Conditional Random Fields*.

est un couple de graphies ne prenant sens que dans la mesure où elles sont associées (蝴 HÚ '∅'). Il se peut que l'une des graphies du couple puisse exister en tant que morphe porteur du même sens (蝶 DIÉ 'papillon'), cependant il n'est jamais autonome (**蝶泳 DIÉ**YǑNG 'nage papillon'). Leur détection est faite à l'aide de leurs particularités statistiques[6]. Lorsqu'il y a une ambiguïté entre différents morphes, le morphe englobant est toujours préféré.

4.2.2 Calcul du score d'association

A chaque étape de l'algorithme, le score d'association de tous les couples de grains est calculé à l'aide du score Poisson-Stirling[7] (Quasthoff & Wolff, 2002) défini ainsi :

$$PS(a,b) = \frac{k \cdot (\log\ k - \log\ \lambda - 1)}{\log\ n}$$

où $\lambda = n \cdot p_a \cdot p_b$ avec p_a et p_b la probabilité d'apparition respective de a et b, k le nombre d'occurrences du couple ab et n le nombre total de couples. Ce score reflète la tendance de a et b à se retrouver plus souvent côte à côte que ne le voudrait une répartition aléatoire du corpus. Plus le score est élevé, plus l'association entre a et b est significative, un score nul indiquant la neutralité. Un lien a été mis en évidence entre les mesures d'association et certains phénomènes linguistiques tels que les relations sémantiques ou syntaxiques (Church & Hanks, 1997).

4.2.3 Sélection du niveau de granularité supérieur : mots-formes ou syntagmes

La sélection du niveau de granularité supérieur consiste à choisir des couples de grains à associer pour identifier des lexies candidates. L'idée directrice de la sélection est que si un couple de signes a un score d'association élevé au sein d'un corpus, alors il existe une relation particulière entre eux. Cette relation peut être d'ordre lexical, syntaxique (en chinois un classificateur est souvent associé à un numéral) ou simplement refléter des relations prédicat-argument communes dans le corpus de travail (副总统报道 'le vice-président a déclaré'). Cependant les relations d'ordre contextuel et syntaxique non lexicalisés ne nous intéressent pas et la faible représentativité de certains signes peut également induire en erreur le processus statistique. Pour écarter les associations non-souhaitables, des critères de filtrage sont ajoutés pour la sélection des couples. Notre méthode travaille sur l'ensemble des segments du corpus. Un segment est une suite ininterrompue de graphies du corpus situé par exemple entre des signes de ponctuation, des caractères alphabétiques ou le début et la fin de ligne.

Soit S_i l'ensemble des segments de sinogrammes du corpus à l'itération i. Soit le segment $s_i^n \in S_i$ avec n le numéro de segment, composé d'une suite $m_1, m_2 \ldots m_{|s_i^n|}$ de grains telle que $\forall\, m_j \in s_i^n\ ,\ m_j\ \in\ Graphie^+$, comme dans l'exemple (1) ci-dessous :

(1) $S_0 = \{\cdots, (\text{ 意·外·身·故·保·险·金 }), \cdots\},\ s_0^n = (\text{ 意·外·身·故·保·险·金 })$
avec $m_1 =$意 et $m_7 =$金.

Le passage d'une granularité courante à une granularité supérieure se fait en appliquant la règle suivante sur l'ensemble du segment.

[6]Fort score d'association, exclusivité d'association d'un ou des composants.

[7]Nous avons également testé l'information mutuelle avec des résultats équivalents modulo un gain de rappel et une perte de précision.

$$\forall\ s_i^n\ \ \forall\ m_x m_y \in s_i^n\ \ \text{si}\ \ \lceil m_x \leftrightarrow m_y \rceil^{s_i^n}$$
$$\text{alors}\ \ m_x m_y\ \ \text{est remplacé par}\ \ m_{xy} = m_x \cdot m_y\ \ \text{dans}\ \ s_{i+1}^n$$

où $\lceil m_x \leftrightarrow m_y \rceil^{s_i^n}$ signifie que « m_x et m_y sont associés dans le contexte du segment n à l'étape i ». Par exemple, si $\lceil$意↔外$\rceil^{s_0^n}$ $\lceil$身↔故$\rceil^{s_0^n}$ $\lceil$保↔险$\rceil^{s_0^n}$ alors $s_1^n =$ (意外·身故·保险·金). La proposition $\lceil m_x \leftrightarrow m_y \rceil^{s_i^n}$ n'est vraie que si les conditions suivantes sont remplies :

1. *Association interne* : $PS(m_x, m_y) > 0$
2. *Liberté droite* : $PS(m_x, m_y) > PS(m_y, m_{y+1})$
3. *Liberté gauche* : $PS(m_x, m_y) > PS(m_{x-1}, m_x)$

et que les conditions de filtrage suivantes sont valides pour le couple $m_x m_y$:

1. *Filtre syntaxique* : Le couple n'appartient pas à la liste des collocations syntaxiques.
2. *Filtre score/fréquence* : Le couple a une fréquence supérieure à 10 (garantit la représentativité statistique) ou un score d'association supérieur à 1 (élimination des cas tangents).
3. *Absence de mot vide* : Aucun grain n'est un mot vide sauf si le couple est dans le lexique.

La liste des collocations syntaxiques et des mots vides a été obtenue à partir de notre corpus de référence. Nous appelons *collocation syntaxique* tout bigramme ayant un score d'association supérieur à zéro dans le corpus de référence mais n'étant pas répertorié comme lexie dans ce corpus. La liste des *mots vides* est composée de l'ensemble des particules modales, prépositions, conjonctions de coordination rencontrées dans les textes écrits du corpus de référence.

4.2.4 Sélection et ordonnancement des lexies candidates

Les lexies sélectionnées sont l'ensemble des associations valides ($\lceil m_x \leftrightarrow m_y \rceil^{s_i^n}$) trouvées à chaque étape de l'algorithme qui ne se trouvent pas dans notre lexique. Des seuils de fréquence et score d'association[8] sont également utilisés pour éliminer les candidats les moins probables en fin de traitement. Nous considérons qu'une lexie est moins pertinente si elle est régulièrement contenue dans une autre ; nous ajustons donc sa fréquence dans le calcul final de son score d'association en lui soustrayant le nombre d'occurrences incluses du nombre d'occurrences total. Les candidats sont ensuite ordonnés par ordre décroissant de score d'association ajusté.

4.3 Protocole d'évaluation

Dans la mesure où le corpus de travail est brut, sans aucune annotation et très volumineux (donc coûteux à annoter), nous utilisons des échantillons pour évaluer les résultats. Les mesures qui nous intéressent sont les scores de précision et de rappel. La précision est définie comme le pourcentage de lexies correctes dans la liste de lexies candidates extraites, et le rappel le pourcentage de lexies inconnues découvertes. Nous estimons le score de précision en sélectionnant aléatoirement un échantillon de 1 000 lexies candidates dans la liste produite par notre système. L'estimation du score de rappel est obtenue en repérant les lexies inconnues présentes dans une

[8] Différents du filtre score/fréquence de la sous-section 4.2.3.

sélection aléatoire de 250 paragraphes dans le corpus de travail, dans lesquels 287 lexies inconnues distinctes ont été identifiées. L'identification et la validation des lexies du corpus ont été effectuées par un locuteur natif du chinois et linguiste sensibilisé à notre problématique en appliquant la définition de la lexie donnée en section 3. Nous attendons que notre système produise une liste de lexies inconnues et nous souhaitons que la distribution de lexies valides soit plus dense en haut de liste qu'en fin de liste.

5 Résultats

Nous avons effectué une expérimentation (tableau 2) en sélectionnant les lexies candidates de fréquence supérieure à 8 et de score d'association supérieur à 10 (voir 4.2.4). Notre méthode extrait après cette sélection une liste de 26 388 lexies candidates parmi lesquelles on estime à 13 880 le nombre de lexies valides (extraits en tableau 3). L'ensemble de ces lexies valides couvre 53,7 % des lexies de l'échantillon du corpus de travail. La précision des résultats décroit avec l'augmentation des centiles considérés mais reste au-dessus de 50 %. La majorité des erreurs (60,7 %) sont des syntagmes non-lexicalisés, c'est-à-dire des combinaisons syntaxiques de mots-formes n'ayant pas de contenu lexical particulier. Ces syntagmes peuvent cependant contenir des lexies inconnues présentes dans les candidats valides ou non. Une petite partie des erreurs (6.5 %) sont des collocations (syntagmes à l'usage contraint). Les collocations ne sont pas pertinentes pour les applications d'analyse de la langue, nous ne les avons donc pas incluses dans les résultats valides. Notons que 25 % des erreurs contiennent des lexies déjà contenues dans la liste des lexies découvertes. Enfin, il est important de remarquer que 92 % des lexies non découvertes ont moins de 8 occurrences dans le corpus de travail. Ceci montre que les lexies peu fréquentes représentent une part significative des lexies du corpus. La diminution du seuil de fréquence des candidats dégradant de façon significative le score de précision, notre méthode n'est actuellement pas en mesure de récupérer efficacement ce type de lexies inconnues.

Précision / Rappel (estimation)				**Détail des erreurs**		
Centile	***Prec.***	***Rapp.***	***Nb Lexies***	***Type***	***%***	***Exemple***
10	71,0 %	29,6 %	1 874	Collocations	6,5 %	Souscrire une assurance 买保险
50	60,2 %	46,0 %	7 943	Syntagmes non lexicalisés	60,7 %	le responsable participe 负责人参加
100	**52,6 %**	**53,7 %**	**13 880**	Erreurs de segmentation	32,7 %	la direction constr... 公司总部设

TAB. 2 – *Résultats et détails des erreurs.*

Lexies	**Rang**	**Fréq.**	**Lexies**	**Rang**	**Fréq.**
institution financière 金融机构	58	536	carte d'assurance maladie 医保卡	9 758	15
Xincheng Assurance Vie 信诚人寿保险有限公司	293	138	ordinateur portable 笔记本电脑	12 615	10
PDG 首席执行官	405	105	politique de soutien à l'agriculture 支农惠农	13 062	10
assurance responsabilité d'agence de voyage 旅行社责任险	5 356	24	procédure de réclamation d'indemnités 索赔流程	24 603	8

TAB. 3 – *Echantillon de lexies valides extraites.*

6 Discussion et conclusion

Notre expérimentation valide le potentiel d'extraction de notre méthode qui découvre un nombre non négligeable de lexies inconnues dans le corpus de travail. Nous avons également confirmé l'utilité du critère d'ordonnancement des candidats, la densité des lexies candidates valides étant plus élevée en début de liste. Pour compléter l'évaluation de notre méthode, il est nécessaire de la comparer à d'autres travaux. Cette comparaison doit cependant se faire au regard de l'influence des genres et usages des corpus respectifs car — comme l'indique Aussenac-Gilles (2002) — le résultat d'une méthode est conditionné par les caractéristiques du contexte d'application : « [Les] *outils* [sont] *construits en référence à des théories spécifiques, qui induisent des contraintes d'utilisation et des biais dans les résultats obtenus [...] leur utilisation* [est] *plus ou moins pertinente en fonction de la nature du corpus et du produit terminologique à construire* ». Ainsi, la méthode de Feng *et al.*, qui s'appuie sur la variation du contexte, est particulièrement adaptée aux dépêches d'agences de presse (rappel 74,2 % et précision 81,2 %). En effet, ce type de corpus abordant de nombreux thèmes, les lexies inconnues provenant de domaines divers sont nombreuses, et les contextes d'apparition plus variés. En revanche dans les corpus spécialisés — où les lexies inconnues sont en partie spécialisées — certaines formulations sont récurrentes et un syntagme non lexicalisé peut apparaître très régulièrement, ce qui génère du bruit (cause de 60,7 % de nos erreurs). Cette difficulté est accentuée dans le discours commercial particulièrement répétitif. Les résultats de la méthode de Fung et Wu sur des comptes-rendus de conseils législatifs (rappel 14,0 % et précision 59,3 %) sont significatifs de la difficulté à traiter ce type de corpus. Face à ce constat, nous pensons qu'il est nécessaire de mettre en place une référence de test commune via un protocole d'évaluation de la tâche d'extraction de lexique en chinois. Ce protocole doit définir clairement l'unité linguistique recherchée et mettre en place un jeu de tests et de métriques tenant compte de la diversité des genres et usages dans les corpus.

Pour conclure, notre objectif est de fournir une aide à la constitution de lexique à partir de corpus. Dans le cadre de cette application, les résultats produits sont corrects aux vues de la quantité et de la qualité des lexies extraites. En relâchant la contrainte de fréquence à 3, nous obtenons un rappel avoisinant les 67,5 % et le nombre de lexies candidates triple (environ 75 000 candidats). Cette amélioration de la couverture se fait cependant au détriment de la qualité des résultats en fin de liste de lexies candidates. Pour autant, grâce à l'ordonnancement des candidats, la qualité des éléments en début de liste n'est que peu influencée. Néanmoins la détection des lexies à faible fréquence est importante car elle représente une part non négligeable du contenu lexical d'un corpus. Par ailleurs, l'extraction de lexies connues à valeur spéciale dans le corpus via le calcul des spécificités lexicales (Drouin, 2004) doit être également envisagée pour compléter nos résultats. Enfin, notre méthode n'organise ni ne distingue les lexies générales et spécialisées. L'intégration d'une analyse des caractéristiques endogènes des lexies permettrait une meilleure organisation, sélection et exploitation des résultats produits. Les futurs développements de notre méthode iront dans ces directions.

Remerciements

Je tiens à remercier spécialement Pierre Zweigenbaum pour ses recommandations pertinentes et ses relectures attentives. Je remercie également chaleureusement mes collègues d'Arisem, en particulier Nicolas Dessaigne et Stéphanie Brizard, pour leur soutien et leurs relectures.

Références

AUSSENAC-GILLES N., CONDAMINES A. & SZULMAN S. (2002). Prise en compte de l'application dans la constitution de produits terminologiques. In *Actes des 2e Assises Nationales du GDR I3*, p. 289–302.

BEIJING UNIVERSITY ICL (2001). PFR china daily corpus.

CHANG J. & SU K. (1997). An unsupervised iterative method for Chinese new lexicon extraction. In *Computational Linguistics*, Computational Linguistics, p. 22–29.

CHURCH K. W. & HANKS P. (1997). Word association norms, mutual information and lexicography. In M. PRESS, Ed., *Computational Linguistics*, volume 16, p. 22–29.

DROUIN P. (2004). Spécificités lexicales et acquisition de la terminologie. In *JADT 2004 : 7ème Journées internationales d'Analyse statistique des Données Textuelles*, p. 345–352.

FENG H., CHEN K., DENG X. & ZHENG W. (2004). Accessor variety criteria for Chinese word extraction. In *ACL*, volume 30, p. 75–93, Cambridge.

FUNG P. & WU D. (1994). Statistical augmentation of a Chinese machine-readable dictionary. In *WVLC-94, Second Annual Workshop on Very Large Corpus*.

L'HOMME M.-C. & POLGUÈRE A. (2008). Mettre en bons termes les dictionnaires spécialisés et les dictionnaires de langue générale. In *Lexicologie et terminologie: histoire de mots. Colloque en l'honneur d'Henri Béjoint*, p. 191–206, Lyon.

LI H., HUANG C., GAO J. & FAN X. (2004). The use of SVM for Chinese new word identification. In *IJCNLP 2004, First International Joint Conference*, p. 723–732.

MCENERY T. & XIAO R. (2004). The Lancaster corpus of Mandarin Chinese.

MEL'CUK I. (1993). *Cours de morphologie générale introduction et première partie : Le mot*, volume 1. Les presses de l'université de Montréal - CNRS Edition.

NAKAGAWA H., KOJIMA H. & MAEDA A. (2004). Chinese term extraction from Web pages based on compound word productivity. In *Third SIGHAN Workshop on Chinese Language Processing*, p. 79–85, Barcelona, Spain.

NGUYEN E. V. T. (2006). *Unité lexicale et morphologie en chinois mandarin vers l'élaboration d'un DEC du chinois*. PhD thesis, Université de Montréal.

PIAO S. S., SUN G., RAYSON P. & YUAN Q. (2006). Automatic extraction of Chinese multiword expressions with a statistical tool. In *Workshop on Multi-word-expressions in a Multilingual Context held in conjunction with the 11th EACL 2006*, Trento, Italy.

POLGUÈRE A. (2003). *Lexicologie et sémantique lexicale. Notions fondamentales*. Université de Montréal.

QUASTHOFF U. & WOLFF C. (2002). The Poisson collocation measure and its application. In *Proceedings of the 20th international conference on Computational Linguistics*.

SPROAT R. & EMERSON T. (2003). The first international Chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*.

SUN M., SHEN D. & TSOU B. K. (1998). Chinese word segmentation without lexicon and hand-crafted training data. In *Proceedings of ACL*, p. 1265–1271.

WU A. & JIANG Z. (2000). Statistically-enhanced new word identification in a rule-based Chinese system. In *Proceedings of the 2nd Chinese Language Processing Workshop*, volume 12, p. 46–51, Hong Kong: ACL.

YANG W. & LI X. (2002). Chinese keyword extraction based on max-duplicated strings of the documents. In *Proceedings of the 25th Annual International ACM*, p. 439–440.