Ressources lexicales chinoises pour le TALN

Huei-Chi LIN, Max SILBERZTEIN
Laboratoire de Sémiolinguistique, Didactique, Informatique (LASELDI) –
Université de Franche-Comté, 30 rue Mégevand 25000 Besançon
lin_huei_chi@yahoo.fr
max.silberztein@univ-fcomte.fr

Résumé. Nous voulons traiter des textes chinois automatiquement ; pour ce faire, nous formalisons le vocabulaire chinois, en utilisant principalement des dictionnaires et des grammaires morphologiques et syntaxiques formalisés avec le logiciel NooJ. Nous présentons ici les critères linguistiques qui nous ont permis de construire dictionnaires et grammaires, sachant que l'application envisagée (linguistique de corpus) nous impose certaines contraintes dans la formalisation des unités de la langue, en particulier des composés.

Abstract. In order to parse Chinese texts automatically, we need to formalize the Chinese vocabulary by using electronic dictionaries and morphological and syntactic grammars. We have used the NooJ software to enter the formalization. We present here the set of linguistic criteria used to construct these dictionaries and grammars, so that they can be used by corpus-linguistic applications. We focus our discussion on the characterization of Chinese linguistic units, specifically compounds.

Mots-clés : ressources linguistiques pour le chinois, linguistique de corpus, NooJ.
Keywords: linguistic resources for chinese, corpus linguistics, NooJ.

1 Introduction

Notre but est de formaliser le vocabulaire de la langue chinoise, plus précisément le mandarin tel qu'on le trouve dans les textes littéraires à partir du XX^{e} siècle, et écrit avec les caractères chinois traditionnels codés avec UNICODE (ce qui représente 70.207 caractères) pour les besoins des applications de linguistique de corpus et d'analyse syntaxique automatique. Nous avons donc entrepris la construction d'un module chinois pour NooJ[1]. Ce travail de recherche nous a conduits à construire des dictionnaires électroniques, des grammaires morphologiques et des grammaires syntaxiques. Toutes les unités du vocabulaire chinois doivent être recensées

[1] Cf. http://www.nooj4nlp.net. NooJ ainsi que le module chinois et d'autres ressources peuvent être téléchargés librement, et les utilisateurs de NooJ peuvent développer leurs propres ensembles de ressources linguistiques pour formaliser divers niveaux des langues : orthographe, lexique, morphologie, syntaxe et sémantique. Ces ensembles de ressources peuvent être rassemblées dans un « module » autonome, qui peut ensuite être chargé par d'autres utilisateurs pour analyser des textes de grande taille.

systématiquement et décrites explicitement. Il est impératif de formaliser tous les types d'unités linguistiques, et pas simplement les mots simples, et aussi de décrire leurs variations lexicales et morphologiques, ce que n'ont pas fait systématiquement les dictionnaires traditionnels chinois jusqu'à présent. Par exemple, le dictionnaire 辭海 (Cíhǎi) contient l'entrée lexicale « 畫冊 » (huàcè) [album de peintures], mais ne la relie pas à sa variante orthographique « 畫冊 » (huàcè) [album de peintures], ni à sa variante morphologique « 畫冊兒 » (huàcèér) [album de peintures]. En chinois, il n'y a pas de blanc séparateur de mots dans les textes. La reconnaissance automatique des mots chinois doit donc passer par la consultation de dictionnaires électroniques ou de grammaires morphologiques ou syntaxiques complets ; ce problème ressemble beaucoup à celui de la reconnaissance des mots composés et expressions figées dans les langues romanes, où l'on ne sait pas a priori où s'arrête un mot composé, et où seule un recensement systématique et une description précise permettent de distinguer les mots composés lexicalisés (par ex. « carte bleue ») des séquences libres de mots (« carte marron »). Nous avons dû adopter une série de critères linguistiques précis et surtout reproductibles, pour décider si une séquence de caractères chinois doit ou non être lexicalisée. Ces critères sont différents de ceux utilisés en linguistique traditionnelle : par exemple, dans le dictionnaire 辭海 (Cíhǎi), on trouve des entrées telles que « 白吃 » (báichī) [prendre gratuitement un repas], que nous n'avons pas de raison de lexicaliser, tandis que nous décrivons explicitement l'entrée « 鋼琴家 » (gāngqínjiā) [pianiste], non répertoriée dans le dictionnaire 辭海 (Cíhǎi).

Nous avons développé le module chinois pour NooJ afin d'analyser automatiquement des textes, de façon similaire aux systèmes CKIP et ICTCLAS ; notre module permet également d'identifier les entités nommées, et d'extraire à volonté des motifs syntactico-sémantiques à partir de requêtes d'utilisateurs. Notre module n'a pas d'application directe en traduction automatique, ni pour la comparaison entre le chinois, le japonais et le coréen. Enfin, nos dictionnaires électroniques ne contiennent pas de synonymes (au contraire de WordNet).

2 Définition des unités

Les Unités Linguistiques Atomiques (*Atomic Linguistic Units*, ou ALUs) de NooJ constituent les unités les plus petites de la langue qui doivent être associées à des informations linguistiques. Formellement, NooJ traite les ALUs en quatre classes formelles distinctes :

1) **Affixes** (préfixe, infixe et suffixe) : Ce sont des séquences de caractères chinois décrites par un composant morphologique, ou qui interviennent dans des opérations lexicales de flexion ou de dérivation. Par exemple, le préfixe « 初 » (chū) et le suffixe « 兒 » (ér)[2].

2) **Mots simples** : NooJ traite a priori chaque caractère chinois comme un mot simple. Par exemple, le caractère « 樹 » (shù) [arbre] constitue un mot simple et est utilisé dans les textes dans des contextes libres.

3) **Mots composés** : Ce sont des séquences de caractères que nous devons lexicaliser, comme par exemple « 蝴蝶 » (húdié) [papillon] et « 嘰哩咕嚕 » (jīlīgūlū) [avoir faim].

[2] 初 (chū) est un préfixe déterminatif que l'on place devant les dix premiers numéros de jours des mois, ce qui correspond au « **er** » dans « le 1er janvier ». 兒 (ér) est un suffixe phonétique pur (qui ne change pas le sens des mots), qu'on utilise après certains noms, verbes, adjectifs ou adverbes.

4) **Expressions figées** : Ce sont des séquences de caractères potentiellement discontinues. Par exemple, à l'intérieur de l'expression « 拖下水 » (tuō xiàshuǐ) [implanter dans l'eau — impliquer quelqu'un dans une situation], on peut insérer des pronoms personnels comme « 我 » (wǒ) [je], « 你 » (nǐ) [tu], « 他 » (tā) [il], etc. :

你要**拖**他**下水**。 (nǐ yào **tuō** tā **xiàshuǐ**) [Tu veux l'impliquer dans cette situation.]

NooJ traite différemment les affixes, mots simples et mots composés, qui sont des séquences insécables constitués d'un ou de plusieurs caractères, des expressions figées qui elles peuvent être discontinues. Il s'agit donc d'une part d'intégrer et de décrire les quatre types d'ALUs du chinois dans des dictionnaires : d'affixes, de mots simples, de mots composés et d'expressions figées ; et d'autre part, de construire des grammaires qui permettent de décrire les conditions d'utilisation et de combinaison de ces ALUs.

Par rapport aux approches lexicographiques traditionnelles, les expressions idiomatiques, les mots polymères chinois (cf. ci-dessous) et les mots surcomposés « en paire » (mots de 2x2 caractères) ont été simplement intégrés à nos dictionnaires puisque ce sont des ALUs comme les autres du point de vue du TALN. Par ailleurs, certaines ALUs ont été formalisées selon deux méthodes : l'une est de les ranger directement dans un dictionnaire ; l'autre est de les représenter à l'intérieur de grammaires locales, morphologiques ou syntaxiques.

Dans l'alphabet chinois traditionnel (au contraire de l'alphabet chinois simplifié), il existe de nombreuses variantes orthographiques, i.e. lorsqu'un mot ou un morphème s'écrit avec deux orthographes ou plus, et se prononce de la même façon dans tous les cas. Plusieurs variantes d'un même caractère peuvent cohabiter dans un même texte. Nous avons donc entré une table d'équivalence qui contient plus de 1.000 paires, telles que 巃 = 龍 et 悪 = 惡, et nous avons modifié l'algorithme de consultation des dictionnaires de NooJ pour qu'il prenne en compte ces équivalences.

3 Critères de lexicalisation

Nous présentons une série de critères qui nous ont permis de formaliser le vocabulaire chinois. Pour ce faire, nous avons dû adapter les critères utilisés avec NooJ pour décrire les ALUs des langues romanes (présentées dans Silberztein 1993) ; de plus il a fallu résoudre les difficultés spécifiques au chinois, principalement :

1) Du point de vue orthographique, la reconnaissance automatique des ALUs chinoises est plus complexe que celle des ALUs pour les langues romanes : un caractère chinois peut soit correspondre à une ALU autonome (comme un « mot simple » français), soit à un composant d'un ensemble d'ALUs productifs (« préfixe » que l'on traite avec des règles morphologiques), soit un composant d'une ALU plus longue (comme un « mot composé ») (Lin 2006).

2) En chinois, bien plus que dans les langues romanes et l'anglais, les mots ont systématiquement plusieurs fonctions syntaxiques. Par exemple, le même mot « 解釋 » (jiěshì) [expliquer ou explication] peut être indifféremment un verbe ou un nom :

他向我**解釋**他昨天缺席的原因。 [Il m'**explique** pourquoi il a été absent hier.]

我接受了他的**解釋**。 [J'ai accepté son **explication**.]

Les ambiguïtés concernent aussi les adjectifs et les adverbes :

他**和藹可親**。 [Il est gentil.] 他**和藹可親**地與我說話。 [Il m'a parlé gentiment.]

les adjectifs et les noms, etc. Pour calculer la fonction d'un mot dans la phrase (ce qui correspond à l'étiquetage pour les langues romanes), il faut donc bien souvent analyser syntaxiquement préalablement la phrase complète. Il est donc impossible de commencer l'analyse d'un texte par une étape d'étiquetage et de levée d'ambiguïtés : l'analyseur lexical des textes chinois produira donc nécessairement un résultat massivement ambigu, qui sera transmis à l'analyseur syntaxique. L'architecture spécifique de NooJ, qui permet d'étiqueter les textes partiellement ambigus en produisant une structure d'annotations potentiellement ambiguë, est donc bien adaptée à l'analyse lexicale des textes chinois.

3.1 Compositionnalité

La majorité des ALUs chinoises sont constituées d'au moins deux caractères. Parmi ces mots, certains ne peuvent pas être décomposés car ils contiennent des caractères non-autonomes, i.e. qu'on ne trouve nulle part ailleurs que dans ces mots. Considérons les mots ci-dessous :

蝴蝶 (húdié) [papillon], 嘰哩咕嚕 (jīlīgūlū) [avoir faim]

Ces deux mots contiennent des caractères qui n'ont pas d'utilisation autonome, par exemple, on ne trouve pas le caractère « 蝴 » en dehors du mot [papillon], et on ne trouve pas le caractère « 嘰 » en dehors de [avoir faim]. Il faut donc recenser et décrire ces mots dans un dictionnaire.

En revanche, certains mots peuvent être constitués de caractères qui peuvent avoir d'autres emplois de façon indépendante. Par exemple, considérons le mot suivant :

白菜 (báicài) [chou chinois]

Ce mot est constitué du caractère « 菜 » (qui signifie « légume ») et du caractère « 白 » (qui signifie « blanc »). Mais le sens du mot [chou chinois] ne peut pas être déterminé à partir du sens de ses deux constituants. En conséquence, il faut absolument lexicaliser ce mot, d'une part pour obtenir la bonne analyse du mot, et d'autre part, pour bloquer une analyse compositionnelle qui produirait le résultat incorrect « légume blanc » (un légume blanc s'écrirait « 白的菜 »).

3.2 Institutionnalisation

Beaucoup de concepts ou objets du monde réel sont désignés ou nommés systématiquement de la même façon par les locuteurs d'une langue, et ce de façon arbitraire. Par exemple, le concept [cœur sensible] s'exprime en chinois par le terme « 豆腐心 », littéralement « cœur de tofu », alors que d'autres expressions très semblables, telles que « 牛奶心 » (« cœur de lait ») ou « 白紙心 » (« cœur de purée ») ne peuvent pas utilisées pour exprimer ce concept.

Les formes dont l'usage est « institutionnalisé » ne sont pas morphologiquement, syntaxiquement ou sémantiquement différentes des autres formes potentielles qui ne sont jamais utilisées par les locuteurs chinois. Il faut donc distinguer formellement les termes vraiment employés par les locuteurs chinois des expressions potentielles qui ne le sont jamais.

Dans certaines applications de la formalisation des langues, telles que la traduction automatique ou l'enseignement des langues secondes, il est nécessaire de lexicaliser ces termes, qui correspondent à la « bonne » traduction, ce qui permet d'éviter des fautes d'analyse ou de traduction. Ainsi, pour traduire correctement l'expression française « cœur sensible », il faut produire « 豆腐心 » sans chercher à analyser les deux constituants de l'expression française ou de sa traduction : on traduit donc le tout « en bloc », ce qui revient à dire que l'expression complète est une ALU lexicalisée. La comparaison systématique entre les termes et des expressions similaires potentielles montre d'une part, que leur structure syntaxique n'est pas spécifique, d'autre part que la construction des termes ne peut pas être calculée par des règles morphologiques, syntaxiques ou sémantiques.

3.3 Structure des mots composés

A chaque fois que certaines propriétés syntaxiques ou sémantiques d'une forme chinoise ne peuvent pas être calculées à partir de celles de ses constituants, il faut lexicaliser cette forme et la traiter en tant qu'ALU, i.e. « en bloc ». Cependant, on ne peut pas ne pas noter que de nombreux termes composés chinois se construisent selon quelques schémas productifs. Nous décrivons ces schémas.

3.3.1 Mots polymères

Les mots polymères se composent d'au moins quatre caractères potentiellement autonomes dont les sens sont « similaires ». Sémantiquement, les polymères chinois sont construits par des mécanismes relevant de la coordination. Les caractères constituant un mot polymère ne peuvent pas être remplacés par d'autres caractères. Cependant, leur ordre d'apparition peut être modifié à l'intérieur du polymère. Par exemple, le terme « 紙墨筆硯 » qui signifie [trousse] peut s'écrire de huit façons différentes :

紙墨筆硯 (zhǐ mò bǐ yàn) [papier, encre, pinceau, encrier], mais aussi :

紙筆硯墨 紙筆墨硯 紙硯筆墨 筆硯紙墨 筆墨硯紙 筆墨紙硯 硯墨紙筆

80% des mots polymères se composent de quatre caractères. Les polymères sont fréquemment employés dans des textes. On distingue traditionnellement trois sortes de polymères :

1) Les quatre caractères ont des sens semblables. Par exemple :

酸甜苦辣 (suāntián kǔlà) [aigre, doux, amer, âcre → aléas de la vie]

Les quatre caractères appartiennent au même champ sémantique.

2) Les quatre caractères sont semblables deux à deux ; les deux premiers forment une paire, et les deux derniers forment une autre paire. Par exemple :

兄弟姊妹 (xiōngdì jiěmèi) [grand frère, petit frère, grande sœur, petite sœur → frères et sœurs]

姊妹兄弟(jiěmèi xiōngdì) [grande sœur, petite sœur, grand frère, petit frère → frères et sœurs]

Ce terme désigne l'ensemble des enfants d'une même famille.

3) Les deux premiers caractères qualifient les deux derniers. Par exemple :

金銀珠寶 (jīnyín zhūbǎo) [or, argent, objet d'une grande valeur → trésor]

Ici il n'y a pas de possibilité de permutation des caractères. « 金 » (jīn) et « 銀 » (yín) appartiennent à la même classe sémantique : métaux précieux, et qualifient le mot « 珠寶 » (zhūbǎo), qui représente la deuxième partie de ce polymère.

3.3.2 Mot radical + suffixe signifiant

Certains termes sont constitués d'un suffixe signifiant associé à un ensemble spécifique de mots de distribution restreinte. Par exemple, le mot « 團 » (tuán) [groupe] peut être combiné avec une centaine de mots selon la règle morphologique productive chinoise : **mot radical + Suffixe signifiant** :

合唱團 (héchàngtuán) [groupe de chanteurs = chorale]

訪問團 (fǎnwèntuán) [groupe d'interviewers = équipe journalistique]

觀光團 (guānguāngtuán) [groupe de touristes]

tandis que d'autres suffixes (匠 (jiàng), 員 (yuán), 商 (shāng), 家 (jiā), etc.) s'utilisent avec d'autres mots. Plutôt que de lexicaliser toutes ces formes dans un dictionnaire, il vaut mieux construire une grammaire locale NooJ qui les décrit de façon unifiée, cf. Figure 1 :

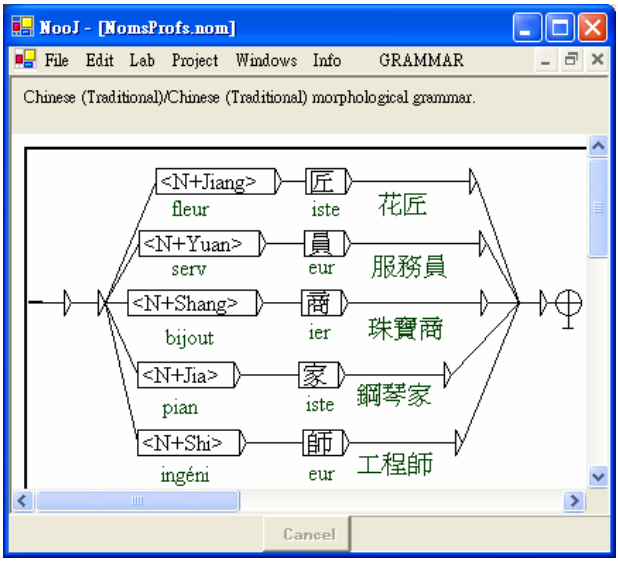

Figure 1 : Une grammaire locale de termes composés

Dans les grammaires de NooJ, les symboles comme <N+Jiang> réfèrent à des informations lexicales et représentent tous les noms (N) associés à la propriété « **+Jiang** » (qui ont été marqués comme pouvant être suivis du suffixe 匠). Noter que la description en extension de

ces ALUs dans une grammaire locale représente le fait que leur mode de construction est productif, mais pas que ces termes sont analysables : cette situation se retrouve en français, où par exemple les termes en « assurance » sont productifs : *assurance maladie*, *assurance vie*, *assurance chômage*, etc. mais chacun des termes construits est non-analysable (par ex. une *assurance vie* n'est pas une assurance contre la vie).

3.3.3 Mots composés de structure **XY**

En chinois, certains mots composés peuvent être créés sans conjonction. Ces mots composés sont classés en trois catégories et doivent être aussi rangés dans un dictionnaire :

1) Les deux constituants sont des lemmes lexicaux semblables. Par exemple :

金漿玉醴 (jīnjiāng yùlǐ) [liqueur d'or + vin de jade → un vin délicieux]

Il s'agit donc d'un mot composé de deux mots « 金漿 » et « 玉醴 » de même classe sémantique qui se combinent.

2) Les deux constituants sont synonymes. Par exemple :

公子哥兒 (gōngzǐ gēér) [jeune homme, garçon → fils de riche]

Ici, le deuxième constituant est un synonyme du premier. En chinois, beaucoup de termes sont construits grâce à des répétitions synonymiques. Ici, les deux constituants « 公子 » [jeune homme] et « 哥兒 » [garçon] sont obligatoirement juxtaposés, et ne peuvent pas être utilisés pour construire le sens de « fils de riche » dans une phrase s'ils sont isolés l'un de l'autre.

3) Chaque constituant décrit une partie d'une image

Lorsqu'ils se juxtaposent, les constituants créent alors une image métaphorique. Dans :

小橋流水 (xiǎoqiáo liúshuǐ) [petit pont, eau courante → un beau paysage]

on trouve les deux constituants « 小橋 » (petit pont) et « 流水 » (eau courante).

3.3.4 Mots composés de structure **AXBX**

Dans certains termes, les deux constituants XX s'intercalent avec deux formes A et B qui sont semblables, synonymes ou antonymes. Par exemple :

上**X**下**X** (shàng X xià X) [dessus X dessous X]　東**X**西**X** (dōng X xī X) [est X ouest X]

好**X**歹**X** (hǎo X dǎi X) [bon X mauvais X]　左**X**右**X** (zuǒ X yòu X) [gauche X droit X]

Trois cas de figure se présentent :

1) Les caractères X sont synonymes, par exemple :

左思右想 (zuǒ sī yòu xiǎng) [gauche réfléchir droite penser → réfléchir pendant longtemps]

2) Les caractères X appartiennent à la même classe sémantique, par exemple :

七手八腳 (qī shǒu bā jiǎo) [septs mains, huit pieds → agitation désordonnée]

Les mots simples « 手 » [main] et « 腳 » [pied] sont des membres du corps.

3) Les caractères X sont identiques, par exemple :

好說歹說 (hǎo shuō dǎi shuō) [bon parler mauvais parler → patiemment]

3.4 Locutions idiomatiques

On trouve beaucoup de locutions idiomatiques dans les textes chinois, et leur structure est plus ou moins régulière. Les locutions idiomatiques peuvent être formalisées avec NooJ de deux façons complémentaires : lorsque ces locutions sont insécables, on peut les ranger dans des dictionnaires, tout comme des mots composés ; lorsqu'elles admettent des insertions possibles (comme en français l'expression « prendre ... en compte »), on doit les traiter avec des grammaires syntaxiques locales, cf. (Silberztein 2007). Par exemple, dans l'expression suivante :

坐冷板凳 (zuò lěng bǎndèng) [s'asseoir sur un banc froid → être mal traité]

les constituants ne sont pas forcément juxtaposés, ce qu'on peut voir dans le texte suivant :

他已**坐**過了**冷板凳**。 (tā yǐ ***zuò*** guò le **lěng bǎndèng**) [Il s'est déjà trouvé dans la situation d'être mal traité]

Par ailleurs, certaines locutions sécables admettent des variantes orthographiques que l'on peut représenter dans les dictionnaires de NooJ. Par exemple, dans les deux expressions synonymes suivantes :

拖下水 (tuō xiàshuǐ) [implanter dans l'eau → impliquer quelqu'un dans une situation]

拉下水 (lā xiàshuǐ) [implanter dans l'eau → impliquer quelqu'un dans une situation]

扯下水 (chě xiàshuǐ) [implanter dans l'eau → impliquer quelqu'un dans une situation]

le verbe « 拖 » (tuō) peut être remplacé par les verbes « 拉 » (lā) et « 扯 » (chě) sans aucun changement de sens.

Certaines locutions admettent des variantes productives, que l'on peut traiter avec le module morphologique de NooJ, un peu comme on traite les variantes orthographiques des langues romanes. Par exemple, la locution suivante :

芝麻綠豆官 (zhīmá lǜdòu guān) [sésame, haricot mungo, fonctionnaire → petit fonctionnaire]

accepte quatre variantes que l'on peut représenter à l'aide d'un automate fini dans NooJ, cf. Figure 2.

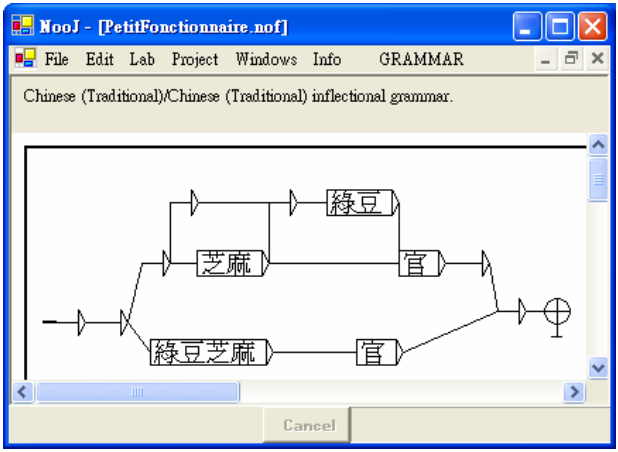

Figure 2 : Dérivation de la locution 芝麻綠豆官 [petit fonctionnaire]

4 Conclusion

Notre formalisation du vocabulaire chinois est fondée sur une classification en quatre classes d'unités linguistiques atomiques : affixes, mots simples, mots composés et expressions figées, qui correspondent à des critères purement orthographiques. Puisque la plupart des unités linguistiques chinoises sont indistinguables des simples séquences de caractères (il n'y a pas de blanc en chinois), il est indispensable de se doter de critères syntaxiques et sémantiques précis et reproductibles pour décider si une forme ou expression doit ou non être décrite dans un dictionnaire, plutôt que d'être traitée comme une séquence analysable de mots.

1) Le **critère de la compositionalité** vérifie si toutes les propriétés d'une forme peuvent ou non être calculées à partir des constituants de la forme.

2) Le **critère d'institutionnalisation** vérifie si une forme ou expression est ou non utilisée de façon systématique et privilégiée par rapport à d'autres expressions potentielles.

Les mots et expressions chinois sont souvent construits sur des schémas morpho-syntaxiques productifs (par exemple les polymères), bien connus des linguistes traditionnels chinois. Nous avons donc classé les termes selon leur structure, ce qui du même coup rend la maintenance de nos dictionnaires plus facile.

Nous avons dû modifier NooJ pour qu'il puisse prendre en compte des phénomènes spécifiques au chinois, comme par exemple la variation systématique de certains caractères (par exemple 羣=群) et les opérations morphologiques de réduplication (快樂 devient 快快樂樂). L'ensemble des outils lexicaux de NooJ (dictionnaires, grammaires flexionnelles, grammaires morphologiques et grammaires syntaxiques locales) ont été utilisés.

Le dictionnaire chinois de NooJ contient actuellement 93.013 entrées lexicales, dont 9 % sont des caractères autonomes (« mots simples ») ou non-autonomes, 27 % sont des mots de deux caractères, 24 % sont des mots composés de trois caractères et 40 % sont des mots de quatre

caractères ou plus. Le module chinois contient aussi des dictionnaires spécialisés, tels que un dictionnaire de noms de famille, un dictionnaire de toponymes, etc. ainsi que plus de vingt grammaires, dont des grammaires morphologiques et des grammaires syntaxiques locales, telles que la grammaire de date, de lieux, de noms de personne, etc.

Nous avons développé, testé et affiné ce module à partir d'un corpus d'une cinquantaine de textes littéraires (par exemple Sìshìtóngtáng de Lao She et *Chéngnán jiùshì* de Lin Haiyin) ; après enrichissement des données, le module couvre la totalité du vocabulaire de ces textes, en incluant la reconnaissance de motifs syntaxiques tels que les noms professionnels. Le module chinois de NooJ peut être téléchargé à partir du site WEB de NooJ : http://www.nooj4nlp.net.

Références

CAO WEI 曹炜. (2004). *Xiandai hanyu cihui yanjiu 现任汉语词汇研究*. Beijing 北京 : Peking University Press 北京大学出版社.

FU HUAIQING 符淮青. (2005). *Xiandai hanyu cihui (Zengding ben) 现代汉语词汇（增订本）*. coll. « yuyanxue jiaocai xilie 语言学教材系列 ». 2nd ed. Beijing 北京 : Beijing daxue chubanshe 北京大学出版社.

GE BENYI 葛本仪. (2002). *Xiandai hanyu cihuixue 现代汉语词汇学*. Jinan 济南 : Shandong renmin chubanshe 山东人民出版社.

KANG SHIYONG 亢世勇. (2004). *Mianxiang xinxi chuli de xiandai hanyu yufa yanjiu 面向信息处理的现代汉语语法研究*. coll. « Yuyan wenzi lilun yu yingyong yanjiu wenku 语言文字理论与应用研究文库 ». Shanghai 上海 : Shanghai cishu chubanshe 上海辞书出版社 and Shiji chuban jituan 世纪出版集团.

LIN HUEICHI 林惠祺. (2006). Les problèmes rencontrés dans le domaine de la catégorisation grammaticale du chinois et leurs solutions proposées. *9th INTEX/NooJ Conférence*.

LU SHICHENG 陸師成. (1992). *Cihai 辭海*. Taipei 臺北 : Wenhua tushu gongsi 文化圖書公司.

SILBERZTEIN MAX. (1993). Les groupes nominaux productifs et les noms composés lexicalisés. *Linguisticæ Investigationes* 2, 405-425.

SILBERZTEIN MAX. (2002). *NooJ Manuel* d'utilisation en anglais que l'on peut télécharger à partir de http://www.nooj4nlp.net (environ 200 pages).

SILBERZTEIN MAX. (2007). Frozen expressions and discontinuous annotations. In the Proceedings of Computational Linguistics 2007. Birmingham.

SUN YINXIN 孙银新. (2003). *Xiandai hanyu cisu yanjiu 现代汉语词素研究*. coll. « Zhonghua xueren congshu 中华学人丛书 ». Beijing 北京 : Zhongguo wenshi chubanshe 中国文史出版社.

ZHU DEXI 朱德熙. (2003). *Xiandai hanyu yufa yanjiu 现代汉语语法研究*. coll. « Shangwu yinshuguan wenku 商务印书馆文库 ». Beijing 北京 : Shangwu yinshuguan 商务印书馆.