UNIVERSITE DE FRANCHE-COMTE

ÉCOLE DOCTORALE «LANGAGES, ESPACES, TEMPS, SOCIETES»

Thèse en vue de l'obtention du titre de docteur en

SCIENCES DU LANGAGE

UN MODULE NOOJ POUR LE TRAITEMENT AUTOMATIQUE DU CHINOIS : FORMALISATION DU VOCABULAIRE ET DES TETES DES GROUPES NOMINAUX

Présentée et soutenue publiquement par

Huei-Chi LIN

Le 15 juin 2010

Sous la direction de M. le Professeur Max SILBERZTEIN

et

la co-direction de Mme la Maître de Conférence H.D.R. Zhitang YANG-DROCOURT

Membres du jury :

Joël BELLASSEN,
Directeur de recherche associé à l'institut National des Langues et Civilisations Orientales, Rapporteur
Andrée CHAUVIN-VILENO,
Professeur à l'université de Franche-Comté
Anaïd DONABÉDIAN-DEMOPOULOS,
Professeur à l'institut National des Langues et Civilisations Orientales, Rapporteur
Max SILBERZTEIN,
Professeur à l'université de Franche-Comté
Zhitang YANG-DROCOURT,
Maître de Conférence H.D.R à l'institut National des Langues et Civilisations Orientales.

REMERCIEMENTS

Je dois avant tout une immense gratitude à Monsieur Max Silberztein, mon directeur de thèse. Je le remercie d'avoir montré tant d'intérêt pour mes recherches et de m'avoir si souvent indiqué la meilleure façon de parvenir à mon but. Il m'a fait découvrir un domaine de recherche dont j'ignorais tout jusqu'alors.

Je tiens à remercier également Madame Zhitang Yang-Drocourt grâce à qui j'ai approfondi mes connaissances en linguistique chinoise. Elle m'a fait connaître de nombreux travaux de recherche consacrés aux phénomènes linguistiques en chinois moderne, elle m'a beaucoup encouragée au moment de la rédaction de cette thèse et m'a donné de nombreux conseils concrets pour l'améliorer.

Je veux remercier aussi Monsieur Joël Bellassen, directeur de recherche associé à l'Institut National des Langues et des Civilisations Orientales ; Madame Andrée Chauvin-Vileno, professeur à l'Université de Franche-Comté ; Madame Anaïd Donabédian-Demopoulos, professeur à l'Institut National des Langues et des Civilisations Orientales. Ils m'ont consacré beaucoup de leur temps et ont formulé sur ma thèse maintes remarques constructives.

Je remercie chaleureusement Monsieur Yuan Qi, chercheur au China Center for Information Industry Development à Pékin ; Monsieur Dong Zhendong, directeur du Language Knowledge Department au Computer & Language Information Research Centre, Chinese Academy of Sciences ; Monsieur Zhan Weidong, professeur à Peking University et Monsieur Song Ro, professeur à Beijing Language and Culture University. Nous avons eu plusieurs discussions concernant le développement d'un module chinois en Traitement Automatique des Langues Naturelles.

Un grand merci à Monsieur Mao Xinian, chercheur de France Télécom à Pékin, qui n'a pas épargné son temps pour m'aider à construire les dictionnaires électroniques.

Je remercie également Madame Françoise Roullier-Singh. Elle a lu une partie de mon manuscrit et m'a soutenue pendant mes années parisiennes. Elle a su me redonner espoir lorsque je traversais des périodes du découragement. Je lui dois de précieux aperçus sur la littérature, la vie à Paris (et la cuisine !)

Je tiens à remercier encore Madame Sabine de Barbuat, qui a accepté de corriger la première version de cette thèse. Ses suggestions m'ont permis d'en clarifier la première rédaction.

J'aimerais aussi remercier Monsieur Gérard Gautier. Sans l'intérêt qu'il a montré, je n'aurais pu terminer cette étude dans les délais prévus. Il m'a donné des conseils précieux pour améliorer mon texte.

Je remercie également mes amies taïwanaises Liu Jiayian 劉佳燕 et Wu Yafen 吳雅楓, qui ont toujours trouvé du temps pour me relire.

Je n'oublie certes pas Madame Chen Feifei 鄭斐斐 de Besançon. Elle m'a donné de nombreux conseils et m'a beaucoup appris sur la vie en France.

Je remercie Wu Yuchen 吳毓淳 grâce à qui j'ai pu résoudre maints problèmes de la vie quotidienne.

Je voudrais dire toute ma reconnaissance à mes amis Chen Yiojun 鄭祐君, Du Yujie 杜雨潔, Lin Yuwen 林郁雯, Chaou Roqian 邱柔千, Zhou Guanbei 周冠貝 et Chen Jiexuan 陳玠璇. Je me souviendrai toujours de l'amical soutien qu'ils m'ont apporté.

Enfin, mes plus sincères remerciements vont à mes parents et à mon frère qui m'ont toujours soutenue moralement malgré l'éloignement. Sans eux, cette thèse n'aurait jamais été terminée. Leur soutien et leurs encouragements m'ont permis de mener à bien ce travail.

林惠祺 Lin Huei-Chi

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

Conventions d'écriture

Pour les exemples en chinois, nous donnons :

1) les caractères chinois traditionnels ;

2) leur translittération en pinyin[1] en italique :

Les règles de translittération sont celles de 中文拼音正词法基本规则 *zhōngwén pīnyīn zhèngcífǎ jīběn guīzé* « Basic Rules for Hanyu Pinyin Orthography », publiées par le Comité national des langues du Ministère de l'Éducation de la République populaire de Chine en 1996.

Pour certains caractères, la prononciation peut être différente selon les normes de la Chine continentale ou de Taïwan. Nous essayons d'en tenir compte par une présentation parrallèle, par exempel, 究 *jiū / jiù*. La première prononciation étant celle de la Chine ;

3) une glose mot-à-mot en français entre chevrons < >, si c'est nécessaire selon le contexte :

Dans la glose mot-à-mot, nous séparons les mots par un tiret, par exemple, <rond-table>, <nous-voir-un-ami>. Nous utilisons parfois directement le pinyin, par exemple, « De » pour noter le subordinateur nominal 的 ou « Le » pour noter la particule modale finale 了. La lettre « Q » représente un quantifieur ;

4) la traduction en français entre guillemets simples ' '.

[1] Nous avons converti les caractères chinois en pinyin de façon automatique à l'aide du service du site Chine-Nouvelle.com. Nous avons ensuite corrigé les erreurs de ton et de segmentation commises par ce service libre, en nous fondant sur les translittérations données par des dictionnaires chinois tels que 康熙字典 *Kāngxī zìdiǎn* '*Dictionnaire de Kangxi*' [1996], 說文解字注 *Shuōwén jiězì zhù* '*Shuowen jiezi et son interprétation*' [2005], 漢語大詞典 *Hànyǔ dàcídiǎn* '*Grand dictionnaire chinois*' [1994], 辭海 *Cíhǎi* '*Lexique chinois*' [1986] (éditions taïwanaises) et 现代汉语规范字典 *Xiàndài hànyǔ guīfàn zìdiǎn* '*Dictionnaire de normalisation du chinois moderne*' [1998].

Abréviations et symboles

Catégories	Abréviations
Mots	
Nom	N
Nom d'entreprises, d'organismes, etc.	NT
Nom propre	NZ
Nom de lieu	S
Locatif	F
Nom de temps	T
Adjectif à valeur distinctive	B
Numéral	M
Quantifieur	Q
Pronom (ou démonstratif)	R
Adjectif à valeur descriptive	Z
Verbe	V
Adjectif à valeur qualificative	A
Adverbe	D
Onomatopée	O
Préposition	P
Conjonction	C
Auxiliaire	U
Particule modale	Y
Exclamation	E
Formule quadrisyllabique	I
Locution	L
Abréviation	J
Affixations	
Préfixe	H
Suffixe	K
Semi-préfixe	SH
Semi-suffixe	SK
Morphèmes non autonomes de différentes propriétés sémantiques	
Nom	NG
Locatif	FG
Nom de temps	TG
Adjectif à valeur distinctive	BG
Numéral	MG
Pronom (ou démonstratif)	RG
Adjectif à valeur descriptive	ZG
Verbe	VG
Adjectif à valeur qualificative	AG
Adverbe	DG
Préposition	PG
Conjonction	CG
Particule modale	YG
Exclamation	EG
Caractères utilisés dans la constitution des onomatopées	OG
Caractères servant à transposer les syllabes de mots étrangers	PHON
Caractères utilisés dans la constitution des prénoms	NPRG

L'astérisque (*) : nous l'utilisons pour noter que tel exemple est grammaticalement incorrect.

Résumé

Cette étude présente le développement du module d'analyse automatique du chinois qui permet de reconnaître dans les textes les unités lexicales en chinois moderne puis les groupes nominaux noyaux. Pour atteindre ces deux objectifs principaux, nous devions résoudre les problèmes suivants :

1) identifier les unités lexicales en chinois moderne ;

2) déterminer leurs catégories ;

3) décrire la structure de syntaxe locale et des groupes nominaux noyaux.

C'est ainsi que nous avons été amenée à constituer d'abord un corpus regroupant des textes littéraires et journalistiques publiés au XX^{e} siècle. Ces textes sont écrits en chinois moderne avec des caractères traditionnels. Grâce à ces données textuelles, nous avons pu recueillir des informations linguistiques telles qu'unités lexicales, structures syntagmatiques ou règles grammaticales. Ensuite, nous avons construit des dictionnaires électroniques dans lesquels chaque unité lexicale est représentée par une entrée, à laquelle sont associées des informations linguistiques telles que catégories lexicales, classes de distribution sémantique ou descriptions formelles de certaines formes lexicales. À ce stade, nous avons cherché à identifier les unités lexicales du lexique chinois et leurs catégories en les recensant. Grâce à cette liste, l'analyseur lexical peut traiter des unités lexicales de différents types, en bloc, sans les découper en composants. Ainsi, on traite les unités lexicales suivantes comme des unités atomiques :

理髮 *lǐfà / fǎ* <arranger-cheveux> 'faire la coiffure'
放假 *fàngjià* <distribuer-vacance> 'être en vacances'
刀子口 *dāozikǒu* <couteau-bouche> 'parole cruelle'
研究員 *yánjiū / jiù yuán* <effectuer des recherches-K> 'chercheur'
翻譯系統 *fānyì xìtǒng* <traduire-système> 'système de traduction'
浪漫主義 *làngmàn zhǔyì* <romantique- -isme> 'romantisme'

Puis, nous avons décrit de manière formelle un certain nombre de syntagmes locaux, ainsi que cinq types de groupes nominaux noyaux. Enfin, nous avons utilisé le module chinois ainsi développé pour étudier l'évolution thématique dans les textes littéraires.

Abstract

This study presents the development of a module for the automatic parsing of Chinese that will allow to recognize automatically lexical units in modern Chinese, as well as central Noun Phrases in texts. In order to reach these two principle objectives, we solved the following problems:

1) identify lexical units in modern Chinese;

2) determine their categories;

3) describe certain local syntactic structures as well as the structure of central Noun Phrases.

Firstly we constructed a corpus regrouping literary and journalistic texts published in the XXth century. These texts are written in modern Chinese with traditional characters. Thanks to textual data, we could collect linguistic information such as lexical units, syntagmatic structures or grammatical rules. Then, we constructed several electronic dictionaries in which each entry represents a lexeme, with which is associated linguistic information such as its lexical category, its semantic distributional class or certain formal properties. At this stage, we tried to identify the lexical units of Chinese lexicon and their categories in order to list them. Thanks to this list, an automatic lexical analyzer can process various types of lexical units in bloc, without deconstructing them in components. For instance, the lexical parser processes the following lexical units as atomic units:

理髮 *lǐfà / fǎ* <operate-hair> ‘have a haircut’
放假 *fàngjià* <distribute-vacation> ‘have vacation’
刀子口 *dāozikǒu* <knife-mouth> ‘straight talk’
研究員 *yánjiū / jiù yuán* <research-K> ‘researcher’
翻譯系統 *fānyì xìtǒng* <translate-system> ‘translation system’
浪漫主義 *làngmàn zhǔyì* <romantic- -ism> ‘romanticism’

Then, we described formally certain local syntagms and five types of central Noun Phrases. Finally, we used this Chinese module to study thematic evolution in literary texts.

Introduction

Motivation

Le **T**raitement **A**utomatique des **L**angues **N**aturelles (**TALN**) est un domaine de recherche pluridisciplinaire qui utilise des méthodes élaborées autour de problématiques diverses ainsi que des programmes informatiques visant à modéliser les langues humaines dans tous leurs aspects. Les méthodes sont fondées sur des connaissances situées à la croisée de multiples disciplines : linguistique, logique, informatique, statistique, Intelligence Artificielle, etc. Il s'agit donc à la fois de construire formellement des règles linguistiques et de traiter des données produites naturellement par les êtres humains.

Dans ce domaine de recherche, on doit en premier lieu, identifier les unités lexicales qui constituent un moyen d'accès aux informations. Ce sont des unités porteuses de sens qu'on appelle très souvent « mots ». À part ces « mots », on peut trouver, dans le lexique d'une langue, d'autres unités. Nous en citerons deux exemples. Les « morphèmes », d'abord, qui correspondent aux plus petites unités porteuses de sens et servent à former des « mots » par composition, flexion ou dérivation. Les « expressions figées », ensuite, qui sont constituées de deux ou de plusieurs « mots ». Elles jouissent d'une autonomie syntaxique et s'utilisent donc souvent comme des mots lors de la constitution des phrases, mais certaines d'entre elles peuvent elles-mêmes être des phrases.

La reconnaissance des unités lexicales est une première étape à partir de laquelle on commence à analyser les textes de façon automatique. Cette reconnaissance s'appuie fréquemment sur une application des dictionnaires qui regroupent les unités atomiques d'une langue.

L'objet de notre recherche consiste donc à réaliser un module d'analyse automatique qui pourra traiter des textes en chinois moderne. Ce module contiendra non seulement un corpus, mais aussi des dictionnaires électroniques regroupant le vocabulaire, ainsi que des grammaires précisant les relations qu'entretiennent entre elles les unités lexicales. La formalisation de ces ressources linguistiques est basée naturellement sur le corpus.

Quelques difficultés relatives à la spécificité du lexique chinois

Dès le début de cette recherche sur le Traitement Automatique des Langues Naturelles, nous avons rencontré des problèmes essentiels :

● En chinois, la notion de « mot » est floue. Contrairement à ce qu'on observe dans les langues indo-européennes, il est assez difficile d'identifier les « mots » chinois dans un texte. En d'autres termes, les unités qu'on extrait peuvent correspondre à des morphèmes, à des « mots », à des expressions figées ou encore à des syntagmes libres. Comment distinguer les unités lexicales des syntagmes libres ? Comment les identifier correctement dans les textes ?

● Si la frontière entre unités lexicales et syntagmes libres demeure instable, comment établir une liste d'entrées qui puisse servir à reconnaître et à annoter les unités lexicales de différents types ? Nous devions élaborer d'abord des dictionnaires électroniques répertoriant les unités lexicales présentées dans notre corpus. Restait à savoir quels critères nous prendrions en compte pour définir les unités lexicales et pour les intégrer dans les dictionnaires électroniques.

● Comment définir les catégories des entrées, quand on sait que les formes dérivationnelles et flexionnelles sont pauvres en chinois. Il n'est pas facile de définir les catégories des unités lexicales. De plus, si une unité lexicale a plusieurs fonctions syntaxiques, est-elle une unité unique appartenant à une seule catégorie ou bien une forme graphique, partagée par plusieurs unités ? Ces dernières sont distinguées par des catégories différentes. Quels sont les critères permettant de déterminer s'il s'agit d'une unité unique appartenant à plusieurs catégories ou alors de plusieurs unités dont chacune appartient à une catégorie différente ?

Cette thèse vise donc à résoudre, en un premier temps, les problèmes lexicaux liés à l'identification des unités lexicales dans les textes chinois et à la détermination de leurs catégories. Ces deux phases de traitement constituent la base du traitement automatique des textes.

But de cette recherche

L'intérêt de cette recherche est de développer un module qui permette d'analyser de façon automatique des textes écrits en chinois moderne avec des caractères traditionnels, c'est-à-dire non simplifiés. Nous avons donc entrepris de développer des dictionnaires électroniques et des grammaires expressément conçus en vue de l'analyse automatique de la langue chinoise moderne. Notons que cet ensemble de ressources linguistiques formalisées peut également servir à filtrer des informations, à caractériser des données ou à extraire des données en fonction de leur grammaticalité spécifiée.

L'unité lexicale en chinois est un élément difficile à définir et à identifier. Il sera donc intéressant de résoudre les difficultés posées par la définition même de « mot », comment, par exemple, identifier correctement un mot ou connaître sa catégorie dans son contexte. Une des originalités de cette étude est de proposer une application informatique d'annotation d'unités lexicales sans que leur soient appliqués des prétraitements manuels. Un prétraitement des textes regroupe des opérations bien communes du Traitement Automatique des Langues Naturelles : introduction des séparateurs, insertion des étiquettes de catégories dans les textes analysés, découpage des textes en mots, etc.

De plus, à l'aide de ce module, nous pouvons rechercher des expressions qui nous intéressent dans le corpus. Enfin, nous tentons également d'utiliser ce module pour étudier des thèmes dans la littérature, d'une manière qui nous permettra d'observer leur évolution pendant une période prédéfinie.

Limites de cette étude

Nous ne traiterons pas ici des problèmes rencontrés dans les textes rédigés en chinois simplifié, ni les cas où deux systèmes d'écriture (caractères traditionnels ou caractères simplifiés) cohabitent. Par ailleurs, en développant un module chinois dans ***NooJ***, nous ne traitons pas tous les phénomènes linguistiques. Cette étude présente donc les limites suivantes :

- Nos dictionnaires électroniques ne contiennent que les unités lexicales trouvées dans le corpus ainsi que leurs usages et les informations linguistiques les concernant [*cf.* Chapitre 5]. Au besoin, des unités manquantes pourront être ajoutées dans ces dictionnaires.

● Nous n'avons pas abordé la syntaxe chinoise dans toutes ses implications, mais seulement ceux de ses éléments qui permettent de lever des ambiguïtés et de répertorier les contraintes qui président à la combinaison des unités lexicales. Parmi les structures syntaxiques, citons celles étudiées en détail au chapitre 7, à savoir : la combinaison numérique, les expressions de temps ou les appellations personnelles.

● Nos grammaires cherchent à décrire de façon formelle la structure des groupes nominaux noyaux en chinois. Ces groupes sont, pour nous, au nombre de cinq. Nous n'accorderons aucune place à la description d'autres structures telles que les groupes verbaux, adjectivaux, adverbiaux, etc.

Plan de la thèse

Cette étude se compose de deux parties et de huit chapitres.

1) Dans la première partie, nous ferons une présentation générale de la langue chinoise moderne. Puis, nous exposerons les théories et les méthodes utilisées dans le domaine du Traitement Automatique des Langues Naturelles. Finalement, nous discuterons des problèmes liés à la construction du corpus, les problèmes rencontrés dans le traitement automatique du chinois et les solutions que nous y avons apportées.

● Chapitre I : Nous présenterons d'abord l'évolution du chinois moderne, les relations entre le son, le sens et la forme graphique ainsi que les éléments phonologiques. Nous décrirons brièvement le lexique, les catégories de mots et les syntagmes. Enfin, nous présenterons le statut de caractères chinois.

● Chapitre II : Nous aborderons le Traitement Automatique des Langues Naturelles. Nous verrons ce qu'ont été les premières théories dans ce domaine, les grammaires formelles, par exemple, et la hiérarchie de Chomsky. Ensuite, nous présenterons trois modèles représentatifs des grammaires d'unification, qui servent à analyser la syntaxe de la langue. Enfin, nous présenterons ***NooJ***, l'outil que nous avons utilisé dans cette recherche.

● Chapitre III : Nous présenterons l'élaboration du corpus et les problèmes qui se posent lorsqu'on tente d'effectuer une analyse automatique, notamment les ambiguïtés dues aux différents niveaux de langue. Les exemples cités dans ce chapitre ressortent du chinois moderne. Ceci nous permettra de résoudre ou de contourner les difficultés rencontrées lors de l'élaboration du module chinois dans

NooJ. Finalement, nous montrerons en détail la désambiguïsation aux niveaux morphologiques et syntaxiques. La désambiguïsation morphologique consiste à déterminer les unités lexicales. La désambiguïsation syntaxique met les unités lexicales en relation avec la structure syntaxique.

2) Dans la deuxième partie, nous décrirons l'élaboration du module chinois dans ***NooJ***. Nous expliquerons comment ont été conçus le corpus et les dictionnaires électroniques, quel traitement ont reçu les différentes formes graphiques d'un même morphème, comment ont été formalisées les règles grammaticales. Enfin nous verrons quelles sont les modalités d'application de ce module.

● Chapitre IV : Nous décrirons la construction du corpus. Après un exposé des divers systèmes de codage appliqués au chinois, nous décrirons les procédures de collecte, de vérification et de correction des textes sélectionnés en justifiant les choix et les décisions que nous avons dû prendre pour résoudre le problème des caractères réfractaires aux systèmes de codage. Finalement, nous présenterons le formatage de notre corpus.

● Chapitre V : Nous traiterons de la construction des six dictionnaires électroniques que nous avons construits à partir du corpus. Nous étudierons tout d'abord la structure interne des mots. Puis, nous préciserons la définition des Unités Linguistiques Atomiques de ***NooJ***, la formalisation des unités lexicales et les critères utilisés pour regrouper les entrées de six dictionnaires construits. Ensuite, nous examinerons les catégories lexicales et les critères qui permettent de les définir. Nous présenterons aussi les informations linguistiques associées à chaque entrée dans le dictionnaire ***ChDic*** (Chinese Dictionary). Ensuite, nous décrirons les cinq autres dictionnaires ***DicBkTitl*** (Dictionary of Book Titles), ***DicChSurn*** (Dictionary of Chinese Surnames), ***DicExpres*** (Dictionary of Expressions), ***GeoDic*** (Geographical Dictionary) et ***DicProNam*** (Dictionary of Proper Names) et les informations qu'ils contiennent. Ces informations, qui se retrouvent dans les grammaires, sont utilisées pour effectuer des requêtes dans notre corpus. Enfin, nous présenterons la production automatique des réduplications à l'aide de ***NooJ***.

● Chapitre VI : Nous présenterons tout d'abord les problèmes liés à la non-correspondance entre les morphèmes et leurs formes graphiques. Nous montrerons ensuite comment choisir des formes graphiques standard parmi toutes les variantes en nous appuyant sur trois critères, tels que définis et publiés par le

gouvernement chinois. La standardisation permettra de relier à chaque forme standard ses variantes graphiques. Elle permettra également à ***NooJ*** de reconnaître et de traiter les variantes de la même façon que la forme standard. Il en résulte que les variantes et leurs formes standard sont traitées comme une seule et unique unité.

● Chapitre VII : Nous présenterons tout d'abord les six principales structures syntagmatiques du chinois moderne. Nous décrirons ensuite deux types de grammaires : les grammaires locales et les grammaires syntaxiques. Les grammaires locales servent à décrire des structures locales telles que la combinaison des chiffres, la composition des expressions temporelles et des appellations personnelles. Les grammaires syntaxiques servent à formaliser cinq types de groupes nominaux noyaux. Enfin, nous procèderons à une évaluation des grammaires syntaxiques en les appliquant au roman de Lao She *Quatre générations sous un même toit*.

● Chapitre VIII : Nous nous intéresserons à une application de nos ressources linguistiques : en nous basant sur l'analyse des textes littéraires de notre corpus, nous y étudierons l'évolution des thèmes au cours du XX^{e} siècle.

Pour conclure, nous ferons un bilan général des résultats de notre étude et un exposé des solutions apportées aux problèmes posés par le traitement automatique du chinois moderne. Enfin, nous dégagerons quelques pistes que nous pourrions exploiter lors de futures recherches :

■ Amélioration de la précision dans la segmentation des séquences de mots, ce qui permettrait de lever plus d'ambiguïtés avant le lancement d'une requête sur le corpus ;

■ Inclusion dans les grammaires syntaxiques de groupes nominaux plus complexes, comme ceux qui sont composés de plusieurs groupes nominaux de différentes structures et ceux qui comportent des groupes verbaux ;

■ Analyse thématique à partir de termes ou, même, de groupes de mots représentatifs.

Première partie : Introduction générale au chinois moderne, au Traitement Automatique des Langues Naturelles et à son application au chinois

Chapitre 1

INTRODUCTION AU CHINOIS MODERNE

Dans ce chapitre, nous proposons un regard rapide sur le chinois moderne. Nous distinguerons d'abord le chinois classique de la langue vernaculaire, ainsi que deux variétés du chinois moderne. Puis, nous examinerons les signes linguistiques en étudiant la relation entre le son, le sens et la forme graphique. Ensuite, nous nous intéresserons à la phonologie, à la morphologie, à la catégorie lexicale et aux syntagmes en chinois moderne. Enfin, nous préciserons ce que représentent les caractères dans le lexique chinois.

1.1 Du *wényán* au *báihuà*

La langue employée dans des textes chinois anciens comme *les Entretiens de Confucius*, par exemple, est une langue nommée 文言 *wényán*, autrement dit, le chinois classique. Son lexique, sa syntaxe et sa locution sont bien éloignés de la langue utilisée aujourd'hui.

Différent du chinois classique, le 白話 *báihuà* est une langue vernaculaire. Elle suit l'évolution de la langue orale parlée à chaque époque. Certaines œuvres littéraires sont écrites dans cette langue vernaculaire.

Il faut attendre les changements sociaux de 1919, connus sous le nom de Mouvement du 4 mai, pour que le *wényán* cède la place, dans tous les domaines écrits, au *báihuà*. Dès lors, on écrira les textes, qu'ils soient littéraires, scientifiques ou journalistiques, en *báihuà*.

On doit distinguer le « *báihuà* ancien » utilisé avant 1919, du « *báihuà* moderne » utilisé à partir de cette date. Le second constitue un héritage du premier, puisque tous deux sont fondés sur le chinois mandarin, langue commune nationale [*cf.* Zhitang Yang-Drocourt, 2007 : 62-68].

1.2 Variétés du chinois moderne

Depuis 1949, du fait de la fermeture de la Chine continentale, le chinois moderne a évolué de façon différente en Chine et à Taïwan, et on distingue deux variétés de chinois moderne : celui de Taïwan et celui de Chine continentale. La différence se traduit

notamment dans le lexique et la prononciation, alors que les syntaxes restent tout de même très proches. Citons quelques termes différents selon la zone géographique :

Termes utilisés à Taïwan	Termes utilisés en Chine	Significations
資訊 *zī xùn* <valeurs-avis>	信息 *xìnxī / xí* <message-respiration>	information
計程車 *jìchéngchē* <compter le trajet-véhicule>	出租汽車 *chūzū qìchē* <loué-voiture>	taxi
部落格 *bùluògé*	博客 *bókè*	blog
軟體 *ruǎntǐ* <souple-forme>	软件 *ruǎnjiàn* <souple-pièce>	logiciel
蕃茄 *fānqié* <tout ce qui n'est pas chinois-aubergine>	西红柿 *xīhóngshì* <ouest-rouge-kaki>	tomate

En dépit des différences de langue, un Taïwanais et un Chinois continental peuvent communiquer sans aucun problème. La langue qu'ils utilisent est commune : c'est dans les deux cas, le chinois moderne, malgré les divergences évoquées ci-dessus.

1.3 Les éléments phonologiques

1.3.1 La structure syllabique

En chinois moderne, il y a vingt et une initiales qui correspondent à toutes les consonnes, sauf *–ng* qui ne peut pas être initiale. Deux consonnes ne se suivent jamais.

Il y a, par ailleurs, trente-neuf finales. Elle peut être composée de façon complexe. Ainsi, les diphtongues réunissant deux voyelles, ou les triphtongues qui en comportent trois, etc.

Les vingt et une initiales et les trente-neuf finales se combinent pour former environ quatre cents combinaisons quasi-syllabiques valides. Le nombre de ces combinaisons varie, selon que l'on compte ou non les interjections consonantiques et certaines formes régionales. On peut combiner une initiale à une finale pour donner un son. Néanmoins, ce son n'existe pas forcément en chinois moderne. C'est la raison pour laquelle il n'existe qu'un petit nombre de combinaisons syllabiques valides.

Le son produit par la combinaison d'une initiale et d'une finale ne permet pas à lui seul d'identifier un morphème chinois. Il faut ajouter un ton à chaque son pour obtenir une forme sonore complète. Il existe quatre tons. Les quatre cents quasi-syllabes, à leur tour, se combinent avec ces quatre tons pour former approximativement mille deux cent cinquante syllabes attestées. Mille deux cent cinquante et non pas mille six cents, comme un rapide

calcul pourrait le faire croire car si, théoriquement, on peut doter une syllabe de quatre tons, certaines combinaisons n'existent pas dans la réalité. Par exemple, on peut ajouter les quatre tons à la forme *ban* : *bān*, *bán*, *bǎn* et *bàn*, mais, la deuxième forme, *bán*, ne correspond à aucun morphème chinois. Notons également que le suffixe 兒 *r* a une prononciation plus courte qu'une syllabe.

En conclusion, reposant sur un système de tonalités, une syllabe chinoise correspond à une unité phonologique unique et insécable, qui ne peut être liée à d'autres, ni insérée entre d'autres. Ainsi, la forme orale se base sur ces mille deux cent cinquante syllabes dont on se sert pour s'exprimer en chinois moderne.

1.3.2 Le syllabisme du chinois moderne

La plus petite unité porteuse de sens est appelée morphème [*cf.* 5.1.1]. Cette unité morphologique peut être représentée sous forme vocale par une seule syllabe et sous forme graphique par un seul carctère. Exemple :

(1) 她時常跳舞。 *tā shícháng tiàowǔ.* <elle-souvent-danser.> 'Elle danse souvent.'

Niveau	**Découpage**				
Syllabes	*tā*	*shí*	*cháng*	*tiào*	*wǔ*
Morphèmes en caractères	她	時	常	跳	舞
Traduction française	elle	temps	souvent	sauter	danse
Mots en caractères	她	時常		跳舞	
Traduction française	elle	souvent		danse	

Tableau 1 : Rapport entre syllabe, morphème et mot en chinois

En tant que syllabe, un son correspond à une unité morphologique. On utilise cette dernière pour former des mots de longueurs différentes. *tā* 她 est un mot monosyllabique ; *shícháng* 時常 et *tiàowǔ* 跳舞 sont des mots dissyllabiques. La phrase donnée montre qu'un morphème chinois est monosyllabique, et que le mot fondé sur le morphème est souvent dissyllabique.

Dans la plupart des cas, un morphème chinois est monosyllabique. Les mots chinois, constitués à partir de morphèmes monosyllabiques, sont généralement dissyllabiques et ce sont eux qui constituent l'essentiel du vocabulaire chinois. Un mot chinois peut comprendre plusieurs syllabes. Néanmoins, du point de vue morphologique, la langue

chinoise peut être considérée comme étant une langue monosyllabique [*cf*. Zhitang Yang-Drocourt, 2007 : 155-156, 213-214].

1.4 L'introduction sur le son, le sens et la forme graphique

Comme nous l'avons étudié dans la section précédente, une syllabe chinoise peut représenter un morphème. Ce dernier s'écrit ave un caractère chinois. Cependant, il n'y a pas de rapport direct entre la syllabe et le caractère. C'est le sens qui permet de construire ce lien. Par conséquent, un caractère chinois est directement lié au sens qui, à son tour, permet de connaître la syllabe correspondante. De ce point de vue, les caractères chinois dénotent non seulement les sens, mais aussi la prononciation de ces sens.

En chinois moderne, il n'existe qu'environ mille deux cent cinquante syllabes. Néanmoins, il compte plus de dix mille morphèmes. Cela signifie que les homophones sont les cas fréquents. Par exemple, la syllabe pái peut se rapporter à sept sens au moins, chacun d'eux s'écrivant avec une graphie particulière :

(2) 俳 *pái* 'attraction théâtrale'
排 *pái* 'aligner'
棑 *pái* 'radeau en bambou'
牌 *pái* 'tableau d'affichage'
狾 *pái* 'chien à poils courts'
簰 *pái* 'grand radeau en bambou'
輫 *pái* 'caisse d'une voiture'

1.5 Le lexique chinois

1.5.1 Concept de "mot" en chinois

Le terme 詞 *cí* qui renvoie à la notion de "mot", n'est apparu en chinois qu'au début du XX^{e} siècle. La langue chinoise commençait alors à être analysée selon les théories linguistiques proposées pour étudier les langues indo-européennes [*cf*. Ma Jianzhong, 1898]. Dès lors, certains linguistes chinois ont cherché à définir le "mot" selon des critères phonologiques, sémantiques ou syntaxiques, c'est ce que firent Li Jinxi [1924], Wang Li [1943], Fu Huaiqing [1985], Ge Benyi [2001], etc. En synthétisant les propositions

émanant de divers confrères, Cao Wei [2003] a donné une définition du mot chinois dans son ouvrage *Recherches sur le vocabulaire en chinois moderne* [2004 : 1-3] :

> *Un mot est une unité phono-sémantique la plus petite. Cette unité contient une forme phonétique relavtivement stable et une convenable longueur lexicale. Elle peut donc être utilisée librement.*[2]

De ce point de vue, le "mot" chinois serait une unité indépendante associée à une prononciation et à un sens particulier. Néanmoins, les théories proposées par les linguistes sont souvent difficilement applicables car, en chinois, la frontière entre un mot composé et un syntagme libre n'est pas facile à définir. Or la distinction entre mots composés et syntagmes libres n'est parfois pas perceptible par les personnes dont le chinois n'est pas la langue maternelle [*cf.* Sun Chaofen, 2006 : 46].

1.5.2 Bref aperçu sur la structure interne des mots chinois

1.5.2.1 Le statut morphologique

À l'oral, une syllabe chinoise constitue une unité indépendante et unique, par exemple, *tā* 'il'. Elle se représente, à l'écrit, par un caractère tel que 典 *diǎn* 'citation'. On peut composer des unités plus complexes à partir de ces unités simples :

1) à l'oral : *xīwàng* <désirer-espérer> 'espérer', *qiǎokèlì* 'chocolat' ou *xīlì-huālā* 'onomatopée du bruit de la pluie qui tombe sur le sol' ;
2) à l'écrit : 彙集 *huìjí* <classer-réunir> 'collectionner', 芭蕾舞 *bāléiwǔ* <ballet-danse> 'ballet' ou 嘰哩咕嚕 *jīlī-gūlū* 'murmurer'.

À première vue, on ne peut pas savoir si ces unités complexes correspondent à des morphèmes, à des « mots », à des syntagmes ou bien à des phrases. L'indépendance de leurs composants (syllabiques ou graphiques) ne permet pas de déterminer leur statut morphologique. De ce point de vue, on ne peut pas identifier directement à quel type d'unité lexicale appartiennent ces unités qui possèdent plus d'une syllabe ou s'écrivent avec plus d'un caractère.

[2] 词是最小的有相对固定的语音形式和适度词长的能独立运用的语音单位。
cí shì zuì xiǎo de yǒu xiāngduì gùdìng de yǔyīn xíngshì hé / hàn shìdù cícháng de néng dúlì yùnyòng de yǔyīn dānwèi.

1.5.2.2 La liberté syntaxique

Comme nous l'avons dit dans la section précédente, les composants des unités lexicales ont une liberté syllabique ou graphique. Mais, ils ne possèdent pas la même liberté syntaxique. On constate que l'unité (3a) a la même liberté syntaxique que l'unité (4a), bien que cette dernière soit plus longue qu'une syllabe. De plus, ces deux unités peuvent entrer dans la composition d'autres unités plus complexes comme (3b, 3c, 4b et 4c).

(2) a. 花 *huā* 'fleur'
b. 花瓶 *huāpíng* <fleur-vase> 'vase à fleurs'
c. 白花 *báihuā* <blanc-fleur> 'fleur blanche'

(3) a. 葡萄 *pútáo* 'raisin'
b. 葡萄酒 *pútáojiǔ* <raisin-vin>'vin de raisin'
c. 酸葡萄 *suānpútáo* <acide-raisin> 'raisin acide : jaloux'

Cependant, dans le lexique chinois, il existe aussi des composants morphologiques qui ne possèdent pas de liberté syntaxique. Néanmoins, ces composants ont aussi une liberté syllabique ou graphique tout autant qu'un sens. C'est le cas de 希 *xī* 'désirer', 典 *diǎn* 'citation', 紀 *jì* 'règle', 語 *yǔ* 'parler' ou 彙 *huì* 'classe'. Ils doivent se combiner avec autres composants pour former des unités lexicales qui servent à constituer des phrases. Mentionnons quelques exemples comportant ces composants : 希望 *xīwàng* <désirer-espérer> 'espérer', 典禮 *diǎnlǐ* <code-cérémonie> 'cérémonie', 紀錄 *jìlù* <règle-registrer> 'consigner par écrit', 語言 *yǔyán* <expression-parole> 'langue' ou 彙集 *huìjí* <classer-réunir> 'collectionner'.

1.6 Les catégories de mots en chinois

Il existe des cas en chinois moderne où un même « mot » semble être utilisé tantôt comme un adjectif, tantôt comme un nom. C'est le cas de 耐心 *nàixīn*[3]. Néanmoins, cela ne signifie pas que tout adjectif puisse devenir un nom et inversement. Il existe d'ailleurs d'autres « mots » qui, de la même manière, peuvent être utilisés tantôt comme verbes, tantôt comme noms. Ce sont, à l'origine, des verbes qui ont syntaxiquement une fonction nominale. Ils n'ont donc pas à deux catégories différentes. Ils se définissent comme des

[3] Ce « mot » signifie 'patient' quand il est employé comme adjectif. Employé comme nom, il signifie 'patience'.

verbes nominalisés [*cf.* Zhu Dexi, 1982 [2004 : 60-61] et Guo Rui, 2002 [2004 : 187]]. C'est le cas de 管理 *guǎnlǐ* 'gérer', 解决 *jiějué* 'résoudre' ou 調查 *diàochá* 'enquêter'.

Aujourd'hui, on discute encore sur le fait de savoir si un « mot » appartient à une ou à plusieurs catégories. Bien que ce type de difficultés ne soit pas encore résolu, on tente de s'accorder sur une définition des catégories et sur les critères qui permettent leur identification.

1.7 Syntagmes chinois

Un syntagme est un élément linguistique qui se situe à mi-chemin entre le mot et la phrase. Il est constitué d'au moins deux mots. Dans un syntagme chinois, les catégories des mots et leur ordre doivent respecter les règles grammaticales. Examinons les exemples suivants :

(4) a. 漂亮的眼睛 *piàoliàng de yǎnjing / jīng* <beau-De-œil> 'yeux beaux'
b. *因為的眼睛 *yīnweì de yǎnjing/ jīng* <car-De-œil>

(5) a. 參觀巴黎 *cānguān Bālí* <visiter-Paris> 'visiter Paris'
b. *巴黎參觀 *Bālí cānguān* <Paris-visiter>
c. *參觀公升 *cānguān gōngshēng* <visiter-litre>

Puisque la conjonction n'est pas utilisée avec le subordinateur nominal 的 *de* pour qualifier ou spécifier le noyau nominal qui le suit, le syntagme (5a) est accepté en chinois. Mais l'exemple (5b) n'est pas un syntagme correct.

Un complément d'objet ne peut pas précéder le verbe. Ainsi, le syntagme (6a) est pertinent en chinois, contrairement au syntagme (6b). Par ailleurs, un complément d'objet ne pouvant pas être assumé par un quantifieur [*cf.* 5.6.3 et 7.1.2], le syntagme (6c) n'obéit pas à la règle grammaticale.

On remarque d'abord que les composants ne sont pas les mêmes suivant les types de syntagmes et que chacun a sa propre structure syntaxique. Autrement dit, chaque composant doit appartenir à une catégorie définie qu'on ne peut pas changer. De plus, pour constituer un syntagme chinois, il est obligatoire d'en respecter la structure.

1.8 Le statut du caractère chinois

Tout caractère ne correspond pas forcément à un mot simple. Nous examinerons en détail, dans la section 5.1, la structure interne des mots et les différents types d'unités lexicales que nous décrivons ci-dessous, à savoir :

1) les caractères non signifiants utilisés pour former des polysyllabes monomorphémiques. Ces caractères n'ont pas de signification. Ils ne constituent donc pas des composants morphologiques de mots. Par exemple, les caractères 葡 *pú*, 萄 *táo*, 蝴 *hú*, 蝶 *dié*, 駱 *luò*, 駝 *túo*, 區 *qū*, 咖 *kā*, 啡 *fēi*, 噗 *pū* ou 通 *tōng* ne sont utilisés que pour leur seule valeur phonétique. Ils sont des composants syllabiques des mots 葡萄 *pútáo* 'raisin', 蝴蝶 *húdié* 'papillon', 駱駝 *luòtúo* 'chameau', 區區 *qūqū* 'minuscule', 咖啡 *kāfēi* 'café' et 噗通 *pūtōng* 'plouf se référant au bruit d'un objet qui tombe dans l'eau'.

On rencontre aussi d'autres cas où les caractères perdent leur sens d'origine. En d'autres termes, il existe une rupture entre le son et le sens. C'est le cas de 巧 *qiǎo*, 克 *kè*, 力 *lì*, 歇 *xiē*, 斯 *sī*, 底 *dǐ* et 里 *lǐ*. Bien que ces caractères représentent graphiquement des morphèmes, ils n'expriment pas de sens, lors de la formation de polysyllabes tels que :

(6) 巧克力 *qiǎokèlì* 'chocolat' ou
歇斯底里 *xiēsīdǐlǐ* 'hystérie'.

En effet, le caractère 力 *lì* 'insister sur' est un composant morphologique du mot 力圖 *lìtú* <insister sur-projeter> 'tendre à'. Le caractère 歇 *xiē* 'se reposer' est un des composants du verbe 歇息 *xiēxí* <se reposer-rester au repos> 'se reposer'. Ainsi, ces deux caractères 力 *lì* et 歇 *xiē* sont à la fois des graphies de transposition phonologique de mots étrangers et des graphies de morphèmes chinois. Ils doivent être traités comme des entrées différentes dans un dictionnaire : caractères non signifiants et morphèmes.

2) des morphèmes non autonomes comme 希 *xī* 'désirer', 典 *diǎn* 'citation', 紀 *jì* 'règle', 語 *yǔ* 'parler' et 彙 *huì* 'classe'. Ils ne sont autonomes que sémantiquement, graphiquement et syllabiquement, mais ils ne correspondent pas à des mots simples, parce qu'ils doivent toujours être utilisés en combinaison avec un autre morphème

pour former un mot. Par exemple, 希望 *xīwàng* <désirer-espérer> ‘espérer’, 典禮 *diǎnlǐ* <code-cérémonie> ‘cérémonie’, 紀錄 *jìlù* <régle-registrer> ‘consigner par écrit’, 語言 *yǔyán* <expression-parole> ‘langue’ ou 彙集 *huìjí* <classer-réunir> ‘collectionner’. En d’autres termes, les morphèmes non autonomes s’écrivant avec des caractères ne sont que des composants de mots et non pas des mots simples.

3) des mots simples monosyllabiques. Ils s’écrivent avec des caractères uniques. Tel est le cas de 他 *tā* ‘il’, 海 *hǎi* ‘mer’, 玩 *wán* ‘jouer’, 看 *kàn* ‘regarder’ ou 筆 *bǐ* ‘stylo’.

4) des caractères utilisés pour former des prénoms chinois, par exemple, 佑 *yòu*, 宜 *yí*, 玫 *méi*, 漢 *hàn*, 曉 *xiǎo*, ou 蘭 *lán*.

5) des caractères qui sont des graphies de noms de famille chinois. Mentionnons quelques exemples : 何 *Hé*, 林 *Lín*, 高 *Gāo*, 陳 *Chén*, 司馬 *Sīmǎ*, 歐陽 *Ōuyáng*, etc.

Chapitre 2

INTRODUCTION AU TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES: METHODES ET MODELES INFORMATIQUES

La modélisation d'une langue s'exerce sur les différents domaines linguistiques traditionnels que sont la phonologie, la morphologie, la syntaxe, la sémantique ou la pragmatique. Elle entraîne le développement de modèles grammaticaux aptes à traiter des productions linguistiques à différents niveaux de la langue. Ce développement des modèles grammaticaux s'inspire de l'étude des grammaires formelles faite par Noam Chomsky. Il existe, actuellement, trois modèles grammaticaux représentatifs de la famille des grammaires d'unification :

1) Grammaire lexicale fonctionnelle (***L**exical **F**unctional **G**rammar, **LFG***) ;

2) Grammaire syntagmatique guidée par les têtes (***H**ead-Driven **P**hrase **S**tructure **G**rammar, **HPSG***) ;

3) Grammaire d'arbres adjoints (***T**ree **A**djoining **G**rammar, **TAG***).

NooJ, nouveau système informatique créé en 2002, est un outil permettant de procéder à différentes applications du Traitement Automatique des Langues Naturelles telles que l'extraction des informations et la traduction automatique.

Nous présenterons ci-dessous les théories principales qui sont à l'origine des progrès effectués dans ce domaine. Puis nous décrirons les trois modèles grammaticaux ci-dessus mentionnés. Nous présenterons enfin le nouvel analyseur, ***NooJ***, utilisé comme outil de notre recherche.

2.1 Grammaires formelles

Une grammaire formelle décrit un ensemble de séquences de symboles sous forme de règles de réécriture, nommées aussi règles de production. Les règles de réécriture se présentent suivant le schéma : $\alpha \rightarrow \beta$. Elles s'appliquent à des constructions syntaxiques incluant le membre gauche **α**. Lorsqu'on remplace le membre gauche par le membre droit, l'objet **α** est transformé en **β**.

Une grammaire formelle est un quadruplet {N, Σ, P, S} dans lequel :

1) **N** est un ensemble fini non vide de symboles non-terminaux notés par des majuscules ;

2) **Σ** est un ensemble fini non vide de symboles terminaux dont les éléments sont notés par des minuscules. **N** et **Σ** sont disjoints ;

3) **P** est un ensemble fini non vide de règles de réécriture : P = {α→β} dans lesquelles **α** et **β** sont des suites de terminaux et de non-terminaux. Chaque règle de réécriture est représentée par une formulation, par exemple, P → aCb ;

4) **S** est un symbole non-terminal de départ qui se nomme axiome.

Une grammaire formelle consiste ainsi à engendrer l'ensemble de toutes les séquences de symboles dans un langage par une dérivation commencée par un axiome. Prenons un exemple, dans lequel un langage **L** est défini par la grammaire formelle suivante :

1) **N** représente l'ensemble de ses symboles non-terminaux : {S, C} ;

2) l'ensemble de ses terminaux : {a, b, e, f} ;

3) **S** le symbole de départ ;

4) **P** les trois règles de réécriture :

$$\left\{\begin{array}{l} 1.\ S \rightarrow C \\ 2.\ C \rightarrow aCb \\ 3.\ C \rightarrow ef \end{array}\right\}$$

Nous commençons par la règle 1. S → C et nous remplaçons **C** du membre droit par la règle 2, nous aurons le symbole **aCb**. Le processus peut se répéter jusqu'à ce que toutes les occurrences de **C** soient éliminées et que les symboles des alphabets **e** et **f** se présentent. Cette application répétitive de ces trois règles de réécriture peut être représentée par le schéma simplifié : C → aCb → aaCbb → aaefbb. On constate à travers cette grammaire que le langage **L** possède des séquences de symboles telles que : {ef, aefb, aaefbb, …}.

2.2 La hiérarchie de Chomsky

La hiérarchie de Chomsky permet de classer les différents niveaux de langages décrits par les différentes grammaires sous forme de règles de réécriture. Un langage est généré par

une grammaire formelle à l'aide de laquelle il est défini. On distingue quatre grammaires dont chacune représente un langage. Leur relation peut être décrite comme suit :

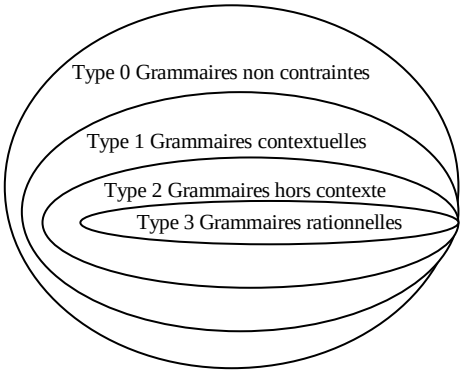


Figure 1 : La hiérarchie de Chomsky

Les langages de type 3 sont des langages rationnels. Ils sont engendrés par des grammaires linéaires gauches ou droites :

1) Lorsque la grammaire est linéaire gauche, ses règles se présentent sous la forme :
   $A \rightarrow Bc$
   $A \rightarrow c$
   $A,B \in N, a \in \Sigma$

2) Lorsque la grammaire est linéaire droite, ses règles se présentent sous la forme :
   $A \rightarrow cB$
   $A \rightarrow c$
   $A,B \in N, a \in \Sigma$

L'ensemble des grammaires linéaires gauches ou droites correspond à l'ensemble des grammaires rationnelles.

Les langages de type 2 sont définis par des grammaires hors contexte, aussi appelées grammaires algébriques ou grammaires de Chomsky. Chaque règle de réécriture se forme de la manière suivante : $A \rightarrow \alpha$, où **A** est un symbole non-terminal, et **α** représente une chaîne composée de symboles terminaux et / ou de symboles non-terminaux.

Les langages de type 1 se définissent par des grammaires contextuelles qui contiennent des règles de réécriture de type : $\alpha A \beta \rightarrow \alpha C \beta$. **α** et **β** construisent un contexte dans lequel le symbole non-terminal **A** est remplacé par la séquence non vide **C**.

Les langages de type 0 sont récursivement énumérables. Un langage est récursivement énumérable à la seule condition qu'il puisse être produit par une machine de Turing. Ces

langages de type 0 sont engendrés par des grammaires non contraintes, par exemple, $\alpha \rightarrow \beta$. **α** et **β** peuvent être des séquences de terminaux et de non-terminaux.

2.3 Les trois grammaires d'unification

2.3.1 La grammaire lexicale fonctionnelle

La grammaire lexicale fonctionnelle (***L**exical **F**unctional **G**rammar, **LFG***) a été développée par Joan Bresnan et Ronald Kaplan pendant les années 70. La structure syntaxique d'une phrase est décrite à la fois par une représentation arborescente et par une structure de traits contenant les informations linguistiques. Le modèle ***LFG*** propose deux outils pour décrire la structure d'une phrase : la structure de constituants (structure c) et la structure fonctionnelle (structure f). La structure de constituants sert à décrire l'ordre des mots, leurs relations de domination et leurs catégories syntaxiques. La structure fonctionnelle sert à indiquer les fonctions grammaticales, les informations morphologiques et les relations morpho-syntaxiques relatives aux mots de la phrase concernée. Cette structure est représentée par un ensemble de traits qui encadrent un attribut et une valeur [*cf.* J. Bresnan et R. Kaplan, 1982 : 173-281].

2.3.2 La grammaire syntagmatique guidée par les têtes

La grammaire syntagmatique guidée par les têtes (***H**ead-driven **P**hrase **S**tructure **G**rammar, **HPSG***) a été inventée par Carl Pollard et Ivan Sag au début des années 80. Ce modèle ***HPSG*** s'inspire principalement de la grammaire syntagmatique généralisée (***G**eneralized **P**hrase **S**tructure **G**rammar, **GPSG***). Il procède du principe d'unification et décrit la structure de phrase à un seul niveau.

Le modèle ***HPSG*** a pour objectif de regrouper, dans un modèle formel, les informations appartenant à différents niveaux linguistiques, par exemple, phonologie, syntaxe, sémantique, etc. Ces informations sont présentées par des traits qui héritent des spécifications du trait supérieur. Ce modèle n'emploie pas d'opérations transformationnelles et se caractérise par une architecture non-dérivationnelle. Il utilise les contraintes déclaratives pour noter les informations syntagmatiques [*cf.* A. Abeillé, 2007 : 115-192].

2.3.3 La grammaire d'arbres adjoints

La grammaire d'arbres adjoints (***T**ree **A**djoining **G**rammar*, ***TAG***), mise au point par Aravind Joshi et Levy Takahashi en 1975 est un formalisme génératif de réécriture qui tente de décrire la syntaxe des langues naturelles. Ce formalisme fonde la description des syntagmes sur des structures représentées par des arbres élémentaires, qui sont de deux types : les arbres initiaux et les arbres auxiliaires. La combinaison d'arbres élémentaires est mise en œuvre à travers deux opérations : la substitution et l'adjonction. L'opération de substitution s'effectue sur les arbres initiaux. L'opération d'adjonction est réservée aux arbres auxiliaires. Ce formalisme consiste à modeler la structure syntaxique des phrases en gérant des contraintes syntagmatiques par une méthode lexicalisée [*cf.* A. Abeillé, 2007 : 193-262].

2.4 *NooJ*

2.4.1 Introduction

Max Silberztein a conçu ***NooJ*** en 2002 s'appuyant sur son expérience de développement d'***INTEX***. Il écrivit ce système à partir de zéro, mais en utilisant les techniques de la **P**rogrammation **O**rientée **C**omposants (**POC**), évolution de la Programmation Orientée Objet. Il implémente ce système en s'appuyant sur le *Framework .NET*.

NooJ permet de décrire des langues en utilisant des dictionnaires électroniques et des grammaires formelles. Les dictionnaires électroniques permettent de décrire quatre types d'Unités Linguistiques Atomiques (***A**tomic **L**inguistic **U**nits, **ALU**s*) : affixes, mots simples, mots composés et expressions figées, que nous détaillerons dans la section 5.2. ***NooJ*** fournit des outils pour décrire la flexion et la dérivation dans ces dictionnaires. Les descriptions grammaticales concernent les niveaux orthographiques, morphologiques, syntaxiques et sémantiques. Elles se présentent sous forme de règles de réécriture semblables aux grammaires présentées en 2.1, ou de graphes que nous présenterons en 2.4.4.

Le système ***NooJ*** est destiné à formaliser les langues naturelles, mais il sert également à traiter des corpus écrits de grande taille. Les résultats de ces analyses textuelles sont

typiquement représentés par des annotations morpho-syntaxiques, des concordances lemmatisées, et peuvent être utilisées pour effectuer des analyses statistiques.

2.4.2 Méthodologie

L'analyse automatique des textes débute par une reconnaissance des Unités Linguistiques Atomiques. Pour déterminer les unités morphologiques ou syntagmatiques, ***NooJ*** emploie trois outils informatiques :

1) Automates à états finis (***F**inite-**S**tate **A**utomata*, ***FSA***) ;

2) Transducteurs à états finis (***F**inite-**S**tate **T**ransducers*, ***FST***) ;

3) Réseaux de Transitions Récursifs (***R**ecursive **T**ransition **N**etworks*, ***RTN**s*).

2.4.2.1 Automates à états finis (*Finite-State Automata*, *FSA*)

Dans ***NooJ***, les automates à états finis (***F**inite-**S**tate **A**utomata*, ***FSA***) constituent un type particulier de transducteurs à états finis (***F**inite-**S**tate **T**ransducers*, ***FST***), qui ne produit pas de résultat en sortie :

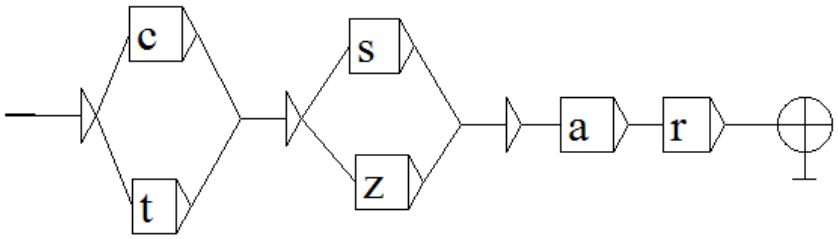


Figure 2 : Automate à états finis décrivant les variantes graphiques du mot français « csar »

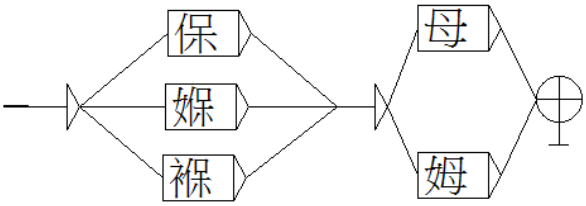


Figure 3 : Automate à états finis décrivant les variantes graphiques du mot chinois « 保姆 »[4]

[4] 保姆 *bǎomǔ* 'nourrice'

2.4.2.2 Transducteurs à états finis (*Finite-State Transducers, FST*)

Lorsqu'on veut analyser des séquences, i.e. les associer à des résultats d'analyse, on utilise des transducteurs à états finis (***F**inite-**S**tate **T**ransducers*). Un transducteur à états finis se présente sous la forme d'un graphe qui permet la reconnaissance de certaines séquences de caractères (morphologie) ou d'***ALU***s (syntaxe) :

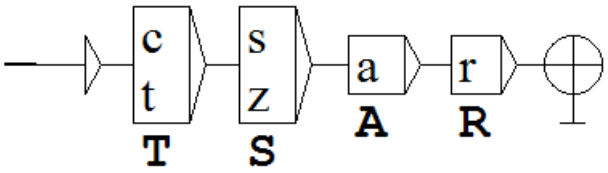


Figure 4 : Transducteur à états finis décrivant les variantes graphiques du mot français « TSAR »

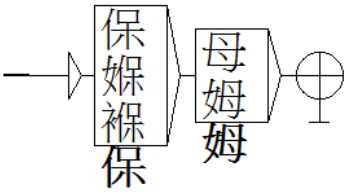


Figure 5 : Transducteur à états finis normalisant les variantes graphiques du mot chinois « 保姆 »

2.4.2.3 Réseaux de Transitions Récursifs (*Recursive Transition Networks, RTNs*)

Les Réseaux de Transitions Récursifs (***R**ecursive **T**ransition **N**etworks, **RTN**s*) constituent des grammaires contenant un ou plusieurs graphes, destinés aux transducteurs à états finis, aux automates à états finis ou aux graphes emballés. Afin de préciser les relations syntagmatiques, on utilise des automates, caractérisés par ***RTN***s, dans lesquels il est possible de créer des automates auxiliaires :

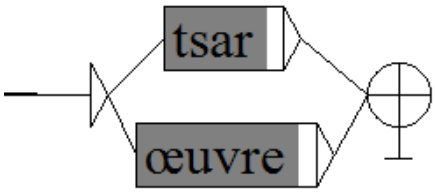


Figure 6 : Un réseau de transitions récursif pour décrire les variantes graphiques des mots français « tsar » et « œuvre »

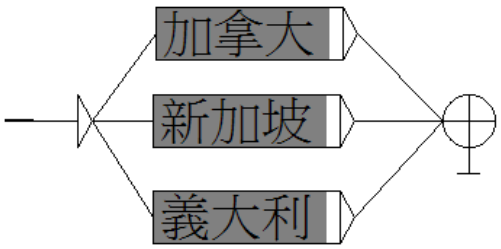


Figure 7 : Un réseau de transitions récursif pour décrire les variantes graphiques des mots chinois « 加拿大 », « 新加坡 » et « 義大利 »[5]

En général, les Réseaux de Transitions Récursifs servent à gérer des ensembles de graphes. Une fois que les graphes sont construits, ils peuvent être réutilisés dans des graphes plus complexes. Par la suite, l'ensemble de ces graphes complexes est à nouveau introduit dans d'autres graphes, selon une approche ascendante (bottom-up).

2.4.3 Construction des dictionnaires électroniques

2.4.3.1 Différences entre les dictionnaires électroniques et les dictionnaires usuels

Il existe une différence de fond entre les dictionnaires électroniques et les dictionnaires usuels. Les dictionnaires électroniques sont conçus pour être utilisés par des analyseurs de texte automatiques. Pour rendre possible l'analyse, toutes les informations linguistiques nécessaires doivent être associées à chaque entrée. Par exemple, au mot français « jumeau » sont associées des informations telles que ses formes flexionnelles, sa catégorie lexicale, etc. Au contraire, les dictionnaires usuels utilisés par les êtres humains, donnent simplement des explications sur des termes de référence, sur un verbe à l'infinitif, par exemple, mais ne donnent pas toujours explicitement les informations de base sur ses formes fléchies ou dérivées. On attend des personnes qui utilisent les dictionnaires usuels qu'elles puissent rapprocher elles-mêmes des informations supplémentaires à l'entrée qu'elles consultent.

Les dictionnaires électroniques doivent expliciter toutes les informations concernant leurs entrées, par exemple, les catégories grammaticales, les informations de distributions,

[5] 加拿大 *Jiānádà* 'Canada'
新加坡 *Xīngjiāpō* 'Singapour'
義大利 *Yìdàlì* 'Italie'

les formes flexionnelles et dérivationnelles. Les explications données dans les dictionnaires électroniques exigent une cohérence absolue, l'emploi, par exemple, d'un seul code pour tous les adverbes. C'est ainsi que les adverbes en chinois sont codés par le seul symbole **D**.

2.4.3.2 Dictionnaires électroniques de *NooJ*

Dans les dictionnaires électroniques de ***NooJ***, les informations linguistiques sont associées aux entrées. Nous devons relier chaque entrée à sa **catégorie**, par exemple **N** pour les noms et **V** pour les verbes. On peut aussi lui associer :

1) des **paradigmes flexionnels et dérivationnels**, comme, par exemple, la conjugaison des verbes, la formation d'adjectifs et d'adverbes à partir de noms, ou inversement, de formes nominalisées à partir de verbes ;

2) des **propriétés syntaxiques**, par exemple **+it** pour les verbes intransitifs et **+t** pour les verbes transitifs ;

3) des **distributions sémantiques**. Ainsi, **+Hum** est un classement associé à des noms humains, comme « artiste », « étudiant », « professeur », etc.

Les descriptions linguistiques associées aux entrées peuvent avoir deux formes :

1) La première est une représentation binaire : une entrée est associée à un ou plusieurs traits, par exemple, **+Hum** ou **+Hum+Sg**. Le dernier décrit que l'entrée appartient à la classe "humaine" et qu'elle est une forme du singulier ;

2) La deuxième exprime, sous forme de paire, une propriété **attribut = valeur**, par exemple, **+temps=présent**.

Les deux notations peuvent être mises en équivalence grâce à un système de définition des propriétés (fichier *property* de ***NooJ***). Ainsi, **+Sg** et **+Nombre=Sg** sont équivalents.

La propriété spéciale **FLX** permet de décrire le paradigme flexionnel de l'entrée lexicale, par exemple, **aider,V+FLX=AIMER** permet de dire que le verbe « aider » se conjugue selon le module de conjugaison « AIMER ».

De même, la propriété **DRV** permet de décrire le paradigme dérivationnel de l'entrée lexicale, par exemple, **jouer,V+FLX=AIMER+DRV=able** décrit la conjugaison du verbe « jouer » selon le paradigme « AIMER » ainsi que sa dérivation en adjectif « jouable ».

2.4.4 Représentation des grammaires

NooJ utilise des grammaires pour décrire des phénomènes orthographiques, morphologiques, ou syntaxiques. Dans le système ***NooJ***, il existe trois sortes de grammaires :

1) Les **grammaires flexionnelles ou dérivationnelles** sont sauvegardées dans les fichiers **.nof**. Ces grammaires servent à représenter les flexions ou les dérivations de certaines entrées des dictionnaires, sous forme de graphes ou de règles. Par exemple, les formes flexionnelles de certains noms sont décrites ainsi :

TABLE = <E>/singulier + s/pluriel;

Cette règle sert à produire la forme plurielle en ajoutant un « s » à l'entrée.

BANLU = <E>/singulier + 們/pluriel;

Cette règle sert à produire la forme plurielle de noms humains chinois en ajoutant un « 們 » à l'entrée.

2) Les **grammaires orthographiques, morphologiques, lexicales ou terminologiques** présentées dans les fichiers **.nom** sont des grammaires destinées à analyser des formes de mots :

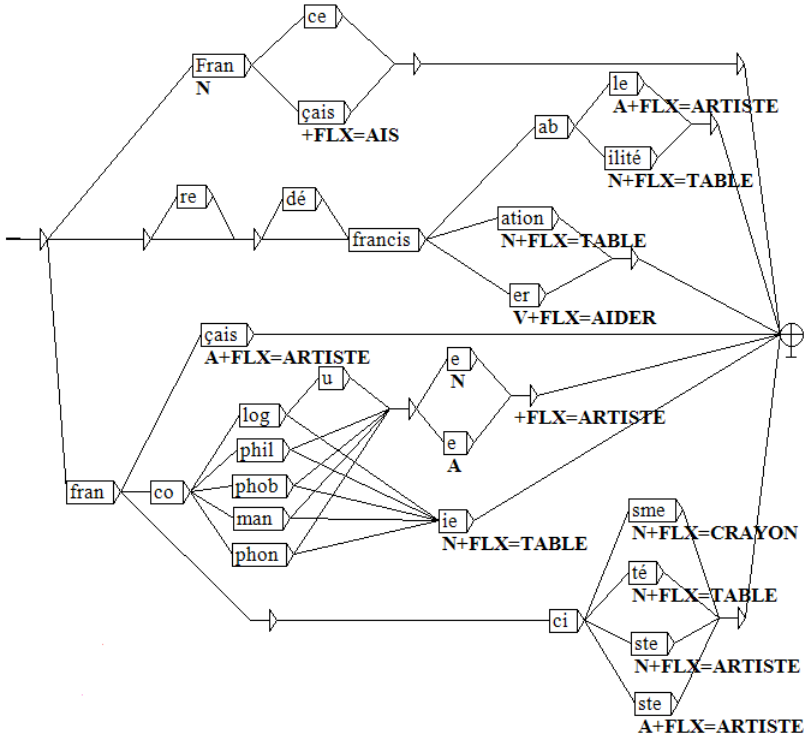


Figure 8 : Grammaire morphologique du mot français « France »

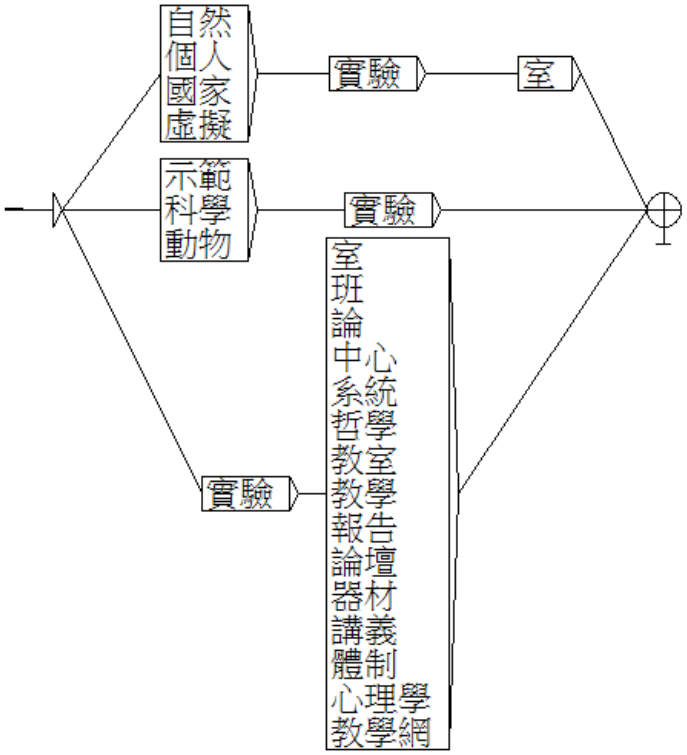


Figure 9 : Grammaire morphologique du mot chinois « 實驗 »[6]

3) Les **grammaires syntaxiques** sont rangées dans des fichiers **.nog**. Ces grammaires sont utilisées pour analyser automatiquement des expressions et les annoter. Elles peuvent être employées pour décrire des structures syntaxiques comme on le voit dans les exemples ci-dessous :

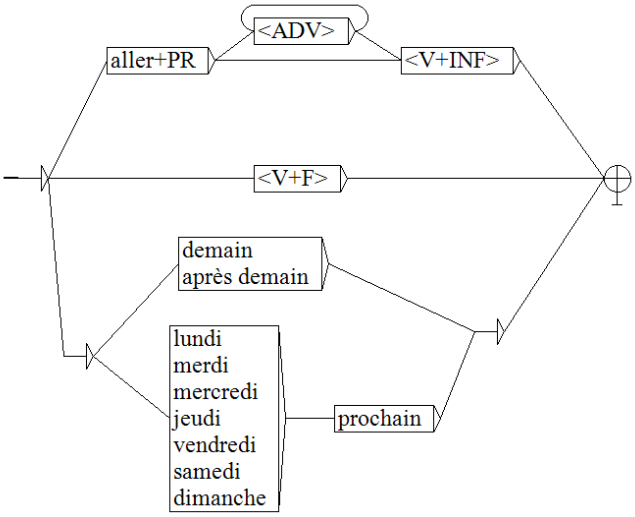


Figure 10 : Description du futur en français

[6] 實驗 *shíyàn* '(une) expérience scientifique'

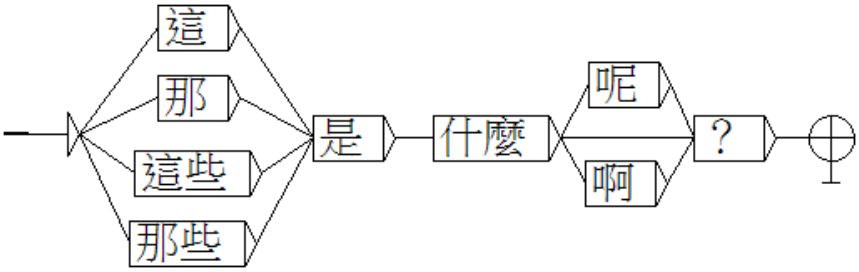


Figure 11 : Description des phrases interrogatives chinoises

2.4.5 Conclusion

NooJ permet aux utilisateurs de créer leurs propres ressources linguistiques : dictionnaires et grammaires. Les utilisateurs peuvent annoter des expressions auxquelles ils s'intéressent dans les textes. C'est la raison pour laquelle nous avons décidé de choisir le système ***NooJ*** comme support pour développer un module capable d'analyser de façon automatique un corpus en chinois moderne.

Chapitre 3

CADRE THEORIQUE : LE TRAITEMENT AUTOMATIQUE DU CHINOIS MODERNE

Les langues humaines sont considérées comme des “productions collectives spontanées” [*cf.* J. Allen, 1987 : 1-2] ; elles se définissent comme des langues naturelles, et se différencient des langues artificielles utilisées dans la programmation informatique ou en logique mathématique par leur constitution. Les éléments d’une langue naturelle sont communs à une communauté humaine. L’usage de cette langue n’est souvent ni explicite ni prescrit.

L’informatique, par les techniques de recherche automatique qu’elle propose pour exploiter efficacement la masse des informations, est à présent devenue un outil indispensable pour l’étude des langues dites naturelles.

Dans ce chapitre, nous insisterons d’abord sur la nécessité d’un corpus. Nous décrirons ensuite les difficultés rencontrées dans le traitement automatique du chinois moderne. En conclusion, nous montrerons comment la modélisation avec le système ***NooJ***, à différents niveaux, du chinois moderne a pu traiter ces difficultés.

3.1 Nécessité d’un corpus

Les corpus de textes fournissent une base d’unités lexicales pour la langue étudiée, ses règles morphologiques et syntaxiques, ainsi que l’ensemble de relations morpho-syntaxiques. L’ensemble des dictionnaires électroniques et des grammaires, traité comme un ensemble de ressources linguistiques, peut être amélioré par l’utilisation des corpus.

3.2 Ambiguïtés relatives à la langue chinoise

Les ambiguïtés de la langue sont à la source des difficultés les plus fréquemment rencontrées lors de son traitement automatique. En chinois, elles apparaissent à différents niveaux, comme en témoignent les deux exemples suivants.

3.2.1 Ambiguïté lexicale

En chinois, il est possible qu'un mot représente plusieurs sens. Il s'agit de polysémies. Par exemple, le mot 後門 *hòumén* <derrière-porte> renvoie à deux significations. L'une est le sens propre de 'porte de derrière', l'autre constitue le sens figuré de 'voie secrète'. Dès lors, le syntagme 走後門 *zǒuhòumén* / *zǒu hòumén* <marcher-derrière-porte> peut être une locution signifiant 'effectuer ses affaires de manière secrète' ou un syntagme verbal libre, ayant le sens de 'passer par la porte de derrière'.

3.2.2 Ambiguïté sémantique

Du point de vue syntagmatique, une unité lexicale peut avoir un sens ambigu. 吃飯 *chīfàn* / *chī fàn* <manger-riz>, par exemple, peut être utilisé comme un verbe ou comme un syntagme verbal libre. En tant que verbe, cette unité lexicale signifie 'prendre un repas'. En tant que syntagme libre, elle signifie 'manger du riz'. Comparons les deux phrases suivantes :

(7)他正在吃飯。 *tā zhèngzài chīfàn.*
<il-être en train de-prendre un repas.>
'Il est en train de manger (prendre un repas).'
(8)你要吃飯、吃麵或是吃蛋糕？ *nǐ yào chī fàn, chī miàn huòshì chī dàngāo ?*
<tu-vouloir-manger-riz, manger-nouille-ou-manger-gâteau ?>
'Tu veux manger du riz, des nouilles ou du gâteau ?'

Dans la phrase (8), le sujet, représenté par le pronom personnel *tā* 他, est en train de prendre un repas, et cette action est décrite par le verbe *chīfàn* 吃飯. Par contre, dans la phrase (9), l'action de manger *chī* 吃 se réfère à trois aliments spécifiques : *fàn* 飯, *miàn* 麵 et *dàngāo* 蛋糕. De ce point de vue, l'unité lexicale *chī fàn* 吃飯, qui a la signification de 'manger du riz', constitue plutôt un syntagme verbal qu'un mot verbal.

3.3 Désambiguïsation

Les deux ambiguïtés mentionnées ci-dessus peuvent être représentées en formalisant la langue à deux niveaux dans le système ***NooJ*** :

- Au niveau morphologique : nous déterminerons, en premier lieu, les unités lexicales. Elles entreront dans la construction des dictionnaires électroniques. Lors

de l'élaboration des dictionnaires, des propriétés linguistiques seront associées à chaque entrée.

● Au niveau syntaxique : nous décrirons les règles morphologiques, les structures syntaxiques et la relation morpho-syntaxique.

Au premier niveau, le vocabulaire de la langue est formalisé en développant des dictionnaires électroniques dont chaque entrée est assortie d'informations linguistiques. Au deuxième niveau, on décrit de façon formelle les règles syntaxiques en précisant la relation entre les unités lexicales. Les informations présentées dans le dictionnaire électronique sont utilisées lors du développement des grammaires.

3.3.1 Niveau morphologique

La reconnaissance des unités lexicales peut aboutir à un résultat qui prenne en compte les ambiguïtés affectant des unités lexicales comme :

● 吃飯 *chīfàn* / *chī fàn*, qui peut être un verbe ou un syntagme verbal ;

● 後門 *hòumén,* qui a un sens concret ou un sens abstrait ;

● 走後門 *zǒuhòumén* / *zǒu hòumén,* qui est une locution ou un syntagme verbal.

Pour les désambigüiser, elles doivent être considérées comme des unités lexicales différentes. Celles-ci sont lemmatisées dans le dictionnaire électronique qui en recense tous les usages possibles. Formalisées de manière à mettre en évidence leurs propriétés syntaxiques multiples, les trois unités mentionnées ci-dessus seront présentées comme suit :

吃,V
吃飯,V

後門,N+Conc
後門,N+Abst

走,V
走後門,L

飯,N[7]

[7] 吃,V *chī* 'manger'
吃飯,V *chīfàn* 'prendre un repas'

Grâce à cette formalisation des entrées de dictionnaires, les morphèmes peuvent être identifiés : le verbe 吃飯 *chīfàn*, le verbe 吃 *chī* et le nom 飯 *fàn*.

Dans le cas de 後門 *hòumén*, cette formalisation du lexique chinois permettra aussi de différencier les deux sens, de même graphie 後門 *hòumén*, chaque morphème étant traité comme une entrée indépendante avec attribution d'une catégorie lexicale suivie d'une distribution sémantique : N+Conc ou N+Abst. **Conc** représente « Concret », tandis que **Abst** renvoie au « Abstrait ».

D'ailleurs, le verbe 走 *zǒu* peut être reconnu, pusiqu'il représente aussi une entrée de dictionnaire.

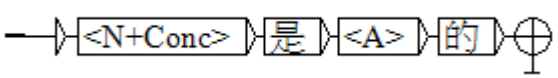
En tant que locution, 走後門 *zǒuhòumén* est égalment traité comme une entrée à laquelle est associée sa catégorie lexciale, indiquée par **L**.

L'ambiguïté au niveau des mots peut parfois être résolue grâce à des propriétés syntaxiques de ces mots. Prenons comme exemple l'adjectif 紅 *hóng* 'rouge' :

(9) 蘋果是紅的。 *píngguǒ shì hóng de.* <pomme-être-rouge-De.>
'La pomme est rouge.'
(10) *喜歡是紅的。*xǐhuān shì hóng de.* <aimer-être-rouge-De.>
'Aimer est rouge.'

La phrase (10) est correcte, car l'adjectif 紅 *hóng* qualifie un objet concret. Cependant, on ne peut pas l'utiliser pour caractériser une action exprimée par un verbe, comme dans l'exemple donné ci-dessus, 喜歡 *xǐhuān* [*cf.* (11)].

Pour désambiguïser les structures phrastiques incluant des adjectifs, on symbolise les règles qui régissent leur emploi par un schéma grammatical qui réutilise les catégories présentées dans le dictionnaire :



後門,N+Conc *hòumén* 'porte de derrière'
後門,N+Abst *hòumén* 'voie secrète'

走,V *zǒu* 'marcher'
走後門,L *zǒuhòumén* 'effectuer ses affaires de manière secrète'

飯,N *fàn* 'riz'

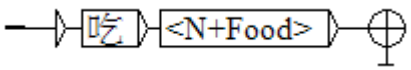
Cette grammaire peut éliminer automatiquement des phrases incorrectes comme (11). Elle vise donc à établir une équivalence entre unités substituables pour préserver la structure syntaxique.

Le dictionnaire doit donc contenir des informations, concernant les catégories ou les distributions sémantiques, qui permettront de rejeter des phrases du type de (11).

3.3.2 Niveau syntaxique

La description syntaxique vise à améliorer l'analyse des combinaisons d'unités lexicales, combinaisons qui possèdent un sens plus large et plus complexe que celui de chacune des unités qui les compose. Elle s'attache à définir les contraintes qui gouvernent les successions lexicales. Il s'agit d'établir des règles aptes à décrire la relation morpho-syntaxique. C'est la description syntaxique qui permettra de juger de la validité d'une séquence de mots.

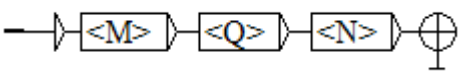
Comme nous l'avons mentionné, le verbe 吃飯 *chīfàn* 'prendre un repas' est reconnu car il figure comme une des entrées dans le dictionnaire. Par contre, pour reconnaître des syntagmes verbaux composés du verbe 吃 *chī* 'manger' suivi d'un nom d'aliment, on peut construire la grammaire locale suivante :



Cette grammaire vise à formaliser la structure de syntagmes libres tels que

吃飯 *chī fàn* <manger-riz> 'manger du riz',
吃麵 *chī miàn* <manger-nouille> 'manger des nouilles' ou
吃蛋糕 *chī dàngāo* <manger-gâteau> 'manger du gâteau'.

On peut établir une grammaire pour reconnaître des syntagmes, mais aussi pour éliminer des compositions inacceptables. Ainsi, la grammaire ci-dessous, sert à décrire des syntagmes nominaux composés d'un numéral (<M>), suivi d'un quantifieur (<Q>) et d'un nom (<N>) [*cf.* 5.6.3] :



Cette grammaire permet de reconnaître des syntagmes nominaux tels que

一顆蘋果 *yī kē píngguǒ* <un-Q-pomme> ‘une pomme’,
三本書 *sān běn shū* <trois-Q-livre> ‘trois livres’ ou
五隻貓 *wǔ zhī māo* <cinq-Q-chat> ‘cinq chats’.

Elle permettra donc d’éliminer des syntagmes inacceptables, par exemple,

*一蘋果 *yī píngguǒ* <un-pomme>,
*三書 *sān shū* <trois-livre> ou
*五貓 *wǔ māo* <cinq-chat>.

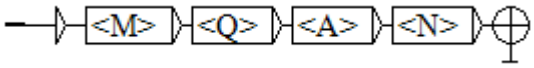
On peut faire de ce syntagme nominal simple un syntagme complexe en ajoutant au nom des adjectifs qui le qualifient. Les adjectifs peuvent ou non être suivis du subordinateur nominal 的 *de*, comme le montrent les exemples suivants :

(11) M – Q – N
一顆蘋果 *yī kē píngguǒ* <un-Q-pomme> ‘une pomme’

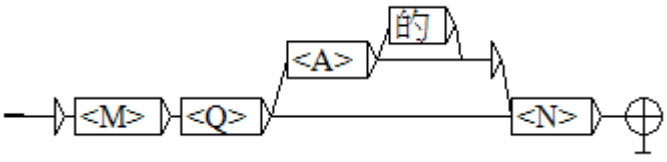
(12) M – Q – A – N
a. 一顆紅蘋果 *yī kē hóng píngguǒ* <un-Q-rouge-pomme> ‘une pomme rouge’
b. 一顆小蘋果 *yī kē xiǎo píngguǒ* <un-Q-petit-pomme> ‘une petite pomme’

(13) M – Q – A – 的 – N
a. 一顆紅的蘋果 *yī kē hóng de píngguǒ* <un-Q-rouge-De-pomme>
‘une pomme rouge’
b. 一顆小的蘋果 *yī kē xiǎo de píngguǒ* <un-Q-petit-De-pomme>
‘une petite pomme’

Dans les syntagmes (13), les deux adjectifs 紅 *hóng* et 小 *xiǎo* servent à qualifier le nom 蘋果 *píngguǒ*. L’adjectif est alors ajouté à la grammaire mentionnée précédemment, pour donner la représentation formelle grammaticale suivante :



En chinois, l’adjectif précède toujours les noms, et peut être suivi du subordinateur nominal 的 *de* [*cf.* (14)]. La grammaire est alors ainsi présentée :



Cette grammaire permet de reconnaître des syntagmes nominaux de trois structures différentes [*cf.* (12)-(14)].

3.4 Conclusion

Dans cette recherche, le traitement automatique du chinois moderne commence par le développement d'un module chinois dans le système ***NooJ***. Ce développement se fait à partir des données d'où l'on extraira des informations linguistiques. Nos connaissances linguistiques vont être utilisées dans la construction de dictionnaires et de grammaires. Cet ensemble de ressources linguistiques permet d'effectuer des recherches d'informations dans notre corpus. Il permet aussi de lever les ambiguïtés découvertes à différents niveaux. Par ailleurs, l'analyse des données permet ultérieurement d'améliorer les performances des analyseurs par ajout d'informations aux entrées de dictionnaire et de précisions concernant les structures syntaxiques.

Deuxième partie : Développement du module chinois dans *NooJ*

Chapitre 4

CONSTITUTION DU CORPUS

Pour développer le module chinois de ***NooJ***, il nous fallait disposer d'un corpus de textes représentatif de la langue chinoise moderne. Nous avons donc choisi d'établir un corpus, à l'intérieur duquel nous avons distingué deux types de textes distincts par leur genre et par leur niveau de langue : l'un est constitué de textes littéraires, l'autre, d'articles parus dans la presse écrite. Ces textes nous fournissent diverses informations linguistiques, aussi bien morphologiques et lexicales que syntaxiques. En nous appuyant sur ce corpus, nous pourrons lemmatiser les unités morphologiques dans les dictionnaires, puis les catégoriser, décrire les règles morphologiques et syntaxiques et formaliser les relations entre les mots et la syntaxe des phrases. L'intérêt de l'utilisation d'un tel corpus est donc d'affiner les performances des lemmatiseurs et des analyseurs morpho-syntaxiques.

La collecte des textes s'effectue par voie électronique. Il s'agit de recueillir des textes numériques sur Internet. Ces textes sont enregistrés sous le format texte brut (.txt), spécifié par une implémentation d'*Unicode*, *UTF-8* [*cf*. infra. § 4.1.2.1].

Dans ce chapitre, nous présenterons les étapes qui ont présidé à la constitution de notre corpus :

1) Nous examinerons les différents systèmes de codage qui permettent de représenter les textes numériques ;

2) Nous montrerons quels ont été les travaux relatifs à la construction de notre corpus ;

3) Nous présenterons les tâches que nous avons dû réaliser pour résoudre les nombreux problèmes posés par les caractères chinois, et qui sont survenus à la suite de la collecte des textes ou de la conversion d'un système de codage à l'autre ;

4) Enfin, nous discuterons de la mise en œuvre du formatage de notre corpus.

4.1 Systèmes de codage

4.1.1 Systèmes de codage des caractères chinois définis à Taiwan et en Chine

4.1.1.1 *Big-5*

En 1984, l'Institut d'Information Industrielle (« Institute for Information Industry »)[8] de Taiwan a défini la méthode de codage appelée *Big-5* (ou *Big5*)[9], 'Codage des cinq grands', pour supporter les cinq principaux logiciels utilisés à cette époque : logiciels de gestion documentaire, logiciels tableurs, logiciels graphiques, bases de données et logiciels de communication. Ayant recours à l'utilisation de doubles-octets, ce système de codage prend en charge les caractères chinois traditionnels. Néanmoins, il ne comporte pas tous les caractères fréquemment employés, c'est pourquoi les auteurs de ce codage ont développé leurs propres extensions. À présent, ce système de codage contient 13 053 caractères, mais la plupart des caractères y possèdent deux codes. Par exemple, le caractère 兀 *wù* 'solitaire' est codé à la fois par A461 et par C94A. Par ailleurs, le caractère 慪 *hù* 'vomir' est également représenté par deux codes différents : DCD1 et DDFC. Ainsi, ce codage ne comporte en réalité que 8 058 caractères.

Le gouvernement de Hong Kong a également adopté ce jeu de caractères chinois *Big-5* pour coder le cantonais. Mais *Big-5* ne prend pas en compte le grand nombre de caractères spécifiques au cantonais. On lui a donc ajouté en 1995 une première extension, appelée « *Government Chinese Character Set* », puis une deuxième en 1999, « *Hong Kong Supplementary Character Set* », qui contient 4 702 caractères spécifiques au cantonais [*cf.* http://www.cmex.org.tw/index.html].

[8] Nous avons traduit le titre anglais en français.

[9] L'Institut d'Information Industrielle, 財團法人資訊工業策進會 *cáituán fǎrén zīxùn gōngyè cèjìnhuì*, a collaboré avec les cinq principales compagnies — 宏碁 *hóngjī* '*Acer*', 神通 *shéntōng* '*MiTAC*', 零壹 *língyī* '*Zero One*', 大眾 *dàzhòng* '*FIC*' et 佳佳 *jiājiā* '*Jia Jia*' — pour élaborer le codage *Big-5*.

4.1.1.2 *CNS[10] 11643*

CNS 11643 est un système de codage des caractères chinois reconnu comme standard officiel à Taiwan. Il est connu sous l'appellation anglaise « *Chinese Standard Interchange Code* », et sous l'intitulé chinois « 中文標準交換碼 *zhāngwén biāozhǔn jiāohuànmǎ* ». En 1992, ce système de codage incluait seize plages qui comportent 48 027 caractères chinois. Ces seize plages ont été étendues, jusqu'à atteindre quatre-vingt plages en 2004. Cependant, ce codage est beaucoup moins connu que le *Big-5*, qui jouit d'une importante popularité, en raison notamment de son usage répandu dans les systèmes d'exploitation *Microsoft* et *Apple* [*cf.* http://www.cns11643.gov.tw/AIDB/welcome.do].

4.1.1.3 *CCCII*

Le traitement automatique des textes écrits en langues de l'Asie orientale fut une nécessité importante, en particulier pour la recherche des documents dans les bibliothèques. L'Université Stanford en Californie, a tenu, en 1979, une conférence, au cours de laquelle il a été, dans un premier temps, proposé d'adopter le code *JISC*[11] *6226*, qui était alors utilisé pour décrire le japonais. Cependant, comme ce codage spécifique aux caractères kanji n'est pas adapté aux caractères chinois, la Bibliothèque sino-américaine de l'Asie orientale[12] (« *Chinese American East Asian Library* ») et les représentants taïwanais ont refusé l'adoption du codage japonais. De ce fait, les représentants taïwanais se sont engagés à proposer un jeu de caractères chinois, afin de le comparer avec le codage *JISC 6226*.

De retour à Taïwan, les représentants ont fait leur rapport au Conseil national scientifique[13] (« *National Science Council* ») et à l'Association sino-américaine[14] (« *Chinese American Association* »). C'est ainsi qu'a été fondé le Groupe d'analyse des caractères chinois[15] (« *Chinese Character Analysis Group* ») et qu'il a élaboré le projet de codage « *Chinese Character Code for Information Interchange* ».

[10] ***CNS*** est l'abréviation anglaise de ***C****hinese* ***N****ational* ***S****tandard*.

[11] Le sigle anglais ***JISC*** signifie ***J****apanese* ***I****ndustrial* ***S****tandards* ***C****ommittee*. Ce comité s'occupe de toutes les normalisations nationales et internationales au Japon.

[12] *cf.* n. 8.

[13] *cf.* n. 8.

[14] *cf.* n. 8.

[15] *cf.* n. 8.

Ce système de codage *CCCII*, contenant 4 808 caractères, a été présenté lors de la conférence de l'*Asian Study Association* en 1980. Dans le but de coder davantage de caractères chinois, des extensions ont été développées. Actuellement, ce codage *CCCII* permet de représenter 75 684 caractères chinois, incluant les caractères simplifiés. Il est appliqué dans de nombreuses bibliothèques à travers le monde [*cf.* http://www.cns11643.gov.tw/AIDB/welcome.do].

4.1.1.4 *GB*[16]

Le gouvernement chinois a défini le *GB* en 1980, afin de coder les caractères chinois simplifiés. Ce système de codage contient 6 763 caractères simplifiés. Cependant, il n'était pas complet. Un autre jeu de caractères chinois, *GBK*[17], a été développé en 1995, avec 21 003 caractères. Complétée par des extensions, la synthèse *GB 18030* contient 70 224 caractères. En raison de sa compatibilité avec *Unicode*, ce codage gère aussi bien les caractères chinois simplifiés que les caractères chinois traditionnels [*cf.* http://www.moe.edu.cn/edoas/website18/17/info16417.htm].

4.1.2 *Unicode* et ses implémentations

Unicode est un codage développé par le consortium *Unicode*. Cette organisation a pour objectif de rassembler dans un même standard tous les caractères des langues du monde entier. Dans ce contexte de développement, *Unicode* donne à tout caractère un nom et un seul identifiant numérique. *Unicode* a été adopté à la fois par les organismes de standardisation et les développeurs qui ont fait le choix au début d'unifier la norme *ISO/IEC*[18] *10646* en cours de conception et *Unicode* [*cf.* http://www.Unicode.org].

4.1.2.1 *UTF-8*

Inventé par Kenneth Thompson de *Bell Laboratoires* en 1992, *UTF-8* (« *Universal character set Transformation Format 8 bits* ») permet de coder les caractères par une suite

[16] Le sigle ***GB*** correspond aux initiales de 国家标准 en pinyin ***G**uójiā **B**iāozhǔn* 'standard national'.

[17] Pour ***GBK***, la lettre « ***K*** » provient de la première lettre du pinyin 扩展 ***K**uòzhǎn*, signifiant 'extension' en français.

[18] ***ISO*** est l'acronyme d'« ***I**nternational **O**rganization for **S**tandardization* ». Au 31 décembre 2006, cette organisation de normalisation se composait de représentants de 158 pays.
IEC désigne « ***I**nternational **E**lectrotechnical **C**ommission* ». Cet organisme, complémentaire de l'Organisation Internationale de Normalisation, s'occupe de domaines électrotechniques tels que l'électricité, l'électronique et les techniques analogues.

d'un à quatre octets. Autrement dit, les caractères sont représentés par des séquences d'octets de longueurs variables. Les caractères numérotés de 0 à 127 sont codés sur un octet dont le bit de poids fort est nul. Ces 128 caractères sont identiques à ceux de 7 bits définis par *ASCII*. Donc, un texte qui contient seulement ces caractères est codé de la même façon en *ASCII* et en *UTF-8*.

UTF-8 est un format de codage très utilisé pour les applications d'Internet ou les langages de programmation, par exemple, le langage C#.

4.1.2.2 *UTF-16*

L'implémentation *UTF-16* (« *Universal character set Transformation Format 16 bits* ») code les caractères avec un ou deux mots de seize bits. Plus précisément, le Plan Multilingue de Base (***B**asic **M**ultilingual **P**lane, **BMP***) est réservé aux écritures alphabétiques, syllabiques, idéographiques ainsi qu'aux chiffres et aux symboles. Ces caractères sont codés par un seul mot de seize bits, c'est-à-dire, deux octets. Les caractères d'autres plans sont codés par deux mots de seize bits. Ainsi, *UTF-16* code les 128 caractères de numéro 0 à 127 avec deux octets. Ce codage est, par conséquent, moins efficace pour traiter les documents qui comportent un grand nombre de caractères représentables par le code *ASCII*, car il utilise plus de mémoire pour stocker les caractères codés. Par exemple, pour coder le caractère **A**, le codage *ASCII* utilise seulement un octet de valeur : 65 alors qu'*UTF-16* le code avec deux octets : 0 – 65.

En *UTF-16*, chaque caractère étant représenté par deux ou quatre octets successifs, il existe une ambiguïté concernant l'ordre dans lequel ces octets sont placés en mémoire : l'octet de poids fort d'abord ou l'octet de poids faible. Pour transmettre correctement les documents entre systèmes hétérogènes, il faut résoudre cette ambiguïté dans l'ordonnancement des octets. De ce fait, il est nécessaire de préciser les noms des protocoles utilisés : *Big-Endian* (*BE*) et *Little-Endian* (*LE*). L'architecture *Big-Endian* enregistre l'octet de poids fort dans l'adresse mémoire la plus petite. L'architecture *Little-Endian* enregistre l'octet de poids faible dans l'adresse mémoire la plus petite. Par exemple, la valeur 1A2B3456 peut être enregistrée selon ces deux architectures :

Endian ordre	Octet 1	Octet 2	Octet 3	Octet 4
Big-Endian	1A	2B	34	56
Little-Endian	56	34	2B	1A

Tableau 2 : Exemple de stockage des données *Big-Endian* et *Little-Endian*

Si l'ordonnancement des octets n'est pas précisé, des difficultés ou des ambiguïtés dans l'affichage des caractères seront constatées après la transmission des données d'une machine à une autre. Par exemple, la valeur hexadécimale du caractère chinois 羊 *yáng* 'mouton' est 7F8A. Mais la manière de la stocker physiquement diffère : 7F8A en *Big-Endian* ; 8A7F en *Little-Endian*. Si l'ordonnancement de ces deux octets (7F et 8A) n'est pas préalablement déterminé, des ambiguïtés se produiront dans la représentation de caractères. La valeur hexadécimale 7F8A de 羊 *yáng* peut être lue par erreur comme 8A7F, correspondant à 誖 *guà* 'erreur' car la valeur hexadécimale du caractère 誖 *guà* est précisément la transposée de celle du caractère 羊 *yáng* : 8A7F en *Big-Endian* ; 7F8A en *Little-Endian*.

Une marque peut servir à indiquer la manière de coder les données. Cette marque, qui ne représente aucun caractère, est implantée en tête des fichiers à transmettre ou à lire. Le code U+FEFF spécifie que le texte qui suit est à interpréter en *Big-Endian*, et le code U+FFFE spécifie qu'il est à interpréter en *Little-Endian*.

4.1.2.3 *UTF-32*

UTF-32 (« *Universal character set Transformation Format 32 bits* ») se définit comme une implémentation d'*Unicode* où chaque caractère est codé par un mot de trente-deux bits, ce qui équivaut à quatre octets. Considéré comme un format de codage simple, il exige une mémoire importante pour représenter les caractères. Ainsi ce format peu économique, car à partir d'une certaine longueur de texte, il nécessite davantage de temps de lecture et d'espace de stockage sur le disque ou en mémoire vive, ce qui diminue les performances pendant les processus de traitement. En effet, la quantité de mémoire nécessitée par *UTF-32* est le double de celle d'*UTF-16*. Par conséquent, les traitements textuels utilisent plus souvent l'implémentation *UTF-16* qu'*UTF-32*.

Comme l'implémentation *UTF-16*, l'implémentation *UTF-32* possède aussi deux schémas de codage normalisés : *UTF-32BE* et *UTF-32LE*. Au moment de la transmission textuelle entre les systèmes, il est donc nécessaire de préciser le schéma de codage pour éviter toute incompatibilité.

4.1.2.4 Applications des implémentations d'*Unicode*

Nous tentons ici de rendre plus claires les applications de ces trois implémentations d'*Unicode*. Prenons l'exemple d'un caractère chinois, 樹 *shù* 'arbre'. Le codage *Unicode* lui attribue la valeur décimale de 27 – 193. D'après la définition des trois implémentations d'*Unicode*, cette valeur numérique se transforme en trois autres valeurs différentes. *UTF-8* code ce caractère 樹 *shù* avec trois octets représenté par 230 – 168 – 185. *UTF-16 LE* le code avec deux octets représentés par 57 – 106. Si cette valeur est représentée selon l'architecture *Big-Endian*, elle s'écrit 106 – 57. *UTF-32 LE* le code avec quatre octets représentés par 57 – 106 – 0 – 0. Elle s'écrit 0 – 0 – 106 – 57 selon l'architecture *Big-Endian.*

4.1.3 Unification des caractères chinois en chinois, en japonais, en coréen et en vietnamien

Les Japonais, les Coréens et aussi les Vietnamiens ont adopté une partie des caractères chinois dans leurs écritures. Les caractères chinois utilisés dans les écritures de ces trois langues ainsi que dans celle de la langue chinoise sont définis comme « Han » par le Projet *Unihan* (« *Project Unification Han* »). Ce projet est porté par le *Research Libraries Group*, *Xerox*, *Taligent* et *Apple*.

Unicode intègre, en tant que sous-ensemble, les caractères Han de ces quatre écritures dans son jeu de caractères. Les caractères Han constituent une plage de caractères nommée *CJKV*[19]. Ces caractères Han concernent les caractères employés dans le chinois traditionnel, le chinois simplifié, le japonais, le coréen et le vietnamien. Ces caractères Han n'incluent ni les syllabaires hiragana et katakana du japonais, ni l'alphabet hangûl du coréen, ni l'écriture latinisée du vietnamien. Selon la formulation d'*Unicode*, les différences entre quatre langues ne sont que graphiques ; autrement dit, elles sont représentées par des polices ou des glyphes différents.

[19] ***CJKV*** est l'abréviation officielle anglaise, définie par l'*ISO/IEC 10646*, qui signifie ***C****hinese* – ***J****apanese* – ***K****orean* – ***V****ietnamese*.

4.1.4 Difficultés de la représentation graphique des caractères chinois en *Unicode*

Le terme « glyphe » désigne une représentation graphique particulière d'une forme graphique. Comme nous l'avons mentionné ci-dessus, les caractères chinois sont aussi bien employés en chinois qu'en japonais, en coréen et en vietnamien, mais leurs glyphes sont différents dans ces langues.

Les Japonais ont pris l'habitude d'écrire certains caractères chinois d'une manière légèrement différente de celle utilisée à Taïwan ou en Chine ("coups de pinceau"), même si formellement les traits sont identiques. Ces différences subtiles ont ensuite été rendues électroniquement au travers de polices de caractères créées pour refléter ces habitudes locales. C'est ce qui permet à un connaisseur des caractères Han de dire, sans voir l'ensemble du texte et donc la syntaxe de la langue utilisée, que telle représentation imprimée provient probablement d'un journal imprimé au Japon, à Taïwan ou en Chine. Ces différences de "glyphes" peuvent donc être importantes.

Or, les caractères *CJKV* sont codés de façon unifiée en *Unicode*. Plus précisément, le codage *Unicode* code les caractères, et non pas leurs glyphes. Par conséquent, une police de caractères appropriée est indispensable pour représenter les glyphes. Si nous ignorons dans quelle langue un caractère est écrit, ou quel style est utilisé pour le représenter correctement, ce caractère deviendra inadapté à son contexte lors de la transmission entre différentes plateformes. Il en résulte que la représentation graphique des caractères doit dépendre du contexte où il se trouve.

Prenons l'exemple du caractère 情 *qíng* 'sentiment'. Son écriture diffère selon qu'elle soit en chinois traditionnel, en chinois simplifié, en japonais et en coréen. Les différences graphiques sont minimes, comme le montre le tableau suivant :

Valeur décimale d'*Unicode*	Chinois traditionnel	Chinois simplifié	Japonais	Coréen
24 – 773	情	情	情	情

Pour afficher correctement les caractères dans les langues *CJKV*, il est important de prendre en compte les polices de caractères utilisées, qui doivent correspondre aux pratiques scripturales des langues concernées.

Par ailleurs, le jeu spécifique des caractères chinois en *Unicode* s'est beaucoup élargi au fur et à mesure du développement des extensions. Malgré cette évolution, ce jeu est encore loin de coder tous les caractères chinois. Ce fait rend impossible la transmission de documents dans lesquels il y a des caractères chinois qui ne possèdent pas de codes *Unicode*.

Comme nous l'avons mentionné plus haut, certains caractères de ces langues asiatiques sont quasiment identiques, mais possèdent une minime différence dans leurs formes graphiques (voir l'exemple donné plus haut, où ces différences résident dans la manière de joindre — ou non — deux traits etc.). De ce fait, causé par l'unification de caractères appartenant à différentes langues, la sélection du glyphe correct est encore en cours de résolution par la norme *Unicode*. Chaque auteur rédige ses textes dans la police de caractères qu'il préfère. Une solution temporaire a été proposée.

4.1.5 Système de codage *TRON*

Puisqu'*Unicode* n'intègre pas tous les caractères des langues *CJKV*, plusieurs jeux multilingues de caractères ont été développés au Japon. Parmi ces nouveaux jeux de caractères, le système *TRON* est le plus réputé. Le projet *TRON* a été proposé au début des années 1980 afin de coder les polices de caractères déjà apparues ou à venir. Dans l'environnement multilingue *TRON*, les caractères ne sont pas unifiés selon la méthode mise en œuvre par *Unihan*. Actuellement, une implémentation commerciale de *TRON*, Cho Kanji 3, intègre 171 500 caractères [*cf*. http://www.tron.org/].

4.1.6 Conclusion

Unicode permet de représenter la plupart des caractères des langues dans le monde. L'implémentation *UTF-32* nécessite une mémoire importante pour représenter les caractères. Ce format de codage est très simple mais il exige davantage de temps de lecture et de place d'écriture sur le disque. À l'inverse, *UTF-8* est le codage le plus économique, et est très utilisé dans les systèmes informatiques actuellement.

Nous avons décidé d'adopter *UTF-8* pour coder nos données de recherche. Cette implémentation affiche correctement les caractères tout en nécessitant moins de mémoire qu'*UTF-32*, ce qui favorise le traitement des documents.

4.2 Description du corpus

Notre corpus est conçu comme un révélateur d'usages en chinois moderne du XX^{e} siècle. Il doit donc contenir des ressources linguistiques en quantité suffisante. Il regroupe trois types de textes :

1) Textes littéraires du XX^{e} siècle [*cf.* Annexe 1] :

Nous avons recueilli des textes littéraires de styles variés, datés des années 1910 aux années 1990. Ils proviennent des sites suivants :

1) 文學視界 *wénxué shìjiè* : http://www.white-collar.net
2) 雲臺書屋 *yúntái shūwū* : http://www.b111.net/index.htm
3) 龍騰世界書庫 *lóngténg shìjiè shūkù* : http://www.millionbook.net

Nous avons choisi des textes d'auteurs représentatifs tant Chinois continentaux que Taïwanais, par exemple, Bing Xin [1919], Yu Dafu [1921], Lu Xun [1923], Shen Congwen [1934], Xu Dishan [1934], Lin Yutang [1939], Lin Haiying [1960], Qiong Yao [1963], Bai Xiangyong [1971], etc. Les trente-neuf œuvres sélectionnées couvrent une période de plus d'un demi-siècle, période pendant laquelle la langue n'a guère subi de changements significatifs sur le plan syntaxique [*cf.* He Jiuying, 2005 : 109-139]. Ils comprennent un volume de données d'environ 15,8 mégaoctets, soit approximativement 7 300 000 caractères.

Ces ouvrages littéraires reflètent des évènements sociaux qui non seulement motivèrent la naissance de divers thèmes, mais influencèrent le choix du vocabulaire et des expressions dans la littérature. Selon la chronologie, les textes littéraires sélectionnés peuvent être distribués en trois groupes. Le tableau ci-dessous donne la répartition en pourcentage de chacun de ces groupes au sein de l'ensemble [*cf.* Chen Sihe, 1999] :

Époques	Thèmes élaborés	Répartitions	
		Écrivains taïwanais	Écrivains chinois
1919 — 1949	Famille, société, personnalités caractéristiques, avenir, résistance chinoise aux japonais, guerre civile, etc.	10,26 %	28,20 %
1949 — 1978	Instrumentalisation du régime, influence de la révolution culturelle, des pensées de gauche, des mouvements sociaux, souvenirs de la vie d'autrefois, vie de petits personnages, etc.	20,51 %	7,69 %
1978 — 1993	Description des différentes générations, variété des points de vue personnels portant sur la société, la vie, l'avenir, etc.	10,26 %	23,08 %

2) Textes journalistiques publiés dans le 聯合電子報 *liánhé diànzǐbào* '*United Daily News*' du 1er juillet 2007 au 30 juin 2008 à Taïwan :

http://paper.udn.com/UDN/Subscribe/subscribe

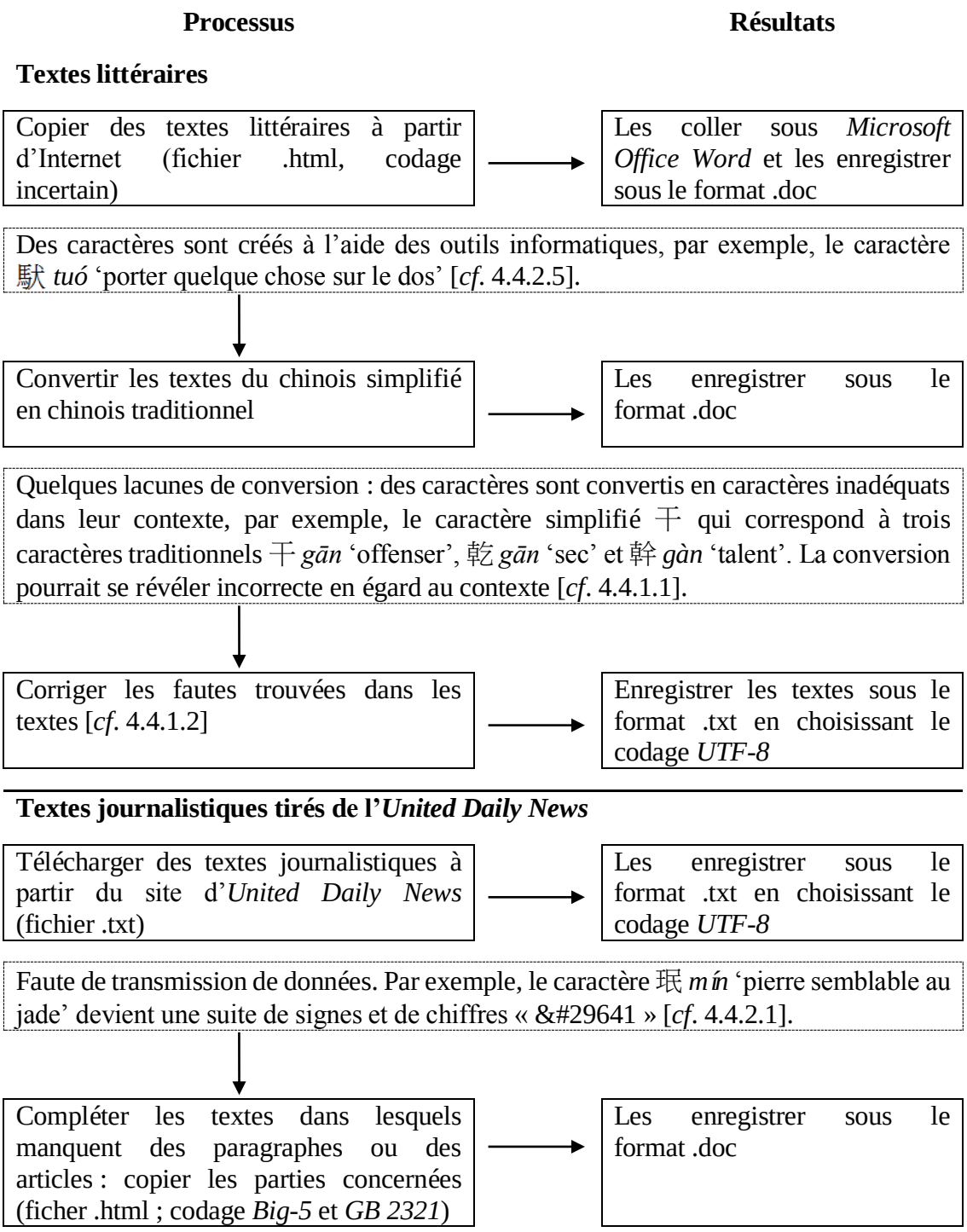
De par la variété des thèmes abordés et la personnalité de chaque journaliste, la presse propose un large éventail de styles. Les textes de ce type sont riches en vocabulaire et offrent une vision assez complète des règles qui gouvernent l'usage de la langue chinoise. L'ensemble des articles que nous avons extraits de l'*United Daily News* compte approximativement 7 000 000 de caractères. Cette section du corpus correspond à peu près à 15,2 mégaoctets.

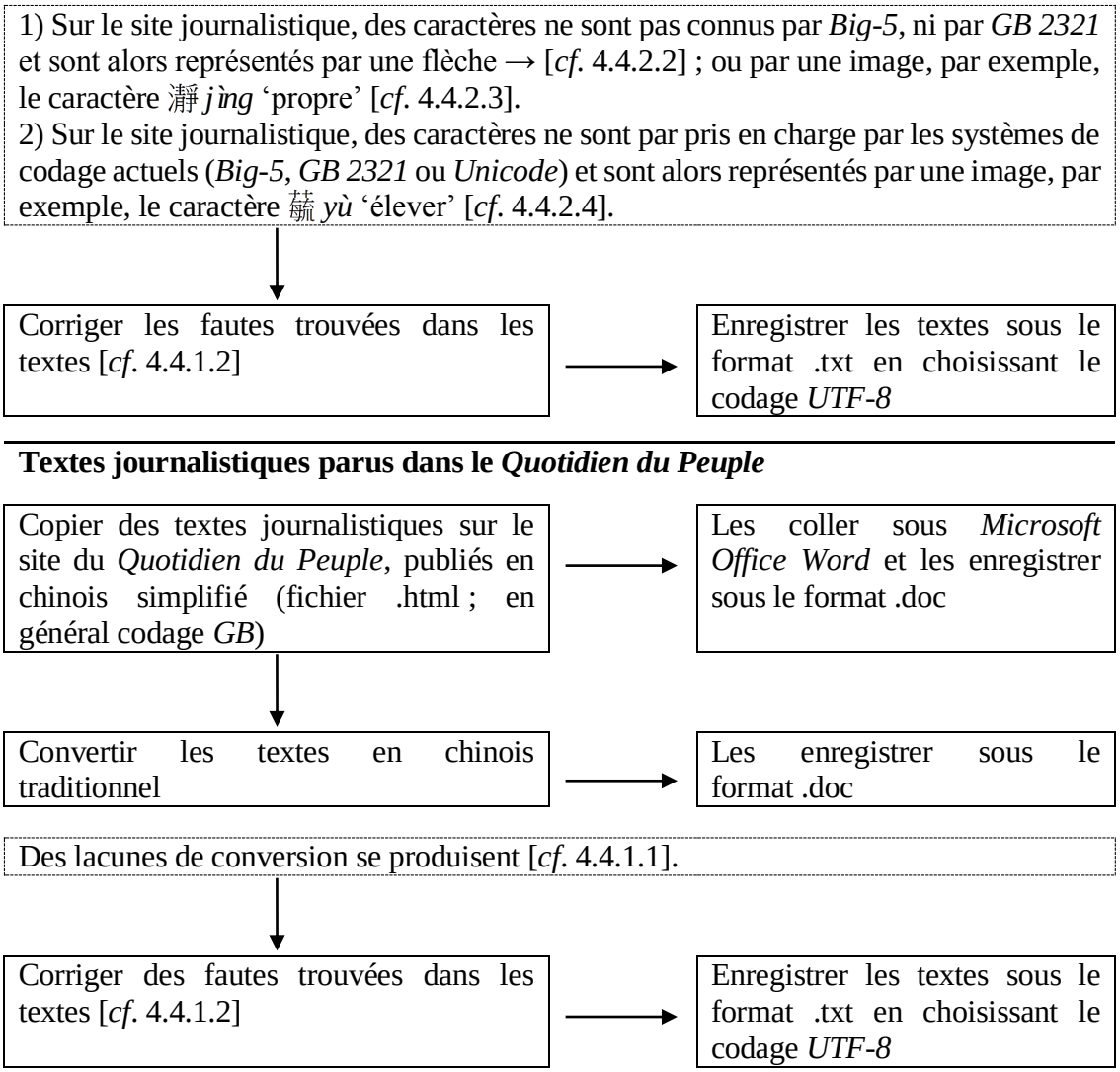
3) Textes journalistiques du 人民日報 *rénmín rìbào* '*Le Quotidien du Peuple*' publiés du 1er au 31 mars 2008, en Chine continentale : http://www.people.com.cn

Le *Quotidien du Peuple* est un journal officiel en République populaire de Chine. Les articles de ce journal que nous avons regroupés contiennent environ 3 600 000 caractères. Le tout représente environ 7,78 mégaoctets. Les articles de cette presse sont écrits en caractères simplifiés. Pour cette étude, nous les avons convertis en caractères traditionnels car l'objectif de cette recherche est de développer un module d'analyse du chinois basé sur l'écriture des caractères traditionnels.

4.3 Collecte des textes

Notre corpus contient ces trois types de textes. Ci-dessous, nous présentons le protocole suivi pour recueillir des textes sur Internet :





4.4 Correction du corpus

Nous avons constaté de nombreuses erreurs dans les textes recueillis sur Internet, et nous avons dû les corriger. Nous détaillons ci-dessous les travaux de correction réalisés pour que notre corpus soit analysable automatiquement.

4.4.1 Premières corrections des textes

4.4.1.1 Conversion du chinois simplifié en chinois traditionnel

La plupart des textes littéraires qu’on a scannés sont écrits en chinois simplifié. Notre recherche portant sur l’analyse du chinois traditionnel, nous avons dû convertir ces textes en chinois traditionnel à l’aide d’une fonction de conversion du chinois simplifié vers le chinois traditionnel incluse dans *Microsoft Office Word*. Néanmoins, cette technique de conversion comporte des limites en raison d’ambiguïtés graphiques, un caractère simplifié

pouvant correspondre à plusieurs caractères traditionnels. Ainsi le caractère simplifié 干 représente trois caractères traditionnels, qui sont 干 *gān* ‘offenser’, 乾 *gān* ‘sec’ et 幹 *gàn* ‘talent’. Lors des choix de simplification des caractères faits par le gouvernement chinois pendant les années 50 et 60, ces trois caractères traditionnels différents ont été unifiés.

Afin de résoudre ce défaut de conversion, nous avons vérifié les phrases où ils apparaissaient. On peut observer ci-dessous, trois mots écrits en chinois simplifié, leurs mauvaises conversions, marquées par un astérisque (*), ainsi que leurs formes correctes après correction :

Mots écrits en chinois simplifié	Formes lexicales reçues après la conversion	Formes lexicales correctes après correction
干净 *gānjìng* <sec-propre> ‘propre’	*幹淨 *gànjìng*	乾淨 *gānjìng* <sec-propre> ‘propre’
能干 *nénggàn* <talent-talent> ‘talent’	*能乾 *nénggān*	能幹 *nénggàn* <talent-talent> ‘talent’
树干 *shùgàn* <arbre-tranche> ‘tronc’	*樹乾 *shùgān*	樹幹 *shùgàn* <arbre-tranche> ‘tronc’

Les erreurs issues de la conversion ont été corrigées à l’aide de la fonction de remplacement proposée dans *Microsoft Office Word*. Pour mieux construire des textes formatés et pour relever ces ambiguïtés concernant les caractères chinois, nous avons relu tous les textes. Cette lecture nous a permis d’avoir des textes numériques “propres” et utilisables.

4.4.1.2 Rectification des caractères chinois inadéquats

Il arrive souvent qu’une faute de frappe ou une erreur de numérisation provoque l’apparition de caractères inadéquats. Nous avons corrigé les fautes de frappe [*cf.* Figure 12] et les erreurs de numérisation (erreur lors du passage au scanner) [*cf.* Figure 13].

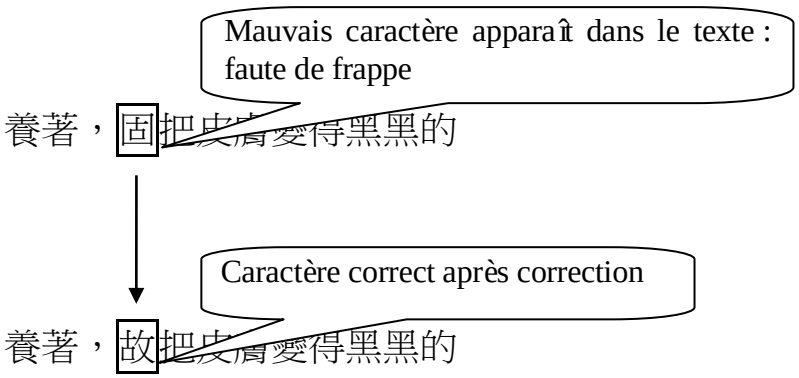


Figure 12 : Faute de frappe

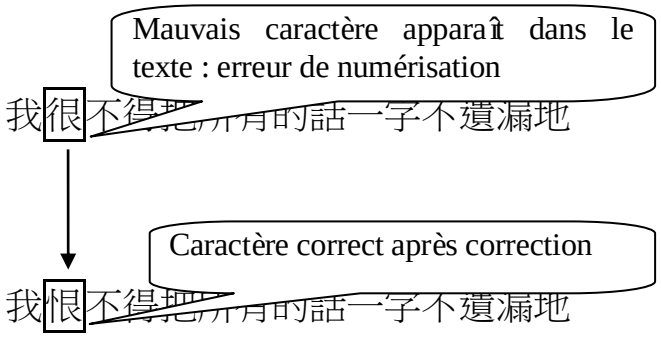


Figure 13 : Erreur de numérisation

Ces deux types d'erreurs sont les plus fréquemment rencontrés lors de la construction de notre corpus. Pour les éliminer nous avons relu chaque texte, et comparé en cas de doute notre version avec la version d'origine sur papier pour obtenir des textes numériques convenant aux applications dans ***NooJ***.

4.4.2 Corrections des caractères chinois dans les textes en chinois traditionnel

Après les premières corrections de textes, nous avons été amenée à résoudre le problème de l'absence de certains caractères chinois dans les systèmes de codage. Actuellement, les systèmes de codage n'assignent pas un code à tous les caractères chinois existants ; manquent en particulier, les caractères rares employés pour noter les noms propres, et les caractères utilisés dans des langues régionales (comme le cantonais). Par conséquent, certains outils ou logiciels, comme *TrueType* ou 中文全字庫 *zhōngwén quánzìkù* « *Master*

Ideographs Seeker for CNS 11643 Chinese Standard Interchange Code »[20], ont été inventés pour former des caractères. Ces logiciels combinent les caractères chinois déjà existants, ce qui permet de générer des nombreux caractères.

Pour pallier cette défaillance — apparition inexacte de caractères chinois — nous avons choisi de remplacer ces caractères par d'autres caractères, pris en charge par *Unicode*.

Nous avons rencontré cinq situations dans lesquelles les caractères ne sont pas correctement présentés. Nous détaillons ci-dessous nos choix de modification et de correction des caractères.

4.4.2.1 Caractères devenant *mojibake* sous le format texte

Certains caractères chinois s'affichent de manière correcte sur la page Web, mais deviennent illisibles lors de l'ouverture des fichiers sous le format texte. Nous avons donc corrigé ces défaillances de codage en nous référant aux textes originaux sur Internet. C'est ce que montre l'exemple suivant :

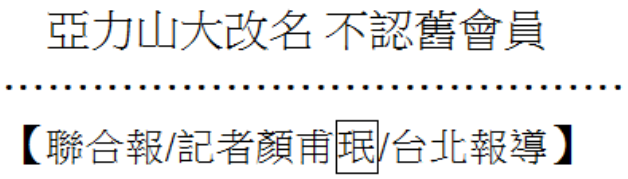


Figure 14 : Affichage correct du caractère 珉 *mín* sur le site journalistique du 16 janvier 2008

Le caractère 珉 *mín* 'pierre semblable au jade', entouré d'un carré noir, se présente de façon correcte sur le site de l'*United Daily News*. Les textes présentés sur ce site de presse sont codés en *Big-5* et *GB 2321*. Ce caractère 珉 *mín* est bien représentable avec ces deux systèmes de codage.

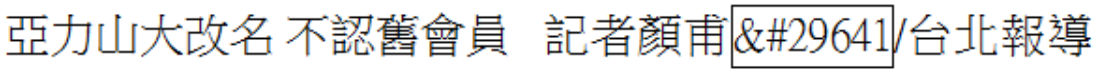


Figure 15 : Affichage d'une suite de signes et de numéros sous le format texte brut

Nous avons enregistré le journal du 16 janvier 2008 sous le format texte brut, puis nous l'avons ouvert. Nous avons constaté que ce caractère 珉 *mín* était devenu une suite de

[20] *cf.* C'est une application du codage *CNS*. Voir : http://www.cns11643.gov.tw/web/index.jsp.

signes et de numéros : &#29641. 29 – 641 est la valeur décimale d'*Unicode*. Cette mauvaise présentation du caractère est désignée par le terme *mojibake*. Cependant, le caractère 珉 *mín* s'affiche correctement sur le site de presse, codé par deux codages, *Big-5* et *GB 2312*. En effet, sur la page Web, les deux codages utilisés sont explicitement spécifiés par le biais d'une balise *HTML*. Cette balise *HTML* spécifie le codage auquel correspond le caractère, et le programme d'affichage des pages Web l'affiche correctement. Pourtant, il est mal traduit lorsqu'il passe en format texte brut (*Unicode*).

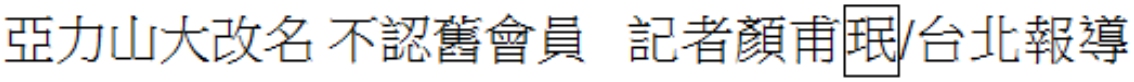


Figure 16 : Intégration du caractère 珉 *mín*

Ce phénomène de *mojibake* se produit si le système d'exploitation ne reconnaît pas le codage avec lequel ce caractère est écrit. Par conséquent, nous avons remplacé le code &#29641, par le caractère correct : 珉 *mín*, après vérification sur le site du journal électronique *United Daily News*.

4.4.2.2 Caractères présentés par des flèches

Certains caractères chinois se présentent incorrectement sous forme de flèches sur la page Web, ainsi qu'en format texte. Nous avons cherché les caractères corrects sur Internet, en confrontant plusieurs journaux et sites numériques. Puis nous avons corrigé les représentations erronées en les remplaçant par les caractères corrects, comme dans l'exemple suivant :

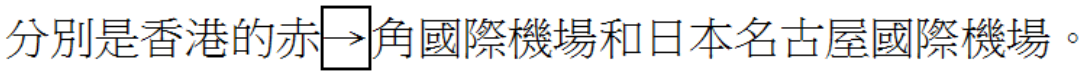


Figure 17 : Affichage d'une flèche sur le site journalistique du 27 février 2008 et sous le format texte brut

Le caractère 鱳 *liè* n'est pas intégré dans les plages de caractères chinois du codage *Big-5*, ni dans celles du *GB 2312*. Comme nous l'avons mentionné dans la section 4.1.1.1, le *Big-5* a été adopté par le gouvernement de Hong Kong afin de coder le cantonais, lequel s'écrit avec de très nombreux caractères spécifiques au cantonais. Or, il faut savoir qu'un certain nombre de ces caractères n'a pas été pris en compte par l'un ou l'autre de ces codages. Il en résulte que ce caractère 鱳 *liè* se présente sous forme de flèche sur le site

journalistique du 27 février 2008. Heureusement, *Unicode* permettant de coder les symboles, la flèche se retrouve dès lors dans le texte codé (fichier .txt).

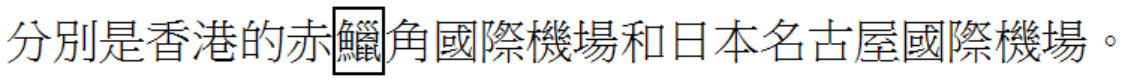


Figure 18 : Correction manuelle du caractère 鱱 *liè*

Après avoir cherché le nom exact de l'aéroport international de Hong Kong, nous avons remplacé la flèche par le caractère 鱱 *liè*, qui le désigne. Dès lors, il s'affiche ainsi dans l'article enregistré sous le format texte brut : 赤鱱角 *chìlièjiǎo*, et non par sa représentation incorrecte 赤→角 [*cf.* Figure 17].

4.4.2.3 Caractères se présentant sous forme d'image

Certains caractères chinois, identifiés comme des caractères créés, sont représentés par des images. Les images produites peuvent s'afficher coorectement sur le site journalistique. Néanmoins, ces caractères transmis en images deviennent des objets non reconnus par *Unicode*. Exemple :

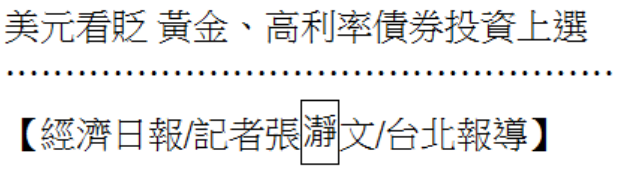


Figure 19 : Affichage du caractère 澾 *jìng* sous forme d'image sur le site journalistique du 10 février 2008

Le caractère 澾 *jìng* 'propre' étant rarement utilisé, il n'est pas pris en charge par le codage *Big-5* ni par le codage *GB 2312*. Pour être correctement représenté, il est créé et transformé en image sur le site du journal *United Daily News*. Après l'enregistrement de ce texte au format texte brut, ce caractère disparaît.

La suite de caractères, 張澾文 *Zhāng Jìngwén* représente le nom et le prénom de la journaliste. Le premier caractère est son nom de famille, l'ensemble du second et du troisième caractère forme son prénom. Représenté par une image, le second caractère va disparaître sous le format texte brut. Par conséquent, la composition originale 張澾文 *Zhāng Jìngwén* devient 張文 *Zhāng Wén*, par omission du second caractère, 澾 *jìng*, en tant qu'image sur le site journalistique.

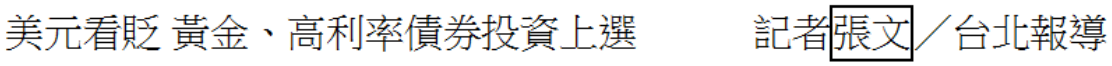


Figure 20 : Affichage du nom incorrect de la journaliste sous le format texte brut : 張文 *Zhāng Wén*

Afin d'obtenir des textes analysables, nous avons remplacé cette image, par le caractère D'*Unicode* (28 – 702) : 澾 *jìng*. Ainsi, nous avons récupéré la représentation correcte de l'ensemble du nom et du prénom de cette journaliste : 張澾文 *Zhāng Jìngwén*.

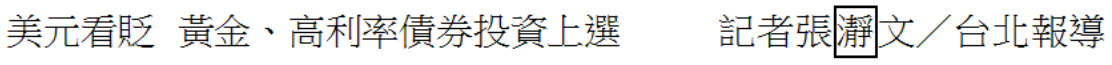


Figure 21 : Correction manuelle du caractère 澾 *jìng*

Il est indispensable que les documents de recherche soient présentés correctement. Il nous faut donc remplacer les caractères concernés par des caractères intégrés et reconnus par *Unicode*.

4.4.2.4 Caractères non pris en charge par les systèmes de codage

Plusieurs caractères chinois, dont l'apparition dépend nécessairement des logiciels d'édition, se présentent sous forme d'images. Nous les avons remplacés par l'indication <<INVALID>>. En témoigne l'exemple suivant, qui est le nom d'un personnage décrit dans le fait divers :

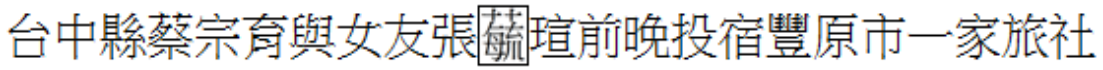


Figure 22 : Affichage de l'image 𦰡 *yù* sur le site journalistique du 7 novembre 2007

Ce caractère 𦰡 *yù* 'élever' a été créé en "serrant" ensemble deux caractères séparés : ⺾ *cǎo* 'herbe' et 毓 *yù* 'élever'. Cette opération a été rendue possible avec l'aide de certains programmes informatiques, comme *TrueType* ou 中文全字庫 *zhōngwén quánzìkù*, que nous avons déjà cités. Donc visible sous forme d'image, il disparaît lorsque le texte est transféré au format texte. Nous venons d'observer, dans la section précédente, un problème similaire pour le cas du caractère 澾 *jìng*. L'ensemble du nom et du prénom du personnage se présente ainsi : 張𦰡瑄 *Zhāng Yùxuān*.

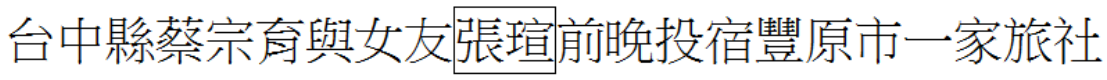


Figure 23 : Affichage incorrect du nom du personnage sous le format texte brut : 張瑄 *Zhāng Xuān*

Le deuxième caractère apparaît en tant qu'image. Il disparaît lors de l'ouverture du texte sous format texte brut : 張瑄 *Zhāng Xuān*. Puisque les textes doivent pouvoir être traités et analysés, et que les images ne sont pas dans un format pris en charge par *Unicode*, toutes les images trouvées doivent être remplacées par les caractères pris en compte par ce système, ou alors par l'indication <<INVALID>>.

台中縣蔡宗育與女友張<<INVALID>>瑄前晚投宿豐原市一家旅社

Figure 24 : Remplacement par l'indication <<INVALID>> : 張<<INVALID>>瑄

Ce caractère 蘟 *yù* n'existe pas dans les systèmes de codage actuels. Nous avons donc décidé de le remplacer par l'indication <<INVALID>>. La composition du nom et du prénom du personnage est représentée de la façon suivante : 張<<INVALID>>瑄 *Zhāng* <<INVALID>> *xuān*.

Ce choix de mode de traitement permet d'analyser un texte sans changer la syntaxe originale d'une phrase. L'indication <<INVALID>> désigne la place qu'occupait le caractère chinois manquant, ce qui facilite la disparition des ambiguïtés et la reconnaissance de la bonne structure textuelle.

4.4.2.5 Caractères inventés par les programmes

Certains caractères sont inventés en manipulant des caractères chinois déjà existants et des polices de caractères, c'est-à-dire des glyphes. Exemple :

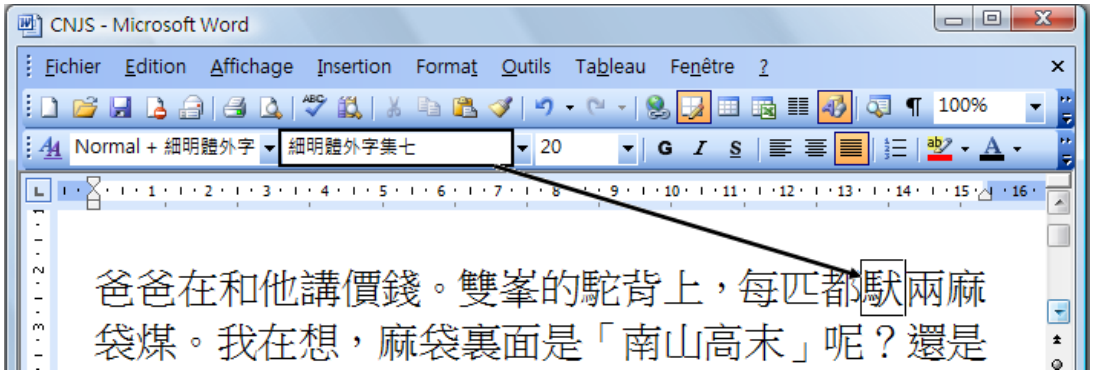


Figure 25 : Caractère inventé à travers une manipulation des fontes

Le caractère sélectionné est un caractère inventé par le programme 中央研究院漢字部件檢字系統 *zhōngyāng yánjiū / jiù yuàn hànzì bùjiàn jiǎnzì xìtǒng* « *Chinese Character Component Searching System of Academia Sinica* »[21]. Pour faire correctement apparaître un caractère ne possédant pas de code (point de code) officiel, il est nécessaire d'appliquer la police de caractères 細明體外字集七 *xìmíngtǐ wàizìjí qī*.

馱

Figure 26 : Résultat de l'invention

Le caractère 馱 *tuó* 'porter sur le dos' est une forme graphique variante du caractère 駄 obtenue en remplaçant 犬 *quǎn* 'chien' par 大 *dà* 'grand'. Il a été inventé grâce à la formule 馬⿰犬 , définie par les auteurs de ce programme, et consiste en la juxtaposition horizontale des deux caractères : 馬 *mǎ* 'cheval' et 犬 *quǎn* 'chien', qui sont deux caractères chinois. Cette combinaison de deux caractères existants est signifiée par le symbole ⿰, opérateur de concaténation mis en œuvre dans le programme « *Chinese Character Component Searching System of Academia Sinica* ». C'est une manipulation de la police pour obtenir les caractères chinois non traités par *Unicode*. Le résultat de cette manipulation est donné par la Figure 26.

Les caractères inventés grâce à une manipulation des polices ne peuvent pas apparaître sous le format texte brut, qui dépend des systèmes de codage. Nous avons décidé de remplacer le caractère inventé 馱 *tuó* par sa variante graphique alternative la plus fréquente, 駄 *tuó*, de même sens et de même prononciation. Le caractère 駄 *tuó*, quant à lui, est bien

[21] *cf.* http://cdp.sinica.edu.tw/cdphanzi/

pris en charge par le codage *Unicode*. Ce remplacement offre une bonne représentation des caractères et préserve la syntaxe de la phrase en question. En témoigne la présentation suivante :

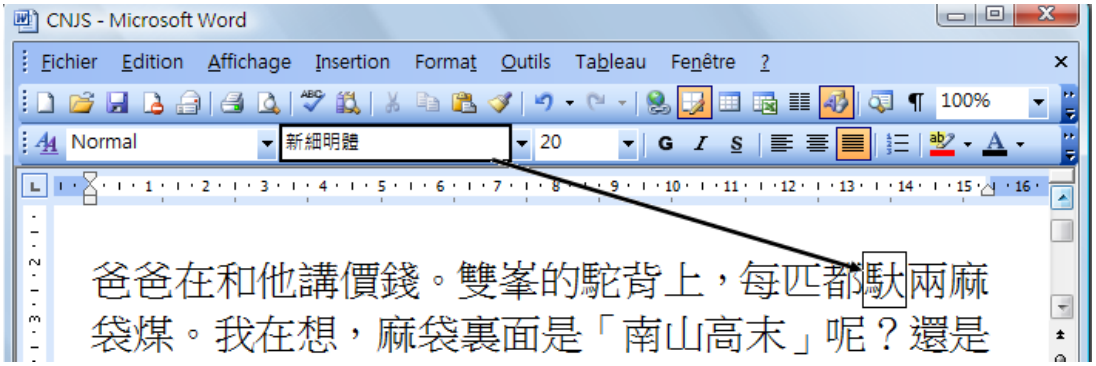


Figure 27 : Variante alternative du même caractère : 馱 *tuó*

Le caractère inventé 馱 *tuó* doit être remplacé par le caractère 馱 *tuó*, assigné dans l'implémentation d'*Unicode*, *UTF-8*. C'est la raison pour laquelle nous avons remplacé certains caractères inexistants dans *Unicode* par leur forme graphique variante, qui possède une représentation dans *Unicode*.

4.4.2.6 Conclusion

Nous avons rencontré cinq cas dans lesquels des remplacements ont été effectués, afin de rendre analysables les textes. Ils sont présentés ci-dessous :

Affichage sur textes originaux	Représentation incorrecte sous format texte brut	Résultat après correction
珉	&#29641 Le texte spécifie littéralement la valeur de codage *Unicode*.	珉
→ Le caractère apparaît sous forme de flèche sur la page Web.	→ Le caractère apparaît également sous forme de flèche sous le format texte brut.	鱷
瀞	Ce caractère sous forme d'image disparaît sous le format texte brut, car l'image n'est pas codée par les systèmes de codage.	瀞 Le caractère peut être représenté en *Unicode* (donc en *UTF-8*).
𦰡 Ce caractère a été créé en utilisant un outil qui permet de faire apparaître les caractères chinois qui ne sont pas pris en charge par le système de codage, en "serrant" ensemble deux caractères séparés : ⁺⁺ *cǎo* 'herbe' et 毓 *yù* 'élever'.	Ce caractère sous forme d'image disparaît sous le format texte brut, les systèmes de codage ne prennent pas en charge l'image.	<<INVALID>>
駄 Ce caractère est créé au moyen d'un outil informatique qui permet de faire apparaître les caractères chinois non pris en charge par les systèmes de codage. Cette opération est faite à l'aide de la manipulation de glyphes. Ainsi, ce caractère est créé à partir de deux caractères chinois placés côte à côte, 馬 *mǎ* 'cheval' et 犬 *quǎn* 'chien', sous le format .doc.	馬犬 Le caractère résultant n'est pas représentable en *Unicode / UTF-8.*	駄 Celui-ci est une variante plus fréquente autorisant une substitution "sémantiquement neutre".

4.5 Formatage des textes

Constituer un corpus en corrigeant les erreurs trouvées dans les textes, puis en formatant ceux-ci, a réclamé un travail et un temps très importants. Nous avons choisi de présenter notre corpus pour cette étude formellement sous format texte. Ce choix permettrait par la suite de transférer notre corpus au format spécifié la ***T**ext **E**ncoding **I**nitiative* (***TEI***).

Nous avons établi un système de marques pour indiquer la structure des textes :

1) La fin d'une partie est marquée par ***** ;
2) La fin d'un chapitre par ##### ;
3) La fin d'une section par §§§ ;
4) La fin d'une sous-section par ※※※.

Nous avons aussi inséré deux retours à la ligne entre chaque paragraphe, soit deux entrées sur le clavier de l'ordinateur, pour bien les séparer.

Au début de chaque texte, nous avons introduit les informations essentielles qui le concernent en les présentant de la manière suivante :

1) Dans le cas d'un texte littéraire, nous spécifions la date de la parution, le nom de l'auteur, son pseudonyme, le titre et la longueur du texte. Ainsi par exemple, pour la nouvelle *My Memories of Old Beijing*, de Lin Haiyin, nous avons précisé les informations suivantes en introduction du texte :

<Date=1960>
<Author=林含英>
<Pseudonym=林海音>
<Title=城南舊事>
<Characters=67518>

2) Dans le cas d'un texte journalistique, nous indiquons la date de sa parution, son numéro, le nom du journal et la longueur du texte.

Après correction, mise au propre, et communication des informations liées aux textes, nous les avons sauvegardés sous format texte (.txt) selon codage *UTF-8*. Les textes codés sous ce format peuvent être transportés d'une plateforme à l'autre et peuvent être traités par ***NooJ***.

4.6 Conclusion

Dans ce chapitre, nous avons présenté les différentes étapes — correction, formatage et codage de corpus — qui visent à rendre les textes analysables par un analyseur linguistique automatique tel que ***NooJ***.

Les travaux de correction permettent de faire un premier tri et d'agencer les textes numériques de façon structurée. Ils se déroulent en trois phases :

1) Collecte des textes littéraires et journalistiques à partir d'Internet ;

2) Nettoyage des fichiers numériques : l'objectif de cette deuxième phase est de corriger les fautes de frappe, les coquilles de numérisation ainsi que les défaillances de la conversion ;

3) Formatage des fichiers : lors de cette phase, nous intégrons à chaque texte les singes de séparation.

Le travail de codage permet de transporter correctement les textes d'une plateforme à une autre. Ce travail se réalise en deux étapes :

1) Lors de la première, on remplace les caractères chinois dont l'apparition, sous forme d'image, n'est pas utilisable. D'une part, certains caractères peuvent trouver leur forme graphique correcte dans *Unicode*. D'autre part, en raison de l'absence de certains caractères chinois dans les systèmes de codage actuels, quelques caractères apparaissent sous forme d'image ou à l'aide des polices de caractères spécifiques. Puisqu'il s'agit de l'enregistrement sous le format texte, nous avons choisi de substituer à ces caractères manquants l'indication <<INVALID>>.

2) La deuxième étape est consacrée au codage des textes. Dans notre corpus, tous les textes sont codés en *UTF-8*.

À partir de ces documents numériques, nous avons pu extraire des informations linguistiques qui se prêtent à la caractérisation des entrées dans les dictionnaires, et qui peuvent être appliquées à la composition de grammaires morpho-syntaxiques et syntaxiques. Nous allons expliquer en détail la constitution des dictionnaires électroniques et les descriptions morpho-syntaxiques, représentés par les graphes dans les chapitres suivants.

Chapitre 5

DEVELOPPEMENT DES DICTIONNAIRES ELECTRONIQUES

Le traitement morphologique constitue la première étape dans la gestion des données. Centré sur des informations morphologiques, il permet l'élaboration de dictionnaires électroniques. Un dictionnaire électronique peut être défini comme le regroupement des Unités Linguistiques Atomiques d'une langue. À chaque entrée sont associées des informations linguistiques exploitables par les applications du traitement automatique des textes. Le dictionnaire électronique est donc conçu pour donner des informations linguistiques. Il est également utilisé pour annoter les données et affiner leur accès.

Nous exposerons d'abord la structure interne des mots chinois et les définitions des quatre types d'Unités Linguistiques Atomiques selon ***NooJ***. Puis, nous décrirons en détail les deux types de critères que nous avons adoptés pour établir la liste des entrées. Le premier permettant de lemmatiser des caractères qui transcrivent les caractères non signifiants, les morphèmes non autonomes, les mots simples monosyllabiques, les affixes, les semi-affixes, etc. Le deuxième servant à lemmatiser les mots polysyllabiques, qu'ils soient des mots simples, des mots composés, des expressions figées, etc. Ensuite, nous expliquerons pourquoi il est si difficile de classer un mot chinois dans telle ou telle catégorie. Nous recenserons les critères utilisés par divers linguistes avant de présenter les catégories proposées par Guo Rui [2002], catégories que nous avons adoptées dans cette étude. Nous présenterons aussi les six dictionnaires essentiellement construits à partir de notre corpus. Nous finirons par la production automatique des formes réduplicatives des noms, des adjectifs, des adjectifs à valeur descriptive, des quantifieurs, etc.

5.1 La structure interne des mots chinois

5.1.1 Définition des morphèmes chinois

5.1.1.1 Morphèmes monosyllabiques et morphèmes polysyllabiques

En chinois moderne, les morphèmes ne sont pas toujours monosyllabiques. Ils peuvent aussi être dissyllabiques ou polysyllabiques.

5.1.1.1.1 Morphèmes monosyllabiques

Les morphèmes monosyllabiques s'écrivent avec un seul caractère et se prononcent en une seule syllabe. Ils se divisent en deux groupes, selon qu'ils sont ou non autonomes [*cf.* 5.1.1.2]. Citons quelques exemples :

● Morphèmes monosyllabiques autonomes :

(14) 水 *shuǐ* 'eau'
海 *hǎi* 'mer'
高 *gāo* 'grand'

● Morphèmes monosyllabiques non autonomes :

(15) 匯 *huì* 'confluer'
希 *xī* 'rare ; espérer'
視 *shì* 'regarder ; examiner'

5.1.1.1.2 Morphèmes polysyllabiques

Les morphèmes polysyllabiques s'écrivent avec deux ou plusieurs caractères correspondant chacun à une unité syllabique. Appelés polysyllabes monomorphémiques ils se classent en quatre groupes :

1) Les morphèmes dits « mots formés de deux syllabes liées » (en chinois 聯綿詞 *liánmiáncí*). Ils sont de trois types :

1.1) Les morphèmes dissyllabiques débutant par une même consonne qui suivent des voyelles différentes. Exemples :

(16) 倉促 *cāngcù* 'à la hâte'
參差 *cēncī* 'irrégulier'
坎坷 *kǎnkě* 'hérissé de difficultés'
彷彿 *fǎngfú* 'semblable ; sembler'
琉璃 *liúlí* 'lazurite'
蜘蛛 *zhīzhū* 'araignée'

1.2) Les morphèmes dissyllabiques dont les deux syllabes ont la même voyelle mais qui se différencient par leurs consonnes. Exemples :

(17) 從容 *cōngróng* 'tout doucement'
徘徊 *páihuái* 'aller et venir : hésiter à avancer'
玫瑰 *méiguī* 'rose'
窈窕 *yǎotiǎo* 'jeune fille belle et vertueuse'
蜻蜓 *qīngtíng* 'libellule'
駱駝 *luòtuó* 'chameau'

1.3) Les morphèmes dissyllabiques qui n'ont ni voyelles ni consonnes communes. Exemples :

(18) 玻璃 *bōlí* 'verre'
芙蓉 *fúróng* 'lotus des Indes'
蝴蝶 *húdié* 'papillon'
鷓鴣 *zhègū* 'francolin perlé'
蜈蚣 *wúgōng* 'scolopendre'

Les caractères qui servent à écrire ces morphèmes dissyllabiques ne représentent que des syllabes. Il ne s'agit nullement de l'association de deux morphèmes monosyllabiques chacun doté de sens. Par exemple, *fú* 芙 et *róng* 蓉 ne sont que les deux syllabes d'un seul morphème *fúróng* 芙蓉 qui signifie 'lotus des Indes'. Pris individuellement, ces caractères n'ont aucun sens.

2) Les morphèmes issus de la réduplication. Par exemple,

(19) 依依 *yīyī* 'inséparable'
姍姍 *shānshān* '(démarche) souple et lente'
惶惶 *huánghuáng* 'agité et apeuré'
楚楚 *chǔchǔ* 'brillant ; beau'
茫茫 *mángmáng* 'immense'
遲遲 *chíchí* 'tard ; hésiter'
關關 *guānguān* 'onomatopée imitant les cris d'oiseaux qui s'appellent et se répondent'

3) Les morphèmes provenant de la transposition phonologique de mots étrangers [*cf.* Zhitang Yang-Drocourt, 2007 : 251-255]. On les classe en trois groupes, en fonction de la méthode employée pour modéliser phonologiquement les mots étrangers syllabe par syllabe.

3.1) Les morphèmes issus de la transposition phonologique. Ce sont des morphèmes polysyllabiques qui résultent de l'imitation phonologique de mots étrangers. Les caractères qui les composent ne renvoient à aucune signification et tentent seulement de transposer au plus près la prononciation en mandarin du mot étranger. Exemples :

(20) 咖啡 *kāfēi* 'café'
沙發 *shāfā* 'sofa'
三明治 *sānmíngzhì* 'sandwich'
巧克力 *qiǎokèlì* 'chocolat'

歇斯底里 *xiēsīdǐlǐ* ‘hystérie’

3.2) Les morphèmes issus de la transposition phonologique et de l’arrangement sémantique. Les graphies de ces polysyllabes servent à la fois à transposer la prononciation de mots étrangers et à rendre compte de leur signification. À l’écrit, on cherche à adopter des graphies (caractères) qui permettent d’expliquer le sens de ces mots, en même temps qu’elles transposent leur prononciation. Par exemple,

(21) 基因 *jīyīn* <essence-cause> ‘gène’
繃带 *bēngdài* <serrer-cordon> ‘bandage’
俱樂部 *jùlèbù* <tout-joyeux-groupe> ‘club’
維他命 *wéitāmìng* <maintenir-il-vie> ‘vitamine’
烏托邦 *wūtuōbāng* <soleil-supporter-pays> ‘utopie’

3.3) Les morphèmes formés à la fois par une transposition phonologique et par l’ajout d’un hyperonyme chinois. La première partie de ces morphèmes transpose les syllabes des mots étrangers. La deuxième partie constitue un hyperonyme qui permet de préciser le sens d’un mot traduit en chinois. Exemples :

(22) 啤酒 *píjiǔ* <bière-vin> ‘bière’
羽毛球 *yǔmáoqiú* <plume-balle> ‘badminton’
芭蕾舞 *bāléiwǔ* <ballet-danse> ‘ballet’
霓虹燈 *níhóngdēng* <néon-lumière> ‘enseigne lumineuse au néon’
摩托車 *mótuōchē* <moto-véhicule> ‘motocyclette’

4) Les onomatopées polysyllabiques. Ce sont des morphèmes qui essaient d’imiter des sons de la nature ou des bruits. Exemples :

(23) 噗通 *pūtōng* ‘plouf se référant au bruit d’un objet qui tombe dans l’eau’
嘩啦啦 *huā--lā--lā* ‘onomatopée du bruit d’un liquide — pluie ou eau du robinet — qui tombe sur une surface solide’
淅瀝嘩啦 *xīlì-huālā* ‘onomatopée du bruit de la pluie qui tombe sur le sol’

5.1.1.2 Morphèmes autonomes et morphèmes non autonomes

5.1.1.2.1 Morphèmes autonomes

Les morphèmes qui ont une liberté syntaxique et s’utilisent en tant que mots sont dits autonomes. Ces morphèmes peuvent former d’autres mots en s’associant à d’autres morphèmes autonomes ou non autonomes. Certains peuvent également constituer des phrases. La plupart des morphèmes autonomes sont monosyllabiques, par exemple, 人 *rén*

‘personne’, 才 *cái* ‘seulement’, 手 *shǒu* ‘main’, 我 *wǒ* ‘je’, 書 *shū* ‘livre’, 能 *néng* ‘pouvoir’, 累 *lèi* ‘fatigué’, 就 *jiù* ‘justement’, 筆 *bǐ* ‘stylo’, 說 *shuō* ‘parler’, etc.

Tout morphème polysyllabique est un morphème autonome et s’utilise en tant que mot [*cf.* 5.1.1.1.2].

5.1.1.2.2 Morphèmes non autonomes

Bien que doués d’une autonomie syllabique, graphique et sémantique, les morphèmes non autonomes ne peuvent être utilisés tels quels, en tant que mots, dans une phrase. Ils doivent se combiner avec d’autres morphèmes, autonomes ou non autonomes, pour former des mots. Exemples :

(24) *shū + jí = shūjí*
書 + 籍 = 書籍
livre + cahier = livre

(25) *jí + guàn = jíguàn*
籍 + 貫 = 籍貫
registre + habiter un même lieu de père en fils = lieu d’origine

(26) *dǎ + zì = dǎzì*
打 + 字 = 打字
taper + mot = taper (un texte)

(27) *zì + diǎn = zìdiǎn*
字 + 典 = 字典
mot + citation = dictionnaire

Dans l’exemple (25), le nom *shūjí* 書籍 se compose des morphèmes *shū* 書 et *jí* 籍. Le premier, *shū* 書 est autonome. Par contre, le deuxième, *jí* 籍 est non autonome. Le morphème non autonome peut être associé à un autre morphème non autonome, pour former le nom *jíguàn* 籍貫 [*cf.* (26)]. Dans l’exemple (27), le mot *dǎzì* 打字 est constitué d’un morphème autonome *dǎ* 打 et d’un morphème non autonome *zì* 字. Celui-ci peut, à son tour, s’adjoindre un autre morphème non autonome *diǎn* 典 pour former le nom *zìdiǎn* 字典 [*cf.* (28)].

5.1.2 Discussion sur les affixations chinoises

Contrairement aux langues indo-européennes la langue chinoise ne compte que peu d'affixes [*cf*. Li et Thompson, 1989 : 36]. Cependant, comment distinguer les affixes des morphèmes non autonomes ? Pour identifier les affixes, les linguistes ont adopté les critères suivants [*cf*. Fu Huaiqing, 1985 [2005 : 32-34] ; Ge Benyi, 2001 [2002 : 56-60] ; Sun Yinxin, 2003 : 237-254 et Yang Xipeng, 2003 : 108-118] :

1) Un morphème chinois qui devient affixe perd son sens originel. Ce sens est alors délexicalisé. Un mot dérivé par l'ajout d'un affixe conserve donc sa signification originale. Le mot 竹子 *zhúzi* <bambou-K> 'bambou' dont la dérivation est réalisée à l'aide du suffixe 子 *zi* conserve le sens du morphème 竹 *zhú*, mais l'affixe 子 *zi* perd le sens d''enfant' qu'il avait en tant que morphème 子 *zǐ*. On pourrait encore citer 員 *yuán*, 家 *jiā*, 性 *xìng* ou 化 *huà*. La confusion repose sur leur apparente ressemblance aux morphèmes 員 *yuán* 'membre' dans 成員 *chéngyuán* <devenir-membre> 'membre', 家 *jiā* 'maison' dans 搬家 *bānjiā* <déplacer-maison> 'déménager', 性 *xìng* 'nature ; caractère' dans 屬性 *shǔxìng* <appartenir à-caractère> 'attribut' ou 化 *huà* 'se transformer' dans 變化 *biànhuà* <changement-se transformer> 'changement ; transformation'. Ces quatre morphèmes ont un sens particulier, lorsqu'ils entrent dans la constitution des mots. Néanmoins, lorsqu'ils se transforment en affixes, ils perdent le sens qu'ils avaient en tant que morphèmes. C'est le cas de 員 *yuán* dans 指導員 *zhǐdǎoyuán* <guider-K> 'guide ; entraîneur', de 家 *jiā* dans 藝術家 *yìshùjiā* <art-K> 'artiste', de 性 *xìng* dans 感受性 *gǎnshòuxìng* <être impressionné-K> 'sensitivité' ou de 化 *huà* dans 工業化 *gōngyèhuà* <induistrie-K> 'industrialiser ; industrialisation'.

2) La position d'un affixe dans le mot est fixée une fois pour toutes. S'il est préfixe, il ne pourra jamais être utilisé comme suffixe et inversement.

3) Un suffixe indique la catégorie ou la classe sémantique d'un mot dérivé. Ainsi, les mots dérivés à l'aide du suffixe 家 *jiā* tels que 鋼琴家 *gāngqínjiā* <piano-K> 'pianiste' ou 舞蹈家 *wǔdǎojiā* <danse-K> 'danseur', sont des noms qui désignent des êtres humains.

Bien que ces critères soient reconnus, il est encore difficile de les appliquer aux cas réels. Il n'existe toujours pas de liste bien précise d'affixes chinois et leur nombre varie suivant les auteurs. Il y a à cela deux raisons.

D'une part, au fil de l'évolution du chinois, certains morphèmes se sont transformés en affixes et leur sens s'est alors délexicalisé. Pourtant, tous les linguistes ne s'accordent pas là-dessus : certains d'entre eux, qui ne les analysent pas comme des affixes, pensent que leur sens n'est pas totalement délexicalisé lors de la constitution de mots. En d'autres termes, chaque linguiste possède un critère sémantique différent pour mesurer la délexicalisation du morphème lorsqu'il est utilisé en tant qu'affixe. Les morphèmes concernés sont par exemple, 化 *huà*, 性 *xìng*, 員 *yuán*, 師 *shī* ou 手 *shǒu*. Selon les critères sémantiques adoptés par les linguistes, ils sont considérés soit comme des morphèmes, soit comme des affixes nouveaux en chinois moderne [*cf.* Ge Benyi, 2001 [2002 : 56-60], Packard, 2000 [2006 : 69-73] et Yang Xipeng, 2003 : 108-118].

D'autre part, en raison de la forte production de néologismes, un certain nombre de morphèmes jouent un rôle important dans le lexique du chinois moderne. Citons 超 *chāo*, 新 *xīn*, 多 *duō*, 反 *fǎn*, 非 *fēi*, 零 *líng*, 壇 *tán*, 熱 *rè*, 感 *gǎn*, 戶 *hù*, 型 *xíng*, 制 *zhì* ou 族 *zú*. À l'inverse des affixes proprement dits, ces morphèmes conservent une autonomie morphologique et sémantique. En d'autres termes, on peut les considérer comme des composants morphologiques. De plus, leur signification n'est pas complètement délexicalisée quand on les utilise pour former des mots. Compte tenu de leur nature linguistique, certains linguistes les traitent comme des semi-affixes [*cf.* Sun Yinxin, 2003 : 237-254 et Cao Wei, 2003 [2004 : 81-82]].

Nous avons d'abord étudié les travaux portant sur les affixes et les semi-affixes chinois de linguistes tels que Wang Li [1943], Lü Shuxiang [1942], Chao Yuen Ren [1968], Ge Benyi [2001], Packard [2000] ou Yang Xipeng [2003]. Nous avons, pour étayer notre étude, opté prudemment pour une liste plus neutre. Ainsi, nous avons adopté les critères et la liste d'affixes et de semi-affixes proposés par Yang Xipeng [2003].

Ci-dessous, nous présenterons quelques affixes et semi-affixes figurant dans cette liste.

5.1.3 Affixes

5.1.3.1 Préfixes

● 老 *lǎo*, 小 *xiǎo* et 阿 *ā*

老 *lǎo* et 小 *xiǎo* servent à former des mots dont la plupart sont des noms d'êtres humains, d'objets ou d'animaux. Le premier, 老 *lǎo*, signifie 'vieux' ; le second, 小 *xiǎo*, 'jeune ; petit'. Par ailleurs, 阿 *ā* est utilisé pour formuler des appellations personnelles. Du point de vue lexical, ce sont des morphèmes porteurs de sens. Exemples :

(28) 老師 *lǎoshī* <H-maître> 'professeur'
老鼠 *lǎoshǔ* <H-souris> 'souris'
小姐 *xiǎojiě* <H-sœur> 'mademoiselle'
小孩 *xiǎohái* <H-enfant> 'enfant'
阿伯 *ābó* <H-frère aîné du père> 'frère aîné du père ; appellation pour s'adresser à un monsieur âgé'
阿姨 *āyí* <H-tante> 'tante maternelle ; appellation pour s'adresser aux femmes approximativement du même âge que la mère du locuteur'

Dès que ces trois morphèmes deviennent des affixes, ils sont employés pour former des appellatifs familiers ou des surnoms. Exemples :

(29) 老五 *Lǎo Wǔ* <H-cinq> 'Lao Wu : la cinquième personne dans le système hiérarchique'
老陳 *Lǎo Chén* <H-Chen> 'Lao Chen : surnom d'une personne supérieure ou plus âgée dont le nom de famille est 陳 *Chén*'
小周 *Xiǎo Zhōu* <H-Zhou> 'Xiao Zhou : surnom d'une personne inférieure ou plus jeune dont le nom de famille est 周 *Zhōu*'
小玲 *Xiǎo Líng* <H-Ling> 'Xiao Ling : surnom d'une personne dont le prénom est 玲 *líng*'
阿美 *Ā Měi* <H-Mei> 'A Mei : surnom d'une personne dont le prénom est 美 *měi*'

Les appellatifs composés à l'aide du préfixe 老 *lǎo* se réfèrent à des personnes qui occupent une position supérieure au sein d'un système hiérarchique. À l'inverse, les appellatifs ou les surnoms composés avec le préfixe 小 *xiǎo* servent à nommer des personnes occupant un rang inférieur.

Le préfixe 老 *lǎo* peut être associé aux chiffres (à partir de deux) pour exprimer une hiérarchie. Ainsi, le deuxième enfant d'une famille est appelé 老二 *Lǎo Èr* <H-deux> 'Lao Er : deuxième'. Mais l'aîné, qui ne peut être appelé *老一 *Lǎo Yī* <H-un> 一 *yī* signifiant

'un', est appelé 老大 *Lǎo Dà* <H-grand> 'Lao Da : premier', 大 *dà* ayant le sens de 'grand'.

● 第 *dì*

Le préfixe 第 *dì* se place devant un numéral pour former un ordinal. Exemples :

(30) 第一 *dì-yī* <H-un> 'premier'
第二 *dì-èr* <H-deux> 'deuxième'

● 初 *chū*

Le préfixe 初 *chū* se place devant les chiffres de un à dix pour numéroter les dix premiers jours de chaque mois du calendrier lunaire. Exemples :

(31) 初一 *chūyī* <H-un> 'premier jour'
初二 *chūèr* <H-deux> 'deuxième jour'

5.1.3.2 Suffixes

● 子 *zi*

Le suffixe 子 *zi* est, à l'origine, un morphème 子 qui signifie 'enfant' et se prononce *zǐ*. Cependant son sens propre disparaît lorsqu'il est employé en dernière syllabe en tant que suffixe de ton neutre pour former des mots nominaux [*cf.* Li et Thompson, 1989 : 42-43]. Exemples :

■ Employé avec un ou deux morphèmes de nature nominale :
(32) 竹子 *zhúzi* <bambou-K> 'bambou'
影子 *yǐngzi* <ombre-K> 'ombre'
鏡子 *jìngzi* <miroir-K> 'miroir'
兔崽子 *tùzǎizi* <lapin-garçon-K> 'coquin ; moutard'
狗腿子 *gǒutuǐzi* <chien-jambe-K> 'patte de chien : valet ; vendu'
樹腰子 *shùyāozi* <arbre-région lombaire-K> 'projectile en bois'

■ Avec un morphème de nature adjectivale :
(33) 瘦子 *shòuzi* <maigre-K> 'une personne maigre'
瞎子 *xiāzi* <aveugle-K> 'un aveugle'
聾子 *lóngzi* <sourd-K> 'un sourd'
胖子 *pàngzi* <gros-K> 'une personne corpulente'

■ Avec un morphème de nature verbale :

(34) 擔子 *dànzi* <supporter-K> ‘fardeau’
釘子 *dīngzi* <clouer-K> ‘clou’
鎚子 *chuízi* <marteler-K> ‘marteau’

Ce suffixe sert également à former des quantifieurs pour la plupart temporels :

(35) 會子 *huǐzi* <Q-K>
股子 *gǔzi* <Q-K>
陣子 *zhènzi* <Q-K>

● 兒 *r*

Le suffixe 兒 *r* provient du morphème 兒 *ér* qui a le sens d’‘enfant’. Ce suffixe sert à former des mots nominaux nouveaux à partir de morphèmes dont les propriétés diffèrent. Exemples :

■ Avec un morphème non autonome de nature nominale :

(36) 李兒 *lǐr* <prune-K> ‘prune’
桃兒 *táor* <pêche-K> ‘pêche’

■ Avec un nom (morphème autonome de nature nominale) :

(37) 刀兒 *dāor* <couteau-K> ‘couteau’
心兒 *xīnr* <cœur-K> ‘cœur’

■ Avec un adjectif :

(38) 尖兒 *jiānr* <aigu-K> ‘pointe’
甜兒 *tiánr* <sucre-K> ‘sucrerie’

■ Avec un verbe :

(39) 吃兒 *chīr* <manger-K> ‘nourriture’
畫兒 *huàr* <peindre-K> ‘peinture’

Ce suffixe peut aussi être ajouté à d’autres mots sans en changer leur catégorie grammaticale d’origine. Exemples :

■ Avec un verbe :

(40) 玩兒 *wánr* <jouer-K> ‘jouer’

■ Avec un quantifieur :

(41) 朵兒 *duǒr* <Q-K> ‘quantifieur réservé aux fleurs’
會兒 *huǐr* <Q-K> ‘quantifieur réservé au temps’

■ Avec un adverbe :

(42) 悄悄兒 *qiāoqiāor* <silencieusement-K> ‘silencieusement’

Ce suffixe est donc considéré comme un diminutif principalement réservé à des noms, bien qu’il soit possible de le combiner avec des verbes, des quantifieurs ou des adverbes.

Non-syllabique, employé dans le dialecte pékinois, le suffixe 兒 *r* sert à former des syllabes finissant par une prononciation rétroflexe. Au lieu de dire 今日 *jīnrì* <aujourd’hui-jour> ‘aujourd’hui’, les pékinois utilisent le terme 今兒 *jīnr* <aujourd’hui-K>. De la même façon, ils utilisent le terme 這兒 *zhèr* <ce-K>, au lieu de 這裡 *zhèlǐ* <ce-place> pour préciser ‘cette place-ci’. [*cf*. Zhu Dexi, 1982 [2004 : 30-31]].

● 頭 *tou*

Le suffixe 頭 *tou* est un suffixe de ton neutre [*cf*. Li et Thompson, 1989 : 43-44]. C’est, à l’origine, un morphème monosyllabique nominal 頭 *tóu*, signifiant ‘tête’. On le trouve associé à des noms ou à des mots locatifs. Il entre également dans la constitution de noms abstraits dont les morphèmes initiaux peuvent être des adjectifs ou des verbes. Exemples :

■ Avec un nom :

(43) 木頭 *mùtou* <bois-K> ‘morceau de bois’
石頭 *shítou* <pierre-K> ‘pierre’
舌頭 *shétou* <langue-K> ‘langue’
芋頭 *yùtou* <taro-K> ‘tubercule de taro’
枕頭 *zhěntou* <oreiller-K> ‘oreiller’
饅頭 *mántou* <petit pain rond cuit à la vapeur-K> ‘petit pain rond cuit à la vapeur’

■ Avec un locatif :

(44) 下頭 *xiàtou* <dessous-K> ‘dessous’
上頭 *shàngtou* <dessus-K> ‘dessus’
外頭 *wàtou* <extérieur-K> ‘à l’extérieur de’
前頭 *qiántou* <avant-K> ‘avant ; devant’
後頭 *hòutou* <après-K> ‘après ; derrière’
裡頭 *lǐtou* <intérieur-K> ‘à l’intérieur de’

■ Dans la constitution de noms abstraits :

♦ Avec un morphème initial nominal :

(45) 苦頭 *kǔtou* <amer-K> ‘difficulté ; désavantage’
甜頭 *tiántou* <sucré-K> ‘bénéfice ; avantage’

♦ Avec un morphème initial verbal :

(46) 吃頭 *chītou* <manger-K> ‘valoir la peine d’être mangé’
看頭 *kàntou* <regarder-K> ‘chose qui mérite d’être vue, intéressante à voir’
聽頭 *tīngtou* <écouter-K> ‘chose qui mérite d’être entendue’

● 們 *men*

Le suffixe 們 *men* sert exclusivement à former le pluriel de noms ou de pronoms désignant des êtres humains [*cf*. Sun Chaofen, 2006 : 64]. Exemples :

■ Noms d’êtres humains :

(47) 員工們 *yuángōngmen* <employé-K> ‘employés’
孩子們 *háizimen* <enfant-K> ‘enfants’
學生們 *xuéshēngmen* <étudiant-K> ‘étudiants’
朋友們 *péngyǒumen* <ami-K> ‘amis’
老師們 *lǎoshīmen* <professeur-K> ‘professeurs’

■ Pronoms :

(48) 他們 *tāmen* <il-K> ‘ils’,
你們 *nǐmen* <tu-K> ‘vous’
我們 *wǒmen* <je-K> ‘nous’

● 化 *huà*

Le suffixe 化 *huà* souligne un processus d’évolution. Par exemple,

(49) 老化 *lǎohuà* <vieux-K> ‘vieillir’
西化 *xīhuà* <ouest-K> ‘occidentaliser ; occidentalisation’
美化 *měihuà* <beau-K> ‘rendre élégant’
現代化 *xiàndàihuà* <moderne-K> ‘moderniser ; modernisation’
都市化 *dūshìhuà* <ville-K> ‘urbaniser ; urbanisation’
電腦化 *diànnǎohuà* <ordinateur-K> ‘informatiser, informatisation’

● 員 *yuán*

Le suffixe 員 *yuán* entre dans la dérivation de mots qui définissent la profession ou la position d’une personne. Par exemple :

(50) 店員 *diànyuán* <magasin-K> ‘employé de magasin’
教員 *jiàoyuán* <enseigner-K> ‘professeur’
學員 *xuéyuán* <apprendre-K> ‘étudiant’
研究員 *yánjiū / jiù yuán* <effectuer des recherches-K> ‘chercheur’
採購員 *cǎigòuyuán* <faire des achats d’articles-K> ‘acheteur’

運動員 *yùndòngyuán* <sport-K> ‘athlète’

5.1.4 Semi-affixes

5.1.4.1 Semi-préfixes

● 多 *duō*

Le semi-préfixe 多 *duō* introduit l’idée de ‘multi’. Par exemple,

(51) 多媒體 *duōméitǐ* <SH-média> ‘multi-médias’
多目標 *duōmùbiāo* <SH-but> ‘multi-objectifs’

● 半 *bàn*

Le semi-préfixe 半 *bàn* renvoie à la signification de ‘semi’. Par exemple,

(52) 半成品 *bànchéngpǐn* <SH-produit> ‘produit semi-fini’
半導體 *bàndǎotǐ* <SH-conducteur> ‘semi-conducteur’

● 輕 *qīng*

Le semi-préfixe 輕 *qīng* renvoie à la signification de ‘léger’. Par exemple,

(53) 輕工業 *qīnggōngyè* <SH-industrie> ‘industrie légère’
輕音樂 *qīngyīnyuè* <SH-musique> ‘musique légère’

● 重 *zhòng*

Le semi-préfixe 重 *zhòng* renvoie à la signification de ‘lourd’. Par exemple,

(54) 重金屬 *zhòngjīnshǔ* <SH-métal> ‘métal lourd’
重機械 *zhòngjīxiè* <SH-machine> ‘machine lourde’

5.1.4.2 Semi-suffixes

● 學 *xué*

On adjoint le semi-suffixe 學 *xué* aux mots qui désignent les différents niveaux du système éducatif ou encore des noms de disciplines académiques. Exemples :

■ Niveaux du système d'éducation :

(55) 大學 *dàxué* <grand-SK> 'université'
小學 *xiǎoxué* <petit-SK> 'école primaire'
中學 *zhōngxué* <moyen-SK> 'école secondaire ; collège et lycée'

■ Disciplines académiques :

(56) 文學 *wénxué* <composition littéraire-SK> 'études de littérature'
社會學 *shèhuìxué* <société-SK> 'sociologie'
語言學 *yǔyánxué* <langue-SK> 'linguistique'

● 度 *dù*

Le semi-suffixe 度 *dù* souligne différents degrés de mesure. Par exemple,

(57) 尺度 *chǐdù* <règle-SK> 'degré d'acceptation'
高度 *gāodù* <haut-SK> 'hauteur'
速度 *sùdù* <rapide-SK> 'vitesse'
知名度 *zhīmíngdù* <connu-SK> 'degré de réputation'
能見度 *néngjiàndù* <pouvoir apercevoir-SK> 'visibilité'
酸鹼度 *suānjiǎndù* <acide et alcali-SK> 'degré de distinction entre acidité et alcalinité'

5.1.5 Mots simples

Un mot simple correspond à un seul morphème autonome, qu'il soit monosyllabique ou polysyllabique. Les mots tels que 樹 *shù* 'arbre', 說 *shuō* 'parler' ou 高 *gāo* 'grand' sont des mots simples monosyllabiques. Les mots simples monosyllabiques constituent la base du vocabulaire chinois. Ils s'utilisent fréquemment [*cf*. Cao Wei, 2003 [2004 : 57-64]]. Néanmoins, ils forment une minorité au sein du lexique chinois [*cf*. Zhitang Yang-Drocourt, 2007 : 224].

Certains mots simples sont des polysyllabes monomorphémiques. En voici quelques exemples :

(58) Mots simples dissyllabiques : 葡萄 *pútáo* 'raisin' ou 星星 *xīngxing* 'étoile';

Mots simples trisyllabiques : 巧 克 力 *qiǎokèlì* 'chocolat' ou 轟隆隆 *hōng--lóng--lóng* 'grondement' ;

Mots simples quadrisyllabiques : 歇 斯 底 里 *xiēsīdǐlǐ* 'hystérie' ou 嘰嘰喳喳 *jījī-zhāzhā* 'cri-cri'.

5.1.6 Mots composés

En français, un mot composé est un mot constitué d'au moins deux mots simples. Par exemple, « pomme de terre » est un mot composé de trois morphèmes qui correspondent aux trois mots simples : « pomme », « de » et « terre ». En chinois, la plupart des mots se définissent comme des mots "composés" de morphèmes, comme l'exprime le terme chinois 複合詞 *fùhécí* qui signifie qu'une unité lexicale est formée de différents morphèmes. De ce point de vue, les mots composés chinois sont constitués plutôt de morphèmes que de mots simples. Selon Packard, les mots composés chinois peuvent être classés en quatre types [*cf.* Packard, 2000 [2006 :67-79]] :

(59)	morphème autonome + morphème autonome	河馬 *hémǎ* <rivière-cheval> 'hippopotame'
	morphème autonome + morphème non autonome	家庭 *jiātíng* <famille-cour> 'famille'
	morphème non autonome + morphème autonome	橡樹 *xiàngshù* <chêne-arbre> 'chêne'
	morphème non autonome + morphème non autonome	研究 *yánjiū / jiù* <étudier-examiner à fond> 'effectuer des recherches'

5.1.7 Expressions figées

Le lexique chinois comporte un certain nombre d'expressions figées, chacune composée de plusieurs mots. Douées de liberté syntaxique, elles peuvent être utilisées en tant que mots. Elles sont traitées comme des expressions phraséologiques, car elles possèdent une signification complète et ont une structure stable. En chinois, il existe quatre types d'expressions figées : les locutions, les formules quadrisyllabiques, les proverbes et les expressions à double volet [*cf.* Zhitang Yang-Drocourt, 2007 : 256-261].

5.1.7.1 Locutions

Les locutions possèdent un délpacement sémantique. On les désigne en chinois par le terme 慣用語 *guànyòngyǔ*. Elles sont pour la plupart composées de trois syllabes :

(60) 下馬威 *xiàmǎwēi* <descendre-cheval-majestueux>
'(faire) montrer de sévérité à la descente de cheval : à l'entrée en fonction (pour asseoir son autorité) ; vouloir s'imposer à peine arrivé'

(61) 戴高帽 *dàigāomào* <porter-haut-chapeau>
'recevoir des louanges ; se faire encenser'

(62) 敲竹槓 *qiāozhúgàn* <frapper-bambou-levier>
‘extorquer quelque chose par chantage ou menace’

Une petite partie des locutions est formée d’au moins quatre syllabes :

(63) 捅馬蜂窩 *tǒng mǎfēngwō*
<creuser-grande guêpe-guêpier>
‘se fourrer dans un guêpier : s’attirer des ennuis’
(64) 皮笑肉不笑 *píxiào ròu bù xiào*
<peau-rire-chair-ne pas-rire>
‘rieur en apparence seulement : dissimulé’
(65) 穿新鞋走老路 *chuān xīnxié zǒu lǎolù*
<porter-nouveau-chaussure-marcher-vieux-rue>
‘changer l’enveloppe mais pas le contenu : illusion de changement’

La structure des locutions est plus libre que celle des formules quadrisyllabiques, que nous présenterons dans la section suivante. Les locutions se présentent comme des séquences discontinues de mots. On peut modifier leur structure en remplaçant un de leur composant morphologique par un autre, ou en y intégrant d’autres morphèmes :

1) La locution 扯後腿 *chěhòutuǐ* <tirer à soi-pattes de derrière> ‘entraver ; être un boulet au pied de’ peut être transformée en 拉後腿 *lāhòutuǐ* ou 拖後腿 *tuōhòutuǐ* en substituant le morphème 扯 *chě* ‘tirer à soi’ par 拉 *lā* ‘traîner’ ou 拖 *tuō* ‘impliquer’.

2) Il est possible d’ajouter le suffixe 子 *zi* à la fin de la locution (62) 戴高帽 *dàigāomào.* On obtiendra une nouvelle forme de cette locution 戴高帽子 *dài gāomàozi* <porter-haut-chapeau-K> qui garde sa signification d’origine.

3) Les locutions peuvent également constituer la base d’une phrase complète. Il est possible de former une phrase, à partir de la locution (62) :

小明給老陳戴了頂高帽。 *Xiǎo Míng gěi Lǎo Chén dàile dǐng gāomào.*
<Xiao Ming-à-Lao-Chen-porter-Le-Q-haut-chapeau.>
‘Xiao Ming a flatté Lao Chen.’

5.1.7.2 Formules quadrisyllabiques

Figées, les formules quadrisyllabiques offrent un sens complet. À la différence des locutions, elles sont définies comme des mots de quatre syllabes.

La signification de certaines formules quadrisyllabiques se réfère à leurs composants morphologiques, par exemple,

(66) 兩全其美 *liǎngquánqíměi* <deux-satisfaire-son-intérêt>
‘donner satisfaction à l’un et à l’autre côté’

(67) 百發百中 *bǎifābǎizhòng* <cent-lancer (une flèche)-cent-attendre le but>
‘mettre dans le mille à tout coup : succès’

(68) 門庭若市 *méntíngruòshì* <cour-comme-marché>
‘la cour de la maison ressemble à un marché : on entre dans cette maison comme dans un moulin ; maison très fréquentée ; affluence de visiteurs’

Contrairement aux exemples mentionnés ci-dessus, nombre de formules quadrisyllabiques ont un sens figuré. Exemples :

(69) 曇花一現 *tánhuāyīxiàn* <fleur du figuier-soudain-apparaître>
‘la fleur du figuier n’apparaît qu’un moment : fugace comme une fleur éphémère’

(70) 枯木逢春 *kūmùféngchūn* <sec-bois-rencontrer-printemps>
‘un arbre desséché rencontre le printemps : revirement subit de fortune’

(71) 胸有成竹 *xiōngyǒuchéngzhú* <poitrine-avoir-réaliser-bambou>
‘avoir des bambous réalisés dans l’esprit : avoir un plan bien organisé dans la tête’

De nombreuses formules quadrisyllabiques proviennent de textes anciens : ouvrages classiques, histoires classiques, fables, légendes, etc., dans lesquels ils puisent leur signification. Exemples :

(72) 四面楚歌 *sìmiànchǔgē* <quatre-côté-chants du royaume de Chu>
‘les chants du royaume de Chu retentissent de tous côtés : être entouré d’ennemis de tous côtés’

(73) 夸父追日 *kuāfùzhuīrì* <Kuafu-courir-soleil>
‘Kuafu poursuit le soleil : entreprendre une chose au-dessus de ses forces’

(74) 完璧歸趙 *wánbìguīzhào* <complet-tablette de jade-retourner-pays Zhao>
‘la tablette de jade est rendue intacte au pays Zhao : objet restitué fidèlement à son propriétaire légitime’

(75) 愚公移山 *yúgōngyíshān* <Yugong-déplacer-montagne>
‘le vieux sot (Yugong) déplace la montagne : avec du temps et de la patience on arrive à tout’

(76) 揠苗助長 *yàmiáo-zhùzhǎng* <arracher-jeune pousse-aider-agrandir>
‘tirer sur les jeunes pousses pour les aider à croître : tout gâter en voulant forcer la nature’

(77) 見異思遷 *jiànyì-sīqiān* <apercevoir-différent-penser-déménager>
‘en voyant quelque chose de différent, penser changer : être inconstant ; l’herbe est toujours plus verte ailleurs’

5.1.7.3 Proverbes

Les proverbes sont des expressions qui, à l'origine, se transmettent oralement. Ils sont formés d'une ou deux phrases et expriment un conseil ou un jugement. Exemples :

(78) 慢工出細活 *màngōng chū xìhuó*
<doux-travail-offrir-fin-produit>
'le travail lent fait la belle œuvre'

(79) 千金難買寸光陰 *qiānjīn nánmǎi cùn guāngyīn*
<mille-or-difficile-acheter-pouce-temps>
'un pouce d'or ne peut acheter un pouce de temps : le temps est plus précieux que l'or'

(80) 留的青山在，不怕沒柴燒。 *liú de qīngshān zài, bù pà méi chái shāo.*
<réserver-De-vert-montagne-se trouver à, ne pas-craindre-manquer-bois-brûler.>
'tant qu'il y aura des montagnes vertes, on ne craindra pas de manquer de bois de chauffage : tant qu'il y de la vie, il y a de l'espoir.'

5.1.7.4 Expressions à double volet

Les expressions à double volet sont aussi des expressions populaires. Leur signification est figée. Elles se composent d'une amorce et d'une chute. L'amorce exprime le sens figuré ou une métaphore. La chute doit montrer la signification attendue du sens figuré ou de la métaphore. S'enchaînant, ces deux éléments forment une unité sémantique entière. Exemples :

(81) 啞巴吃黃連—有苦難言。 *yǎba chī huánglián — yǒukǔ-nányán.*
<muet-manger-gentiane — avoir-amertume-difficile-parler>
'en mangeant de la gentiane, le muet n'en peut dire l'amertume : ne pouvoir exprimer son amertume intérieure'

(82) 肉包子打狗—有去無回。 *ròubāozi dǎ gǒu — yǒuqù-wúhuí.*
<viande-pain cuit à la vapeur-frapper-chien — avoir-partir-sans-retourner>
'On bat un chien avec un pain farci cuit à la vapeur — faire quelque chose sans espoir de résultat'

Lorsque le sens apparaît dès l'amorce, la chute est omise. Exemples :

(83) 癩蛤蟆想吃天鵝肉。 *làiháma xiǎng chī tiānéròu.*
<crapaud-vouloir-manger-cygne-viande.>
'le crapaud rêve de manger de la viande de cygne.'

L'amorce révélant déjà la chute 'prétention ridicule', celle-ci 異想天開 *yìxiǎng-tiānkāi* <imaginer-ciel-ouvrir> 'prétention ridicule' est sous-entendue.

(84) 黃鼠狼給雞拜年。 *huángshǔláng gěi jī bàinián.*
<belette-à-poule-souhaiter la bonne année.>
'la belette vient souhaiter la bonne année aux poules.'

La signification 'ne pas avoir de bonnes intentions' apparaissant dès cette amorce, la chute 沒安好心 *méi ān hǎoxīn* <ne pas-avoir-bon-cœur> 'ne pas avoir de bonnes intentions' peut être omise.

5.2 Définition des Unités Linguistiques Atomiques selon *NooJ*

Selon ***NooJ***, l'Unité Linguistique Atomique (***A**tomic **L**inguistic **U**nit*, ou ***ALU***) est, dans une langue, la plus petite unité porteuse de sens. Ces unités sont regroupées et associées à des informations linguistiques dans des dictionnaires électroniques. ***NooJ*** classe les Unités Linguistiques Atomiques en quatre classes formelles distinctes. Les **affixes**, les **mots simples** et les **mots composés** sont des séquences insécables, constituées d'un ou de plusieurs morphèmes ou caractères. À l'inverse, les **expressions figées** peuvent être discontinues. Nous présentons ici les quatre types d'Unités Linguistiques Atomiques définis par ***NooJ***, en les illustrant par des exemples :

1) **Affixes** : Ce sont des séquences de lettres non délimitées. Elles se présentent comme des composants morphologiques, qui interviennent dans des opérations lexicales. Elles sont associées aux informations morphologiques. Ce sont, par exemple, les préfixes « anti– », « bi– » ou « re– » ou les suffixes « –able », « –isme » ou « –ment » en français ; ou les préfixes 小 *xiǎo*, 老 *lǎo* ou 阿 *ā* ou les suffixes 子 *zi*, 兒 *r* ou 頭 *tou* en chinois.

2) **Mots simples** : Ce sont des séquences de lettres entre deux séparateurs qui contiennent des informations linguistiques, par exemple, « aisé », « chaise », « emballer », « formation », « natif » ou « pouvoir » en français ; ou 人 *rén* 'personne', 能 *néng* 'pouvoir', 筆 *bǐ* 'stylo', 葡萄 *pútáo* 'raisin', 星星 *xīngxing* 'étoile' ou 巧克力 *qiǎokèlì* 'chocolat' en chinois.

3) **Mots composés** : Ce sont des séquences de lettres avec séparateurs qui doivent être lemmatisées, comme « chemin de fer », « pomme de terre » ou « sac à main » en français ; ou 家庭 *jiātíng* 'famille', 橡樹 *xiàngshù* 'chêne', 圓桌會議 *yuánzhuō*

huìyì ‘réunion autour d’une table ronde’ ou 電視節目 *diànshì jiémù* <télévision-programme> ‘programmes de chaines télévisées’ en chinois.

4) **Expressions figées** : Ce sont des séquences de morphèmes ou de mots potentiellement discontinues, par exemple, « avoir faim », « prendre en compte » ou « tenir au courant » en français ; ou 戴高帽 *dàigāomào* ‘se faire encenser’, 門庭若市 *méntíngruòshì* ‘affluence de visiteurs’, 慢工出細活 *màngōng chū xìhuó* ‘le travail lent fait la belle œuvre’ ou 癩蛤蟆想吃天鵝肉—異想天開。 *làiháma xiǎng chī tiānéròu — yìxiǎng-tiānkāi*. ‘le crapaud rêve de manger de la viande de cygne — prétention ridicule.’ en chinois.

5.3 Formalisation des Unités Linguistiques Atomiques chinoises avec *NooJ*

5.3.1 Quatre programmes pour la morphologique lexicale

NooJ formalise les quatre types d’Unités Linguistiques Atomiques selon quatre méthodes :

1) Les éléments morphologiques délimités peuvent être traités par le programme “affixes”. Ces éléments morphologiques concernent les affixes, les semi-affixes, les caractères non signifiants, les morphèmes non autonomes, les mots simples monosyllabiques, etc. en chinois. De plus, ce programme permet également de traiter les prénoms et les noms de famille chinois dont chaque composant est représenté par un caractère.

2) Le programme “mots simples” permet de traiter les mots simples polysyllabiques chinois.

3) Le programme “mots composés” s’utilise pour traiter les mots composés définis par le terme chinois 複合詞 *fùhécí*, ou les mots constitués d’au moins deux mots composés (mentionnés précédemment).

4) Le programme “expressions figées” sert à traiter les expressions figées décrites en 5.1.7.

Ces quatre programmes peuvent donc analyser toutes les unités atomiques dans les textes rédigés en chinois moderne.

5.3.2 Le traitement informatique du lexique dans *NooJ*

Pour analyser une séquence de caractères, ***NooJ*** propose deux méthodes :

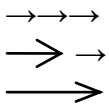
1) Chaque composant de cette séquence de caractères est traité comme une unité atomique ;

2) La séquence de caractères tout entière est traitée comme une unique unité atomique.

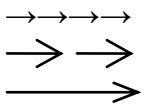
Pour plus de clarté, prenons l'exemple de trois unités lexicales chinoises : 翻譯 *fānyì* 'traduire', 翻譯員 *fānyìyuán* 'traducteur' et 翻譯系統 *fānyì xìtǒng* 'système de traduction'. ***NooJ*** analyse chacune de ces unités lexicales selon les deux méthodes mentionnées qu'on peut schématiser comme suit :

1) 翻譯


Cette séquence de caractères 翻譯 *fānyì* est analysée comme pouvant contenir soit l'unité linguistique atomique 翻譯, soit les deux unités atomiques : 翻 et 譯.

2) 翻譯員


Si une séquence de caractères contient également une autre, mais plus courte, ***NooJ*** la reconnaît aussi comme étant une unité atomique. Ainsi, ***NooJ*** reconnait dans la séquence de caractères 翻譯員 *fānyìyuán* cinq unités atomiques : 翻, 譯, 員, 翻譯 et 翻譯員, selon trois analyses possibles : l'unité atomique 翻譯員, ou alors les trois unités 翻, 譯, 員, ou alors l'unité 翻譯 suivie par l'unité 員.

3) 翻譯系統


Cette séquence de caractères 翻譯系統 *fānyì xìtǒng* peut être analysée de cinq façons :

-- l'unité atomique 翻譯系統

-- l'unité atomique 翻譯 suivie par l'unité atomique 系統

-- l'unité atomique 翻譯 suivie par les deux unités 系 puis par 統

-- l'unité 翻 suivie par l'unité 譯 puis par l'unité 系統

-- la séquence des quatre unités 翻, 譯, 系 puis 統.

Ce système permet de représenter toutes les ambiguïtés lexicales possibles. Pour définir une séquence de caractères comme étant une Unité Linguistique Atomique, on doit la décrire explicitement dans un dictionnaire. On peut cependant bloquer à volonté certaines de ces analyses en associant à une unité linguistique le trait **+UNAMB** : ce trait donne à ***NooJ*** l'ordre de ne pas produire d'autres analyses. Ainsi, dans le dictionnaire, si l'on représente les trois unités lexicales précédentes de la façon suivante :

翻譯,V+UNAMB
翻譯員,N+UNAMB
翻譯系統,N+UNAMB

NooJ ne produira qu'une seule analyse pour ces séquences de caractères : 翻譯 *fānyì*, 翻譯員 *fānyì yuán* et 翻譯系統 *fānyì xìtǒng*.

5.4 Critères de lemmatisation des caractères uniques

Les affixes, les semi-affixes, les caractères non signifiantes, les morphèmes non autonomes, les mots simples monosyllabiques, etc. s'écrivent tous avec un seul caractère. Par ailleurs, même s'ils ont plusieurs composants, les prénoms et les noms de famille chinois s'écrivent également avec des caractères chinois. Dans cette étude, nous les distinguerons graphiquement en leur accordant à chacun une entrée et une étiquette. Ils se présenteront donc de la façon suivante :

1) Les affixes et les semi-affixes :

(85) Les préfixes recevront l'étiquette <H> :
第,H *dì*
老,H *lǎo*

(86) Les suffixes recevront l'étiquette <K> :

子,K *zi*
兒,K *r*

(87) Les semi-préfixes recevront l'étiquette <SH> :
半,SH *bàn*
多,SH *duō*

(88) Les semi-suffixes recevront l'étiquette <SK> :
學,SK *xué*
度,SK *dù*

2) Les caractères non signifiants, qui servent à former des onomatopées, recevront l'étiquette <OG> :

(89) 嘩,OG *huā*
噗,OG *pū*
啦,OG *lā*

3) Les caractères non signifiants, qui servent à transposer les syllabes de mots étrangers, recevront l'étiquette <PHON> :

(90) 巧,PHON *qiǎo*
琵,PHON *pí*
芭,PHON *bā*

4) Les morphèmes non autonomes recevront différentes étiquettes selon leur propriété sémantique, par exemple, <NG> étiquettera les morphèmes non autonomes de nature nominale, <VG> les morphèmes non autonomes de nature verbale, <AG> les morphèmes non autonomes de nature adjectivale, etc. :

(91) 程,NG *chéng* 'mesure'
閱,VG *yuè* 'examiner'
善,AG *shàn* 'bienveillant'

5) Les mots simples monosyllabiques recevront diverses étiquettes selon leur catégorie :

(92) 山,N *shān* 'montagne'
看,V *kàn* 'regarder'
貴,A *guì* 'cher'

6) Les composants des prénoms seront étiquetés <NPRG> :

(93) 宜,NPRG *yí*
漢,NPRG *hàn*
蘭,NPRG *lán*

7) Les noms de famille monosyllabiques recevront l'étiquette <NPL> :

(94) 何,NPL *Hé*
孫,NPL *Sūn*
高,NPL *Gāo*

Parmi ces unités linguistiques du chinois moderne, qui s'écrivent avec un seul caractère, on rencontre des homonymes homographes, qui peuvent être ou non homophones [*cf*. 6.1]. S'il s'agit bien d'homonymes homographes, nous les distinguerons par des entrées différentes, chacune assortie d'une étiquette particulière. Voir les exemples plus loin.

Considérons, par exemple, les morphèmes 相 *xiāng* et 相 *xiàng*. Ce sont des homonymes homographes et non pas homophones, car ils représentent deux morphèmes non autonomes, distingués par leur prononciation *xiāng* ou *xiàng*. En tant que morphème non autonome de nature nominale, il sert à former le mot 丞相 *chéngxiàng* <premier ministre-administrateur> 'premier ministre'. En tant que morphème non autonome, de nature verbale, il entre dans la composition de certains verbes, comme 相似 *xiāngsì* <mutuel-ressembler> 'ressembler', 相信 *xiāngxìn* <mutuel-croire> 'croire', 相傳 *xiāngchuán* <mutuel-transmettre> 'transmettre' ou 相親 *xiàngqīn* <observer-relatif> 'rendez-vous qu'on prend pour faire la connaissance d'une personne éventuellement destinée à devenir un époux ou une épouse'. Il se prononce *xiāng* dans la plupart des cas comme on le constate par les trois premiers verbes donnés en exemples. Pourtant, il se prononce *xiàn* dans un petit nombre de cas, comme dans 相親 *xiàngqīn*. Les deux morphèmes qui s'écrivent avec la même graphie 相 seront différenciés comme suit dans le dictionnaire ***ChDic*** :

相,NG
相,VG

Voici quelques exemples d'homonymes :

1) Homonymes homographes et homophones :

(95) 耐,NG *nài* 'patience'
耐,VG *nài* 'résister à'
耐,AG *nài* 'patient'
(96) 耗,NG *hào* 'nouvelle'
耗,VG *hào* 'dépenser'

2) Homonymes homographes :

(97) 更,VG *gēng* ‘changer’
更,DG *gèng* ‘encore’
更,NG *jīng* ‘veille (chacune des cinq veilles de la nuit, de deux heures chacune, de dix-neuf heures à cinq heures)’

(98) 還,VG *huán* ‘retourner ; rendre’
還,DG *hái* ‘encore’

Dans les exemples qui précèdent, l’homonymie se manifeste toujours au niveau morphologique. Cependant l’homonymie peut se manifester dans différents niveaux de langue. Dans ce cas-là, nous traitons également les homonymes en les distinguant par des entrées et des étiquettes, comme nous l’avons décrit ci-dessus. Par exemple, le caractère 力 *lì* peut être un composant morphologique du mot 力圖 *lìtú* <insister sur-projeter> ‘tendre à’ et il possède alors le sens d’‘insister sur’. Mais ce même caractère 力 *lì* peut aussi être une des graphies utilisées pour transposer la prononciation du mot anglais « *chocolate* » dans l’écriture chinoise : 巧克力 *qiǎokèlì* ‘chocolat’. Ce caractère est donc un homonyme homographe. Par conséquent, nous lui accorderons deux entrées dont chacune recevra une étiquette différente dans le dictionnaire ***ChDic*** :

力,PHON
力,VG

Voici quelques exemples supplémentaires :

1) Homonymes homographes et homophones :

(99) 歇,PHON *xiē*
歇,VG *xiē* ‘se reposer’

(100) 水,NG *shuǐ* ‘liquide’
水,NPRG *shuǐ*
水,N *shuǐ* ‘eau’

2) Homonymes homographes :

(101) 王,NG *wáng* ‘roi’
王,VG *wàng* ‘gouverner’
王,NPL *Wáng*

(102) 朝,NG *zhāo* ou *cháo* ‘matin ; audience impériale’
朝,VG *cháo* ‘rendre visite’
朝,NPRG *zhāo* ou *cháo*

5.5 Critères de lemmatisation des mots polysyllabiques

Par “polysyllabiques”, nous entendons les polysyllabes monomorphémiques, les mots composés, les mots super-composés ou les quatre sortes d’expressions figées. Bien qu’ils s’écrivent tous avec au moins deux caractères, leur statut diffère selon la nature de leur composition morphologique. Nous présentons ci-dessous les critères qui nous ont guidée dans la lemmatisation de ces types de mots. Ils sont similaires aux mots composés de M. Silberztein [*cf.* 1993 : 405-426].

5.5.1 Compositionnalité

On ne peut pas décomposer les polysyllabes monomorphémiques en composants morphologiques, car les caractères avec lesquels ils s’écrivent représentent des syllabes et non pas des morphèmes. De ce point de vue, on doit les traiter comme des unités atomiques, i. e. Unités Linguistiques Atomiques et les décrire par une entrée lexicale particulière. Considérons les mots ci-dessous :

(103) 琵琶 *pípá* ‘luth à quatre cordes’
巧克力 *qiǎokèlì* ‘chocolat’
嘩啦啦 *huā--lā--lā* ‘onomatopée du bruit d’un liquide — pluie ou eau du robinet — qui tombe sur une surface solide’

Ces trois mots polysyllabiques sont formés d’unités syllabiques qui ne sont jamais utilisés de façon autonome. Le seul morphème 琵琶 *pípá* associe à deux syllabes écrites avec deux caracètres 琵 *pí* et 琶 *pá*. La forme graphie 琵 *pí* n’a pas le sens elle-même. Il faut se composer ave la graphie 琶 *pá* pour former le mot 琵琶 *pípá*. Issu de la transposition phonologique, le mot 巧克力 *qiǎokèlì* est formé graphiquement de trois syllabes représentées par les trois caractères 巧 *qiǎo*, 克 *kè* et 力 *lì*. L’onomatopée 嘩啦啦 *huā--lā--lā* se compose de deux caractères, 嘩 *huā* et 啦 *lā*, qui sont censés imiter le bruit d’un liquide qui tombe sur une surface solide. De plus, le caractère 啦 *lā* se répète une fois lors de la constitution de cette onomatopée.

Certains mots peuvent être constitués de morphèmes susceptibles d’être employés indépendamment. En d’autres termes, ces composants peuvent être des morphèmes ou des mots. On peut citer :

(104) 冰糖 *bīngtáng* <glace-sucre> ‘sucre candi’

白菜 *báicài* <blanc-légume> ‘chou chinois’
紅包 *hóngbāo* <rouge-emballer> ‘enveloppe rouge’
馬尾 *mǎwěi* <cheval-queue> ‘queue de cheval’
龍頭 *lóngtóu* <dragon-tête> ‘chef d’une communauté’

Le mot 白菜 *báicài*, par exemple, est constitué de deux morphèmes 白 *bái* et 菜 *cài*. Le sens du mot 白菜 *báicài* n’est pas déterminé par la combinaison du sens de ses deux composants morphologiques. En conséquence, les mots de ce type doivent être lemmatisés, si l’on veut faire connaître leur sens et bloquer une analyse compositionnelle, qui aboutirait à une inexactitude telle que *‘légume blanc’, qui, lui, doit être écrit 白的菜 *bái de cài*.

5.5.2 Institutionnalisation

Certains concepts ou objets sont nommés systématiquement de façon arbitraire. Par exemple, le concept ‘cœur sensible’ s’exprime en chinois par le mot 豆腐心 *dòufuxīn*, littéralement ‘cœur en tofu’. Mais d’autres expressions pourtant très semblables, telles que *牛奶心 *niúnǎixīn* ‘cœur en lait’ ou *白紙心 *báizhǐxīn* ‘cœur en papier’ ne pourraient pas être utilisées pour exprimer ce concept. Citons encore quelques exemples :

(105) 蓮花指 *liánhuāzhǐ* <lotus-doigt> ‘manière délicate et souvent très féminine de bouger les doigts’
珍珠米 *zhēnzhūmǐ* <perle-riz> ‘maïs’
門檻精 *ménkǎnjīng* <seuil-esprit> ‘rusé’
紙老虎 *zhǐlǎohǔ* <papier-tigre> ‘une personne n’ayant que les apparences de la force’
鐵公雞 *tiěgōngjī* <fer-coq> ‘une personne avare’
黃臉婆 *huángliǎnpó* <jaune-visage-vieille femme> ‘épouse d’une personne’

L’association de morphèmes peut devenir un figement. Elle ne sont pas morphologiquement différentes de formes potentielles qui ne sont jamais utilisées par les locuteurs chinois. Il faut donc les distinguer formellement les unes des autres.

5.5.3 Structuration des mots composés et super-composés

Si certaines des propriétés syntaxiques ou sémantiques d’un mot chinois ne peuvent pas être définies à partir des propriétés de ses composants, il est nécessaire de lemmatiser ce mot en le traitant en bloc. Comme on le voit ci-après, en effet, de nombreux mots composés chinois se construisent selon quelques schémas productifs.

5.5.3.1 Mots composés avec des affixes ou des semi-affixes

Nombre de mots du vocabulaire chinois sont combinés de cette manière. Ainsi, le suffixe 化 *huà*, combiné avec un groupe de morphèmes, peut former des mots composés selon la définition qu'en donne la linguistique chinoise, par exemple,

(106) 老化 *lǎohuà* <vieux-K> 'vieillir'
西化 *xīhuà* <ouest-K> 'occidentaliser ; occidentalisation'
美化 *měihuà* <beau-K> 'rendre élégant'
現代化 *xiàndàihuà* <moderne-K> 'moderniser ; modernisation'
都市化 *dūshìhuà* <ville-K> 'urbaniser ; urbanisation'
電腦化 *diànnǎohuà* <ordinateur-K> 'informatiser, informatisation'

Il existe d'autres suffixes, tels que 員 *yuán*, 家 *jiā*, 子 *zi*, 兒 *r*, etc., qui s'utilisent avec d'autres morphèmes. Nous avons lemmatisé cette catégorie de mots dans le dictionnaire principal ***ChDic***. Par exemple, 服務員 *fúwùyuán* 'serveur', 鋼琴家 *gāngqínjiā* 'pianiste', 工程師 *gōngchéngshī* 'ingénieur', 竹子 *zhúzi* 'bambou' ou 李兒 *lǐr* 'prune'.

On ne peut pas déduire le sens d'un mot à partir de l'ensemble de sens de ses composants. Ainsi, le mot 鋼琴家 *gāngqínjiā* 'pianiste' ne signifie pas *'maison de piano'. De même que le sens 'pianiste' ne peut pas être rendu par d'autres compositions telles que *鋼琴兒 *gāngqínr*, *鋼琴員 *gāngqínyuán* ou *鋼琴者 *gāngqínzhě*.

5.5.3.2 Mots polymères

Les polymères chinois sont construits par des mécanismes relevant de la coordination [*cf*. Cao Wei, 2003 [2004 : 117-120]]. Les composants morphologiques constituant un mot polymère ne peuvent pas être remplacés par d'autres composants. Citons trois sortes de polymères :

(107) 酸甜苦辣 *suāntián kǔlà*
<aigre-doux-amer-piquant>
'aigre, doux, amer, piquant : douceurs et amertumes de la vie'

Ici, les quatre composants morphologiques appartiennent au même champ sémantique. Chacun d'eux est un morphème autonome qui peut être utilisé librement. L'ensemble de ces quatre morphèmes forme un mot polymère qui évoque des saveurs alimentaires diverses, et présente un sens figuré.

(108) 兄弟姊妹 *xiōngdì zǐ / jiě mèi*
<grand frère-petit frère-grande sœur-petite sœur>
‘grand frère, petit frère, grande sœur, petite sœur : frères et sœurs’

Ce mot désigne l’ensemble des enfants d’une même famille. Les quatre morphèmes sont regroupés par paires : 兄 *xiōng* et 弟 *dì* ; 姊 *zǐ / jiě* et 妹 *mèi*. Ces deux paires sont des mots composés dissyllabiques. Leur apposition crée un mot dit polymère.

(109) 金銀珠寶 *jīnyín zhūbǎo*
<or-argent-bijou>
‘or, argent, bijou : trésor’

Ce mot polymère se compose de deux parties. La première comprend deux mots simples monosyllabiques 金 *jīn* et 銀 *yín*, qui relèvent de la même classe sémantique, celle des métaux précieux. Ils qualifient le mot composé dissyllabique 珠寶 *zhūbǎo*, qui est la deuxième partie de ce polymère.

En conclusion, les mots polymères doivent être recensés dans le dictionnaire, quelle que soit leur composition. Ils sont traités comme les autres mots composés en tant qu’Unités Linguistiques Atomiques.

5.5.3.3 Mots composés de structure XY

Certains mots chinois sont composés de deux mots composés dissyllabiques **X** et **Y**, et peuvent être considérés comme des mots “super-composés” [*cf.* Cao Wei, 2003 [2004 : 120-124]]. Traités comme des mots en bloc, ils doivent, eux aussi, être décrits par une entrée dans le dictionnaire. Citons :

(110) 香車寶馬 *xiāngchē bǎomǎ*
<voiture parfumée-cheval de prix>
‘voiture parfumée, cheval de prix : bel attelage’

Les deux composants, 香車 *xiāngchē* et 寶馬 *bǎomǎ*, sont deux mots composés dissyllabiques. Ils appartiennent à deux catégories sémantiques différentes : le premier 香車 *xiāngchē* désigne un moyen de transport, le deuxième 寶馬 *bǎomǎ* un animal. Combinés sans conjonction ils forment le mot 香車寶馬 *xiāngchē bǎomǎ*.

(111) 公子哥兒 *gōngzǐ gēr*
<jeune homme-garçon>
‘jeune homme, garçon : fils de riche’

Ce mot contient deux composants morphologiques, 公子 *gōngzǐ* et 哥兒 *gēr*. Séparés l'un de l'autre, aucun des deux n'a le sens de 'fils de riche'.

(112) 小橋流水 *xiǎoqiáo liúshuǐ*
<petit pont-eau courante>
'petit pont, eau courante : un beau paysage'

Ce mot est composé de deux mots composés dissyllabiques 小橋 *xiǎoqiáo* et 流水 *liúshuǐ*. La formation de ces deux mots composés donne un nouveau mot 小橋流水 *xiǎoqiáo liúshuǐ* qui possède une nouvelle signification : un beau paysage.

5.5.3.4 Mots composés de structure AXBY

En chinois moderne, un certain nombre d'unités lexicales sont constituées selon la structure **AXBY** [*cf*. Cao Wei, 2003 [2004 : 120-124]]. Citons quelques exemples :

七 X 八 Y *qī* X *bā* Y 'sept X huit Y'
上 X 下 Y *shàng* X *xià* Y 'dessus X dessous Y'
三 X 五 Y *sān* X *wǔ* Y 'trios X cinq Y'
四 X 五 Y *sì* X *wǔ* Y 'quatre X cinq Y'
左 X 右 Y *zuǒ* X *yòu* X 'gauche X droite Y'
好 X 歹 Y *hǎo* X *dǎi* Y 'bon X mauvais Y'
東 X 西 Y *dōng* X *xī* Y 'est X ouest Y'
前 X 後 Y *qián* X *hòu* Y 'devant X derrière Y'

Les morphèmes **A** et **B** sont des mots qui sémantiquement, appartiennent à la même catégorie, par exemple, 三 *sān* 'trois', 五 *wǔ* 'cinq', 七 *qī* 'sept' ou 八 *bā* 'huit'. Ils peuvent être des antonymes tels que 左 *zuǒ* 'gauche', 右 *yòu* 'droite', 前 *qián* 'devant' ou 後 *hòu* 'derrière'. Les morphèmes **X** et **Y** se combinent de trois manières différentes :

- **X** et **Y** constituent ensemble un mot ;
- **X** et **Y** ne forment pas un mot. Ce sont des morphèmes indépendants ;
- **X** est le même morphème que **Y**.

Donnons trois exemples :

(113) 東張西望 *dōngzhāng xīwàng*
<est-regarder au loin-ouest-regarder au loin>
'regarder à gauche et à droite : regarder de tous côtés'

Le morphème 東 *dōng* est sémantiquement opposé au morphème 西 *xī*. Les deux morphèmes 張 *zhāng* et 望 *wàng* se combinent pour former le verbe 張望 *zhāngwàng* <étendre-regarder au loin> ‘regarder tout autour de soi’.

(114) 左顧右盼 *zuǒgù yòupàn*
<gauche-regarder-droite-regarder>
‘regarder à gauche et à droite : laisser errer ses regards’

Les morphèmes 左 *zuǒ* et 右 *yòu* sont des antonymes. Les morphèmes 顧 *gù* et 盼 *pàn* sont deux morphèmes de nature verbale. Ils ne forment pas un mot.

(115) 好說歹說 *hǎoshuō dǎishuō*
<bon-parler-mauvais-parler>
‘pratiquer la menace et la flatterie pour remporter l’adhésion’

Le morphème 好 *hǎo* est l’antonyme du morphème 歹 *dǎi*. Le verbe 說 *shuō* est à la fois **X** et **Y** lors de la construction du mot 好說歹說 *hǎoshuō dǎishuō*.

5.5.4 Expressions figées

Les expressions figées regroupent les locutions, les formules quadrisyllabiques, les proverbes ou les expressions à double volet. Leur structure est plus ou moins régulière. Elles peuvent être formalisées avec ***NooJ*** de deux façons complémentaires :

1) lorsqu’il s’agit d’expressions insécables, elles peuvent être rangées dans le dictionnaire, tout comme les mots composés ;

2) lorsqu’il s’agit d’expressions qui admettent des insertions comme, en français, l’expression « prendre ... en compte », elles doivent être traitées, d’une part, dans le dictionnaire, d’autre part, dans des grammaires syntaxiques locales. Par exemple, dans l’expression suivante :

(116) 坐冷板凳 *zuò lěngbǎndèng*
<s’asseoir-froid-banc>
‘s’asseoir sur un banc froid : moisir dans un poste quelconque ; avoir un emploi mal rétribué’

Les constituants ne sont pas forcément juxtaposés, comme on peut le voir dans le texte suivant :

(117) 他已坐過了冷板凳。 *tā yǐ **zuò**guòle **lěngbǎndèng**.*
<il-déjà-s'asseoir-U-U-froid-banc>
'Il a déjà moisi dans un poste quelconque.'

Par ailleurs, certaines expressions sécables admettent des variantes morphologiques qui peuvent être représentées dans les dictionnaires de ***NooJ***. Ainsi les trois locutions synonymes suivantes qui toutes signifient 'entraîner quelqu'un dans l'eau : entraîner quelqu'un dans une mauvaise affaire' :

(118) 拖下水 *tuōxiàshuǐ* <implanter-descendre-eau>
(119) 拉下水 *lāxiàshuǐ* <tra înner-descendre-eau>
(120) 扯下水 *chěxiàshuǐ* <tirer à soi-descendre-eau>

Le verbe 拖 *tuō* peut être remplacé par les verbes, 拉 *lā* et 扯 *chě*, sans aucun changement de sens. Donc, chacune de ces trois locutions doit être traitée, dans le dictionnaire ***ChDic***, comme une entrée lexicale particulière.

5.6 Les catégories en chinois

5.6.1 Difficultés d'identification des catégories de mots

En chinois, il est difficile de connaître la catégorie d'une unité lexicale. Les formes, qu'elles soient graphiques ou orales, ne proposent pas d'indications en ce sens. Ainsi l'unité lexicale telle qu'elle apparaît sous la forme graphique 乾脆 *gāncuì* dans le roman de Lao She *Quatre générations sous un même toit* :

(121) 城外頭，**乾脆**沒人管事兒啦！
*chéng wàitou, **gāncuì** méi rén guǎn shìr lā !*
<ville-extérieur-K, **vraiment**-sans-personne-diriger-affaire-K-E !>
'À l'extérieur de la ville, personne ne s'occupe **vraiment** des affaires !'
(122) 「好啦！瑞宣！再見！我喜歡你這麼**乾脆**嘹亮，西洋派兒！」
*「 hǎolā ! Ruìxuān ! zàijiàn ! wǒ xǐhuān nǐ zhème **gāncuì** liáoliàng, xīyángpàir !」*
<« bon-E ! Ruixuan ! au revoir ! je-aimer-tu-si-**net**-clair, occident-branche-K ! »>
'« C'est bon ! Ruixuan ! Au revoir ! J'aime que tu sois **net** et clair comme les Occidentaux ! »'

L'unité lexicale *gāncuì* 乾脆 est adverbe dans la phrase (122), mais adjectif dans la phrase (123). Cette unité lexicale peut donc appartenir à des catégories différentes (adjectif ou adverbe), bien que sa graphie ne propose aucun moyen de les distinguer.

En chinois, une unité lexicale peut être aussi bien un nom qu'un verbe, un adjectif qu'un adverbe, un adjectif qu'un nom. Ce phénomène fréquent est inhérent à la nature de la langue chinoise. Exemples :

(123) a. 比賽游泳 ***bǐsài*** *yóuyǒng* <**concourir**-nager>
'concourir pour la natation : natation de compétition'
b. 演講**比賽** *yǎnjiǎng* ***bǐsài*** <discours-**concours**>
'concours de discours'

(124) a. **負責**採購 ***fùzé*** *cǎigòu* <**occuper**-faire des courses>
'être responsable d'achat d'articles'
b. 小王很**負責**。 *Xiǎo Wáng hěn* ***fùzé***. <Xiao Wang-très-**responsable**.>
'Xiao Wang est très responsable.'

(125) a. 這消息很**意外**。*zhè xiāoxí hěn* ***yìwài***. <ce-nouvelle-très-**surprenant**.>
'Cette nouvelle est très surprenante.'
b. 交通**意外** *jiāotōng* ***yìwài*** <transport-**accident**>
'accident de voiture'

L'unité lexicale *bǐsài* 比賽 est un verbe ou un nom ; l'unité lexicale *fùzé* 負責 un verbe ou un adjectif ; et l'unité lexicale *yìwài* 意外 un adjectif ou un nom. Chacune de ces unités lexicales relève de deux catégories différentes.

Néanmoins, il existe un nombre d'unités lexicales dont les catégories demeurent incertaines. En effet, une unité lexicale, qu'elle soit verbe, adjectif ou adverbe, n'apparaît toujours que sous une seule forme. Il y a à cela des raisons. Les verbes chinois n'ont pas de formes flexionnelles, et rien ne distingue la personne, le temps ou le mode. Il n'existe pas non plus de formes dérivationnelles qui permettraient de distinguer les adjectifs des adverbes ou les noms des adjectifs, etc. Seule la syntaxe des phrases permet de faire ces distinctions. Par ailleurs, bien que les unités lexicales occupent des fonctions syntaxiques diverses, certaines unités ne sont incontestablement que des verbes [*cf*. Guo Rui, 2002]. Ainsi dans les exemples suivants :

(126) a. 他負責學術書籍的**出版**。
tā fùzé xuéshù shūjí de ***chūbǎn***.
<il-occuper-discipline scientifique-ouvrage-De-**éditer**.>
'Il édite des ouvrages scientifiques.'

b. 他已經負責**出版**工作多年。
*tā yǐjīng fùzé **chūbǎn** gōngzuò duōnián.*
<il-déjà-occuper-**éditer**-travail-nombreux-année.>
'Il est responsable de l'édition depuis plusieurs années.'

c. 他最近又**出版**了一本重要的學術著作。
*tā zuìjìn yòu **chūbǎn**le yī běn zhòngyào de xuéshù zhùzuò.*
<il-récemment-encore-**éditer**-Le-un-Q-important-De-discipline scientifique-œuvre.>
'Il a récemment édité un autre ouvrage scientifique important.'

(127) 政府出版品的**訂購**可申請補助。
*zhèngfǔ chūbǎnpǐn de **dìnggòu** kě shēnqǐng bǔzhù.*
<gouvernement-publication-De-**commander**-pouvoir-demander-subvention.>
'Il est possible d'obtenir une aide financière pour commander des livres officiels.'

(128) 大家都盼望春天的**到來**。
*dàjiā dōu pànwàng chūntiān de **dàolái**.*
<tout le monde-tout-espérer-printemps-De-**arriver**.>
'Tout le monde espère l'arrivée du printemps.'

(129)學習的**評估**是重要的事情。
*xuéxí de **pínggū** shì zhòngyào de shìqing / qíng.*
<apprendre-De-**évaluer**-être-important-De-chose.>
'L'évaluation du niveau d'étude est une chose importante.'

Les unités lexicales peuvent occuper différentes fonctions syntaxiques. Mais, elles demeurent essentiellement des verbes. Zhu Dexi dans son ouvrage *Cours de grammaire* [1982], définit les verbes de ce type comme étant des verbes nominalisés. De ce point de vue, les quatre unités lexicales mentionnées relèvent toutes de la catégorie des verbes [2004 : 60-61].

On doit donc se demander si les unités lexicales possédant différentes fonctions syntaxiques appartiennent à des catégories définies ou bien "dépendant du contexte". Cette question suscite encore des discussions, en particulier pour les verbes qui peuvent apparaître comme des noms. Selon la plupart des linguistes chinois, si une unité lexicale d'origine verbale peut servir comme verbe ou syntaxiquement comme nom lors de la construction de syntagmes ou de phrases, cette unité lexicale néanmoins reste un verbe [*cf.* Guo Rui, 2002 [2004 : 184-188]]. De ce point de vue, par exemple, les unités lexicales telles que 分析 *fēnxī* 'analyser', 研究 *yánjiū / jiù* 'effectuer des recherches', 管理 *guǎnlǐ* 'gérer', 解決 *jiějué* 'résoudre', 調查 *diàochá* 'enquêter', 貢獻 *gòngxiàn* 'contribuer à',

sont des verbes qui peuvent être utilisés comme des noms dans certains syntagmes ou phrases.

5.6.2 Critères permettant la catégorisation des unités lexicales chinoises

5.6.2.1 Critère de la propriété sémantique

En 1898, Ma Jianzhong a publié *Chinese Grammar*, ouvrage dans lequel il fonde la définition de catégories de mots sur la propriété sémantique. C'est ainsi qu'il écrit [1923 : 23] :

> *"Ne possédant aucun sens prédéfini, les mots ne sauraient appartenir à des catégories définies. Pour les ranger dans l'une ou l'autre des catégories, il faut obligatoirement connaître le contexte dans lequel ils apparaissent."*[22]

Selon la définition qu'il donne des catégories, les noms désignent des abstractions ou des objets concrets. Ils sont sujet ou objet dans une phrase. Les verbes sont destinés à exprimer des actions. Ils prennent la fonction de prédicat ou de complément d'objet lors de la construction des phrases. Les adjectifs servent à la description. Ils s'utilisent en tant que déterminants dans les phrases. Dans cette perspective, Ma Jianzhong propose de classer les mots chinois en neuf catégories :

1) Cinq catégories réservées aux mots lexicaux (實字 *shízì* 'mots pleins') : 名字 *míngzì* 'nom', 代字 *dàizì* 'pronom', 動字 *dòngzì* 'verbe', 靜字 *jìngzì* 'adjectif' et 狀字 *zhuàngzì* 'adverbe' ;

2) Quatre catégories réservées aux mots grammaticaux (虛字 *xūzì* 'mots vides') : 介字 *jièzì* 'préposition', 連字 *liánzì* 'conjonction', 助字 *zhùzì* 'mot auxiliaire' et 嘆字 *tànzì* 'exclamation'.

S'appuyant sur la théorie de Ma Jianzhong et l'approfondissant, Li Jinxi [1924] décrit ainsi la définition des catégories [1925 : 6] :

[22] "字無定義，故無定類。而欲知其類，當知上下文義何如耳。"
"*zì wú dìngyì, gù wú dìnglèi. ér yùzhī qí lèi, dāngzhī shàngxià wényì hérú ěr.*"

> *"On ne peut déterminer la catégorie des mots chinois par la seule étude des mots, autrement dit, à partir de leur forme graphique. Il est nécessaire de savoir quelle place ou quelle fonction ils occupent dans la phrase avant de déterminer à quelle catégorie ils appartiennent. C'est une différence importante entre la grammaire de la langue chinoise et celle des langues indo-européennes. ..."*[23]

Li Jinxi propose donc de distinguer cinq groupes de catégories de mots [1925 : 6] :

1) Le groupe des mots désignant des objets concrets, 實體詞 *shítǐcí*, rassemble les noms (名詞 *míngcí*) et les pronoms (代名詞 *dàimíngcí*) ;

2) Le groupe des mots de présentation, 述說詞 *shùshuōcí*, rassemble les verbes (動詞 *dòngcí*) ;

3) Le groupe des mots de distinction, 區別詞 *qūbiécí*, rassemble les adjectifs (形容詞 *xíngróngcí*) et les adverbes (副詞 *fùcí*) ;

4) Le groupe des mots de relation, 關係詞 *guānxìcí*, rassemble les prépositions (介詞 *jiècí*) et les conjonctions (連詞 *liáncí*) ;

5) Le groupe des mots de description d'état, 情態詞 *qíngtàicí*, rassemble les mots auxiliaires (助詞 *zhùcí*) et les mots d'exclamation (嘆詞 *tàncí*).

Cette catégorisation des mots constitue une base théorique sur laquelle de nombreux linguistes continuent de proposer des définitions plus complexes et de donner des critères plus précis pour déterminer la catégorie des mots (des unités lexicales).

5.6.2.2 Critère des mots-clefs

Fang Guangdao [1939] propose une approche selon laquelle la forme morphologique permet de déterminer la catégorie d'un mot. Selon lui, l'interrelation ou la composition des mots fournit une piste pour décider de leurs catégories.

D'autres linguistes ont adopté ce critère d'identification de la catégorie d'une unité lexicale à partir de leur constitution syntagmatique. En 1954, Lü Shuxiang [2002 : 233-238]

[23] "國語的詞類，在詞的本身上（即字的形體上）無從分別；必須看他在句中的位置、職務，才能認定這一個詞是何種詞類：這是國語文法和西文法一個大不相同之點。……"
"*guóyǔ de cílèi, zài cí de běnshēn shàng (jí zì de xíngtǐ shàng) wúcóng fēnbié ; bìxū kàn tā zài jù zhōng de wèizhì, zhíwù, cái néng rèndìng zhè yī gè cí shì hé zhǒng cílèi : zhè shì guóyǔ wénfǎ hé / hàn xī wénfǎ yī gè dà bù xiāngtóng zhī diǎn. ...*"

définit l'approche par les mots-clefs qui consiste à utiliser certains morphèmes ou mots représentatifs pour établir des règles grammaticales concernant les mots ciblés. Un des exemples qu'il propose est la règle 很 *hěn* ~ [24]. L'adverbe 很 *hěn* 'très ; tellement' est le mot-clef qui sert à déterminer les adjectifs, qu'il précède toujours. Cependant, certains adjectifs ne peuvent pas être qualifiés par cet adverbe. Lü Shuxiang cite le cas de 正 *zhèng* 'positif', 負 *fù* 'négatif', 反 *fǎn* 'négatif', 副 *fù* 'secondaire', 真 *zhēn* 'vrai', 假 *jiǎ* 'faux', 絕對 *juéduì* 'absolu', 相對 *xiāngduì* 'relatif', 唯一 *wéiyī* 'unique' ou 無限 *wúxiàn* 'infini'[25]. Par ailleurs, cet adverbe peut aussi qualifier des verbes tels que 想念 *xiǎngniàn* 'penser', 感謝 *gǎnxiè* 'remercier', 贊成 *cànchéng* 'approuver', 需要 *xūyào* 'nécessiter', 推崇 *tuīchóng* 'soutenir' ou 懂 *dǒng* 'comprendre' ou des verbes auxiliaires comme 會 *huì* 'être capable de' ou 能 *néng* 'pouvoir'.

De ce fait, cette règle n'est pas applicable à tous les cas adjectivaux parce que les unités lexicales qui suivent l'adverbe 很 *hěn* peuvent être non seulement des adjectifs, mais aussi des verbes ou encore des verbes auxiliaires. En outre, tous les adjectifs ne peuvent pas être déterminés selon cette règle.

L'approche par les mots-clefs est pratique lorsque tous les mots (unités lexicales) d'une même catégorie peuvent obéir à une règle grammaticale unique. Si celle-ci ne s'applique pas à tous les mots d'une même catégorie, il est nécessaire d'employer d'autres règles pour les déterminer. Donc, l'approche par les mots-clefs n'est pas le premier principe par lequel on peut catégoriser les mots chinois.

5.6.2.3 Critère des fonctions syntagmatiques

La théorie de Wang Li s'inspire de la théorie des rangs telle que l'a exposée le linguiste danois Otto Jespersen dans *The Philosophy of Grammar* [1924]. La théorie de Jespersen distingue trois rangs : le rang primaire, le rang secondaire et le rang terminal.

Selon Wang Li [1943], dans un syntagme ou une phrase, les places occupées par les unités lexicales peuvent être classées hiérarchiquement. De ce point de vue, un classement se définit comme un rang. Les classements de ce type constituent des intermédiaires entre

[24] Le symbole ~ placé à la suite d'un morphème ou d'un mot représente tout mot susceptible d'occuper cette place. Ici, tout mot qui peut suivre 很 *hěn*.

[25] À partir de Zhu Dexi [1982], on les traite comme des adjectifs à valeur distinctive.

les catégories d'unités lexicales et les éléments syntagmatiques (fonctions syntaxiques). Par conséquent, il est possible d'établir un rapport entre catégories d'unités lexicales et rangs, ces derniers renvoyant eux aussi aux éléments syntagmatiques (fonctions syntaxiques). De ce fait, les unités lexicales ne changent pas de catégories. Elles changent de rang lors de la constitution des phrases dans lesquelles elles sont incluses. Par exemple, un nom peut se placer au rang primaire, au rang secondaire ou au rang terminal. Dès lors, les dictionnaires peuvent attribuer telle ou telle catégorie à une unité lexicale. Lors de la constitution des phrases, les rangs sont plus importants que les catégories lexicales [1985 : 18-25].

Selon cette théorie de Wang Li, les trois rangs dépendent de structures syntagmatiques. C'est donc le classement de leurs fonctions respectives. Ainsi, la catégorie d'une unité lexicale dont on n'est pas certain pourrait être définie par son rang.

D'autres linguistes considèrent que les trois rangs, en tant qu'intermédiaires entre les catégories lexicales et les éléments syntagmatiques, sont des critères insuffisants pour déterminer la catégorie d'une unité lexicale [*cf*. Guo Rui, 2002 [2004 : 16]].

5.6.2.4 Critère de la fonction grammaticale

Le linguiste Guo Rui [2002] propose le critère de fonction grammaticale pour décider des catégories. Son critère est fondé sur deux particularités linguistiques de la langue chinoise :

> 1) Les structures syntagmatiques : Les catégories se réfèrent à l'ordre adopté par les unités lexicales lors de la formation des syntagmes ou des phrases. Par exemple, entre un numéral et un nom, il est nécessaire d'insérer un quantifieur se référant au nom considéré. (Les termes seront précisés plus loin). Ainsi, en chinois, on dit 三張桌子 *sān zhāng zhuōzi* <trois-Q-table> 'trois tables' et non *三張喜歡 *sān zhāng xǐhuān* <trois-Q-aimer>. Le syntagme formé d'un numéral et d'un quantifieur ne doit être suivi que d'un nom. Donc, il n'est pas correct d'énoncer *三張喜歡 *sān zhāng xǐhuān*, car le mot 喜歡 *xǐhuān* est un verbe. De ce fait, chaque place syntagmatique est occupée par une unité lexicale appartenant à une catégorie précise. Par conséquent, Guo Rui suppose que les places syntagmatiques permettent de reconnaître la catégorie d'une unité lexicale.

2) Les syntagmes chinois (que nous examinerons en 7.1) : À l'intérieur du syntagme, les rôles sont assumés par des unités lexicales de diverses catégories. Ainsi, la catégorie peut être confirmée à l'aide de la reconnaissance d'une unité lexicale dans tel ou tel syntagme. Prenons comme exemple le syntagme du type Modifieur-Tête. Puisque le Modifieur ne peut être assumé que par un adjectif ou un nom, on peut affirmer que les unités lexicales qui occupent la fonction de Modifieur sont un adjectif ou un nom. Ainsi :

(130) 漂亮衣服 *piàoliàng yīfu / fú* <joli-vêtement> 'joli vêtement'
(131) 棉布衣服 *miánbù yīfu / fú* <coton-vêtement> 'vêtement en coton'

Ce type d'analyse permet d'attribuer une catégorie aux unités lexicales qui sont des Modifieurs dans les exemples (131) et (132). Le mot 漂亮 *piàoliàng* est un adjectif, tandis que le mot 棉布 *miánbù* est un nom. Ils servent tous deux à qualifier la Tête nominale 衣服 *yīfu / fú*.

Pour identifier la catégorie d'une unité lexicale, nous avons, dans cette étude, adopté le critère de la fonction grammaticale parce que Guo Rui [2002] a pour ce faire, proposé des règles concrètes compatibles avec l'approche de ***NooJ***.

				Classifications chinoises	Traductions françaises	Abréviations
詞類 *cílèi* Catégories lexicales	可組合詞 *kězǔhécí* Mots composants	實詞 *shící* Mots lexicaux	體詞 *tǐcí* Mots substantifs	名詞 *míngcí*	Nom	N
				處所詞 *chùsuǒcí*	Nom de lieu	S
				方位詞 *fāngwèicí*	Locatif	F
				時間詞 *shíjiāncí*	Nom de temps	T
				區別詞 *qūbiécí*	Adjectif à valeur distinctive	B
				數詞 *shùcí*	Numéral	M
				量詞 *liàngcí*	Quantifieur	Q
				代名詞 *dàimíngcí*	Pronom[26]	R
			謂詞 *wèicí* Mots prédicatifs	代名詞 *dàimíngcí*	Pronom	R
				狀態詞 *zhuàngtàicí*	Adjectif à valeur descriptive	Z
				動詞 *dòngcí*	Verbe	V
				形容詞 *xíngróngcí*	Adjectif à valeur qualificative	A
				副詞 *fùcí*	Adverbe	D
				擬聲詞 *nǐshēngcí*	Onomatopée	O
		虛詞 *xūcí* Mots grammaticaux		介系詞 *jièxìcí*	Préposition	P
				連接詞 *liánjiēcí*	Conjonction	C
				助詞 *zhùcí*	Auxiliaire	U
				語氣詞 *yǔqìcí*	Particule modale	Y
	單獨用詞 *dāndúyòngcí* Interjections			嘆詞 *tàncí*	Exclamation	E

Tableau 3 : Catégories lexicales chinoises

[26] La distinction entre les pronoms substantifs et les pronoms prédicatifs est abordée dans la suite.

5.6.3 Catégories lexicales en chinois moderne

Le système de catégorisation lexicale actuellement utilisé est présenté page précédente. Ce système contient dix-huit catégories définies en fonction des structures syntaxiques. Le groupe des mots pleins comporte deux sous-groupes : les mots prédicatifs et les mots substantifs. Il existe également un groupe lexical minoritaire qui regroupe les interjections, mots qui ne peuvent être combinés avec d'autres mots, mais sont toujours employés seuls dans les phrases ou constituent même, à eux seuls, une phrase.

● 名詞 *míngcí* 'Nom' (N)

Les noms servent à nommer des objets, qu'ils soient concrets ou abstraits, par exemple,

(132) 蘋果 *píngguǒ* 'pomme'
思想 *sīxiǎng* 'pensée'

● 方位詞 *fāngwèicí* 'Locatif' (F)

Les locatifs, eux, marquent une position spatiale ou temporelle relative, par exemple, 外面 *wàimiàn* 'à l'extérieur de' et 裡面 *lǐmiàn* 'à l'intérieur de'. Le tableau suivant expose leur structuration :

Morphèmes chinois / Locatifs simples	面 *miàn* 'face'	頭 *tou* 'suffixe'	邊 *biān* 'côté'	Significations
上 *shàng* 'dessus'	上面 *shàngmiàn*	上頭 *shàngtou*	上邊 *shàngbiān*	dessus
下 *xià* 'dessous'	下面 *xiàmiàn*	下頭 *xiàtou*	下邊 *xiàbiān*	dessous
前 *qián* 'avant ; devant'	前面 *qiánmiàn*	前頭 *qiántou*	前邊 *qiánbiān*	avant ; devant
後 *hòu* 'après ; derrière'	後面 *hòumiàn*	後頭 *hòutou*	後邊 *hòubiān*	après ; derrière
左 *zuǒ* 'gauche'	左面 *zuǒmiàn*		左邊 *zuǒbiān*	à gauche de
右 *yòu* 'droite'	右面 *yòumiàn*		右邊 *yòubiān*	à droite de
裡 *lǐ* 'intérieur'	裡面 *lǐmiàn*	裡頭 *lǐtou*	裡邊 *lǐbiān*	à l'intérieur de
外 *wài* 'extérieur'	外面 *wàimiàn*	外頭 *wàitou*	外邊 *wàibiān*	à l'extérieur de
東 *dōng* 'est'	東面 *dōngmiàn*	東頭 *dōngtou*	東邊 *dōngbiān*	à l'est de
南 *nán* 'sud'	南面 *námiàn*	南頭 *nántou*	南邊 *nánbiān*	au sud de
西 *xī* 'ouest'	西面 *xīmiàn*	西頭 *xītou*	西邊 *xībiān*	à l'ouest de
北 *běi* 'nord'	北面 *běimiàn*	北頭 *běitou*	北邊 *běibiān*	au nord de

Tableau 4 : Composition des locatifs

● 處所詞 *chùsuǒcí* ‘Nom de lieu’ (S)

Les noms de lieu se réfèrent

1) à des lieux géographiques :

(133) 巴黎 *Bālí* ‘Paris’,
倫敦 *Lúndūn* ‘Londres’,
比利時 *Bǐlìshí* ‘Belgique’, etc. ;

2) à des noms communs qui désignent un lieu particulier ou une position spatiale :

(134) 野外 *yěwài* ‘campagne’,
門口 *ménkǒu* ‘seuil’,
郊區 *jiāoqū* ‘banlieue’, etc.

● 時間詞 *shíjiāncí* ‘Nom de temps’ (T)

Les noms de temps indiquent un moment, une époque, une saison, une date ou bien, historiquement, une dynastie. Exemples :

(135) 明天 *míngtiān* ‘demain’,
上午 *shàngwǔ* ‘matin’,
春節 *chūnjié* ‘nouvel an chinois’,
夏朝 *xiàcháo* ‘dynastie des Xia (2207 — 1766 av. J. – C.)’, etc.

● 區別詞 *qūbiécí* ‘Adjectif à valeur distinctive’ (B)

La catégorie dite 區別詞 *qūbiécí* a été traduite par plusieurs termes. Dans cette étude, nous avons adopté la traduction de Zhitang Yang-Drocourt [2007 : 287-291], soit « adjectif à valeur distinctive ». En outre, nous avons adopté son abréviation, « B », bien connue et utilisée dans les domaines linguistiques et informatiques. Nous l’avons employée lors du développement du module chinois dans ***NooJ*** et de la rédaction de cette thèse.

Les adjectifs à valeur distinctive se présentent par paires qui permettent de les distinguer et de les classer en deux groupes, par exemple,

(136) 中式 *zhōngshì* ‘oriental’ et 西式 *xīshì* ‘occidental’ ;
急性 *jíxìng* ‘aigu’ et 慢性 *mànxìng* ‘chronique’.

● 數詞 *shùcí* ‘Numéral’ (M)

Les numéraux servent à former les nombres cardinaux et ordinaux. La constitution numérique se fait selon une structure prédéfinie qui ne peut être décrite de façon exhaustive [*cf*. Guo Rui, 2002 [2004 : 219-222]]. Les numéraux se divisent en quatre groupes :

1) 系數 *xìshù* ‘numéraux cardinaux’

Les numéraux cardinaux comprennent les numéraux invariables et les numéraux quantitatifs :

1.1) Les numéraux invariables :

■ Les chiffres de zéro à dix écrits avec les graphies : 〇 *líng* ‘zéro’, 一 *yī* ‘un’, 二 *èr* ‘deux’, 三 *sān* ‘trois’, 四 *sì* ‘quatre’, 五 *wǔ* ‘cinq’, 六 *liù* ‘six’, 七 *qī* ‘sept’, 八 *bā* ‘huit’, 九 *jiǔ* ‘neuf’ et 十 *shí* ‘dix’,

■ Les chiffres de zéro à dix en grande écriture[27] : 零 *líng* ‘zéro’, 壹 *yī* ‘un’, 貳 *èr* ‘deux’, 參 *sān* ‘trois’, 肆 *sì* ‘quatre’, 伍 *wǔ* ‘cinq’, 陸 *liù* ‘six’, 柒 *qī* ‘sept’, 捌 *bā* ‘huit’, 玖 *jiǔ* ‘neuf’ et 拾 *shí* ‘dix’,

■ L’expression de la moitié d’une unité 半 *bàn* ‘moitié’ ou une variation graphique du chiffre ‘deux’ 兩 *liǎng*.

1.2) Les numéraux quantitatifs incluent des mots comme 多 *duō* ‘nombreux’, 幾 *jǐ* ‘quelques’, 數 *shù* ‘plusieurs’, 大量 *dàliàng* ‘grande quantité’, 好些 *hǎoxiē* ‘un bon nombre’, 好幾 *hǎojǐ* ‘certain’, 若干 *ruògān* ‘un certain nombre’, 許多 *xǔduō* ‘beaucoup’ ou 無數 *wúshù* ‘innombrable’.

2) 位數 *wèishù* ‘numéraux (cardinaux) positionnels’

Ce sont 十 *shí* ‘dix’, 百 *bǎi* ‘cent’, 佰 *bǎi* ‘cent (en grande écriture)’, 千 *qiān* ‘mille’, 仟 *qiān* ‘mille (en grande écriture)’, 萬 *wàn* ‘dix mille’, 億 *yì* ‘cent millions’ et 兆 *zhào* ‘trillion’.

[27] La “grande écriture” ou “écriture complexe” est un style graphique appliqué aux numéraux et qu’on trouve dans les documents officiels tels que textes administratifs, annonces universitaires ou chèques. Il est utilisé en place de l’écriture habituelle notamment pour éviter la falsification. Par exemple, il est facile de transformer un 2 二 *èr*, en un 3 三 *sān*.

3) 系位數 *xìwèishù* ‘numéraux cardinaux positionnels’

Les numéraux cardinaux positionnels sont 廿 *niàn* ‘vingt’ et 卅 *sà* ‘trente’.

4) 數量數 *shùliàngshù* ‘numéraux numériques quantificatifs’

Les numéraux numériques quantificatifs sont 倆 *liǎ* ‘deux’ et 仨 *sā* ‘trois’.

● 量詞 *liàngcí* Quantifieur (Q)

La catégorie dite 量詞 *liàngcí* a deux traductions. Dans le domaine linguistique, elle est traduite par le terme « classifieur » abrégé en « CL ». Dans le domaine du Traitement Automatique des Langues Naturelles, elle est traduite par le terme « quantifieur » et représentée par « Q ». Dans cette étude, nous avons adopté le terme « quantifieur » et le code « Q » en les appliquant à notre développement du module chinois et à notre explication de thèse.

Il existe cinq sortes de quantifieurs :

1) 名量詞 *míngliàngcí* ‘quantifieurs nominaux’

Les quantifieurs nominaux ont pour rôle de mesurer des objets concrets. D’après les différentes méthodes de mesure, ces quantifieurs nominaux se divisent en huit sous-catégories :

1.1) 個體量詞 *gètǐ liàngcí* ‘quantifieurs d’unité’ : 匹 *pī*, 張 *zhāng*, 本 *běn*, 輛 *liàng*, 顆 *kē*, etc. :

(137) 一匹馬 *yī pī mǎ* <un-Q-cheval> ‘un cheval’

1.2) 集合量詞 *jíhé liàngcí* ‘quantifieurs de groupe’ : 副 *fù*, 套 *tào*, 批 *pī*, 組 *zǔ*, 群 *qún*, etc. :

(138) 一披貨 *yī pī huò* <un-Q-marchandise> ‘un lot de marchandises’

1.3) 成形量詞 *chéngxíng liàngcí* ‘quantifieurs de forme’ : 團 *tuán*, 塊 *kuài*, 滴 *dī*, 片 *piàn*, 節 *jié*, etc. :

(139) 一滴水 *yī dī shuǐ* <un-Q-eau> ‘une goutte d’eau’

1.4) 度量詞 *dùliàngcí* ‘quantifieurs de mesure’ : 克 *kè* ‘gramme’, 公分 *gōngfēn* ‘centimètre’, 升 *shēng* ‘litre’, 斤 *jīn* ‘livre’, 米 *mǐ* ‘mètre’, etc. :

(140) 一斤茶 *yī jīn chá* <un-Q-thé> ‘une livre de thé’

1.5) 種類量詞 *zhǒnglèi liàngcí* ‘quantifieurs d’espèce’ : 樣 *yàng*, 種 *zhǒng*, 類 *lèi*, etc. :

(141) 一種水果 *yī zhǒng shuǐguǒ* <un-Q-fruit> ‘une espèce de fruits’

1.6) 等級量詞 *děngjí liàngcí* ‘quantifieurs de degré’ : 等 *děng*, 級 *jí*, 號 *hào*, etc. :

(142) 一級檢定 *yī jí jiǎndìng* <un-Q-examen> ‘premier niveau d’examen’

1.7) 過程量詞 *guòchéng liàngcí* ‘quantifieurs de processus’ : 場 *chǎng*, 盤 *pán*, 頓 *dùn*, etc. :

(143) 一頓飯 *yī dùn fàn* <un-Q-repas> ‘un repas’

1.8) 不定量詞 *bù dìng liàngcí* ‘pseudo-quantifieurs’ : 些 *xiē* ‘quelque’, 點兒 *diǎnr* ‘quelque’, etc. :

(144) 一些糖果 *yī xiē tángguǒ* <un-Q-bonbon> ‘quelques bonbons’.

2) 動量詞 *dòngliàngcí* ‘quantifieurs verbaux’ : 下 *xià*, 回 *huí*, 次 *cì*, 場 *chǎng*, 遍 *biàn*, etc.

Les quantifieurs verbaux s’utilisent souvent avec des verbes :

(145) 看一場電影 *kàn yī chǎng diànyǐng* <regarder-un-Q-film>
‘regarder un film’.

3) 時量詞 *shíliàngcí* ‘quantifieurs de temps’

Ce sont des quantifieurs basés sur les mots qui désignent la temporalité, comme 天 *tiān* ‘jour’, 年 *nián* ‘an’, 秒 *miǎo* ‘seconde’, 分鐘 *fēnzhōng* ‘minute’, 會兒 *huǐr* ‘un petit moment’, 陣子 *zhènzi* ‘un certain temps’, etc. :

(146) 一年工程 *yī nián gōngchéng* <un-Q-travail> ‘un travail d’un an’

4) 臨時量詞 *línshí liàngcí* ‘quantifieurs occasionnels’ :

4.1) 容量量詞 *róngliàng liàngcí* ‘quantifieurs de contenance’ : 杯 *bēi,* 盆 *pén,* 身 *shēn,* 屁股 *pìgǔ,* 黑板 *hēibǎn,* etc. :

(147) 一盆水 *yī pén shuǐ* <un-Q-eau> ‘une cuvette d’eau’

4.2) 臨時動量詞 *línshí dòngliàngcí* ‘quantifieurs verbaux occasionnels’ : Les quantifieurs verbaux occasionnels sont des noms employés comme quantifieurs, par exemple, 口 *kǒu* ‘bouche’, ou des verbes qu’on répète, comme 聽 *tīng* ‘écouter’ ou 跳 *tiào* ‘sauter’. Exemples :

(148) a. 喝一口茶 *hē yī kǒu chá* <boire-un-Q-thé> ‘boire une gorgée de thé’
b. 聽一聽音樂 *tīng yī tīng yīnyuè* <écouter-un-Q-musique> ‘écouter de la musique’
c. 嚇一跳 *xià yī tiào* <être effrayé-un-Q> ‘être effrayé’

5) 複合量詞 *fùhé liàngcí* ‘quantifieurs composés’

Il est possible de regrouper deux quantifieurs pour en créer un nouveau. Il existe trois types de quantifieurs composés en chinois :

5.1) 名量詞 *míngliàngcí* + 名量詞 *míngliàngcí* : 件套 *jiàntào,* 篇部 *piānbù,* etc.

(149) 六件套洋裝 *liù jiàntào yángzhuāng* <six-Q-Q-robe> ‘six robes’

5.2) 名量詞 *míngliàngcí* + 動量詞 *dòngliàngcí* : 人次 *réncì,* 場次 *chǎngcì,* 架次 *jiàcì,* 班次 *bāncì,* 輛次 *liàngcì,* etc.

(150) 三班次火車 *sān bāncì huǒchē* <trois-Q-Q-train> ‘trois trains’

5.3) 度量詞 *dùliàngcí* + 度量詞 *dùliàngcí* : 千瓦時 *qiānwǎ shí* ‘mille watt par heure’, 米每秒 *mǐ měimiǎo* ‘mètre par seconde’, etc.

(151) 二十米每秒 *èrshí mǐ měimiǎo* <vingt-Q-Q> ‘vingt mètres par seconde’

● 代名詞 *dàimíngcí* ‘Pronom’ (R)

Selon la fonction syntagmatique, on distingue deux types de pronoms [*cf.* Zhu Dexi, 1982 [2004 : 80-94]] :

1) les pronoms substantifs tels que 我 *wǒ* ‘je’, 他 *tā* ‘il’, 你們 *nǐmen* ‘vous (à la deuxième personne du pluriel)’ ou 她們 *tāmen* ‘elles’ ;

2) les pronoms prédicatifs tels que 這麼 *zhème* ‘si ; aussi’, 那麼 *nàme* ‘si ; aussi’, 這麼樣 *zhèmeyàng* ‘si ; aussi’ ou 那麼樣 *nàmeyàng* ‘si ; aussi’.

Ci-dessous, nous les présenterons autrement.

Les pronoms et les démonstratifs sont très souvent regroupés puisque certains d’entre eux peuvent appartenir aux deux catégories. Les pronoms et les démonstratifs peuvent être classés en quatre groupes [*cf.* Guo Rui, 2002 [2004 :225-226, 238-240]] :

1) Les démonstratifs (du nom) : ils peuvent ou non précéder le syntagme de forme Numéral-Quantifieur. Ils peuvent se placer directement avant le nom. On compte : 這 *zhè* ‘ce ~ -ci’, 那 *nà* ‘ce ~ -là’, 每 *měi* ‘chaque’, 某 *mǒu* ‘certain’, 另 *lìng* ‘autre’, 任何 *rènhé* ‘n’importe quel’, 其他 *qítā* ‘autre’ ou 唯一 *wéiyī* ‘unique’.

2) Les pronoms personnels représentent les noms humains, des objets, des animaux ou des dieux :

2.1) Les pronoms personnels : 我 *wǒ* ‘je’, 咱 *zán* ‘nous deux’, 你 *nǐ* ‘tu’, 您 *nín* ‘vous de politesse’, 他 *tā* ‘il’, 她 *tā* ‘elle’, 我們 *wǒmen* ‘nous’, 咱們 *zánmen* ‘nous inclusif (incluant la ou les personnes auxquelles s’adresse le locuteur)’, 你們 *nǐmen* ‘vous à la deuxième personne du pluriel’, 他們 *tāmen* ‘ils’ ou 她們 *tāmen* ‘elles’ ;

2.2) Les mots assimilés à des pronoms personnels tels que : 自己 *zìjǐ* ‘soi-même’, 自個兒 *zìgèr* ‘soi-même’, 人家 *rénjiā* ‘les autres’, 別人 *biérén* ‘autrui’, 他人 *tārén* ‘d’autres personnes’, 大家 *dàjiā* ‘tout le monde’, 大伙兒 *dàhuǒr* ‘nous tous’ ou 大傢伙兒 *dàjiāhuǒr* ‘nous tous’ ;

2.3) Les pronoms réservés aux objets : 它 *tā* ‘il’ ou 它們 *tāmen* ‘ils’ ;

2.4) Les pronoms réservés aux animaux : 牠 *tā* ‘il’ ou 牠們 *tāmen* ‘ils’ ;

2.5) Les pronoms réservés aux dieux : 祂 *tā* ‘il’ ou 祂們 *tāmen* ‘ils’.

3) Les pronoms démonstratifs : ils représentent d’autres mots et occupent la fonction de sujet. Il existe cinq types de pronoms démonstratifs :

3.1) Les deux démonstratifs 這 *zhè* ‘ce ~ ci’, 那 *nà* ‘ce ~ là’ peuvent être utilisés comme des pronoms démonstratifs, représentant d’autres mots ;

3.2) Les pronoms démonstratifs de lieu : 這裡 *zhèlǐ* ‘ici’, 那裡 *nàlǐ* ‘là-bas’, 這兒 *zhèr* ‘ici’ ou 那兒 *nàr* ‘là-bas’ ;

3.3) Les pronoms démonstratifs de temps : 這時 *zhèshí* ‘ce moment-là (proche)’, 那時 *nàshí* ‘ce moment-là (lointain)’, 這會兒 *zhèhuǐr* ‘ce moment-là (proche)’ ou 那會兒 *nàhuǐr* ‘ce moment-là (lointain)’ ;

3.4) Les pronoms démonstratifs verbaux : 這樣 *zhèyàng* ‘ainsi ; de la sorte’, 那樣 *nàyàng* ‘ainsi ; de la sorte’, 怎樣 *zěnyàng* ‘comment’ ou 怎麼樣 *zěnmeyàng* ‘comment’.

3.5) Les pronoms démonstratifs adverbiaux : 這麼 *zhème* ‘si ; aussi’, 那麼 *nàme* ‘si ; aussi’, 多 *duō* ‘comme’ ou 為什麼 *wèishénme* ‘pourquoi’.

4) Les pronoms interrogatifs avec lesquels on construit des phrases interrogatives. Exemples : 哪 *nǎ* ‘où’, 什麼 *shénme* ‘quel’, 哪裡 *nǎlǐ* ‘où’, 哪兒 *nǎ'r* ‘où’, 多 *duō* ‘quel’, 幾 *jǐ* ‘combien de’ ou 多少 *duōshǎo* ‘combien’.

● 狀態詞 *zhuàngtàicí* ‘Adjectif à valeur descriptive’ (Z)

Dans cette étude, nous avons adopté le terme français « adjectif à valeur descriptive» proposé par Zhitang Yang-Drocourt [2007 : 287-291] pour désigner le terme chinois 狀態詞 *zhuàngtàicí.*

Les adjectifs à valeur descriptive servent à représenter des objets et des situations. Ils représentent une image ou un degré de description. Certains d'entre eux sont construits à l'aide de la réduplication des adjectifs. Exemples :

(152) 冰凉 *bīngliáng* 'glacial'
雪白 *xuěbái* 'blanc immaculé'
甜絲絲 *tiánsīsī* 'tout sucré'
烏溜溜 *wūliūliū* 'noir brillant'
快快樂樂 *kuàikuài-lèlè* 'très content'
流裡流氣 *liúlǐliúqì* 'fantaisiste'

● 動詞 *dòngcí* 'Verbe' (V)

Les verbes expriment des actions, décrivent l'apparence des objets, les modifications qu'ils subissent et leurs relations entre eux. Ils constituent la catégorie la plus importante du vocabulaire chinois. Exemples :

(153) 洗 *xǐ* 'laver'
跑 *pǎo* 'courir'
跳 *tiào* 'sauter'
主張 *zhǔzhāng* 'préconiser'
工作 *gōngzuò* 'travailler'
準備 *zhǔnbèi* 'préparer'
想念 *xiǎngniàn* 'penser'
解釋 *jiěshìa* 'expliquer'

Il existe également des verbes auxiliaires, comme :

(154) 想 *xiǎng* 'avoir l'intention de'
要 *yào* 'vouloir'
能 *néng* 'être capable de'
可以 *kěyǐ* 'pouvoir'
應該 *yīnggāi* 'devoir'
能夠 *nénggòu* 'avoir la possibilité de'

● 形容詞 *xíngróngcí* 'Adjectif à valeur qualificative' (A)

Les adjectifs à valeur qualificative sont les adjectifs proprement dits. La plupart des adjectifs peuvent être qualifiés par l'adverbe 很 *hěn* 'très ; tellement'. Pour former une négation, on utilise 不 *bù* suivi d'un adjectif. Les adjectifs à valeur qualificative peuvent être qualifiés par des compléments adverbiaux. Par ailleurs, certains adjectifs sont

déterminants d'un nom, complément d'objet ou complément résultatif d'un verbe. Exemples :

(155) 白 *bái* 'blanc'
苦 *kǔ* 'amer'
平淡 *píngdàn* 'ordinaire'
忙錄 *mánglù* 'chargé'
舒服 *shūfú* 'confortable'

● 副詞 *fùcí* 'Adverbe' (D)

L'adverbe a pour fonction principale de circonstancier. Il précède le noyau prédicatif qui est assumé par un adjectif ou un verbe. Par exemple,

(156) 都 *dōu* 'tout'
也許 *yěxǔ* 'peut-être'
到處 *dàochù* 'partout'
剛才 *gāngcái* 'justement'
稍微 *shāowēi /wéi* 'quelque peu'
逐漸 *zhújiàn* 'progressivement'

● 擬聲詞 *nǐshēngcí* 'Onomatopée' (O)

Les onomatopées ont pour objectif d'imiter des sons ou les bruits de la nature. Elles peuvent constituer une phrase en elles-mêmes ou des unités indépendantes à l'intérieur d'une phrase. Exemples :

(157) 咯咯 *gégé* 'onomatopée des bruits produits par des meubles en bois'
噗通 *pūtōng* 'plouf se référant au bruit d'un objet qui tombe dans l'eau'
嘩啦啦 *huā--lā--lā* 'onomatopée du bruit d'un liquid — pluie ou eau du robinet — qui tombe sur une surface solide'
淅瀝嘩啦 *xīlì-huālā* 'onomatopée du bruit de la pluie qui tombe sur le sol'
轟隆轟隆 *hōnglōng-hōnglōng* 'onomatopée du grondement d'une voiture ou d'un train'

Par ailleurs, les onomatopées servent occasionnellement de verbes, par exemple, 哞 *mōu* 'meugler', 咩 *miē* 'bêler' ou 喵 *miāo* 'miauler'.

● 介系詞 *jièxìcí* ‘Préposition’ (P)

En chinois, certaines prépositions peuvent être confondues avec des verbes : c’est qu’elles proviennent, effectivement, de verbes auxquels elles doivent, d’ailleurs, leur sens. Exemples :

Prépositions	**Verbes**
在 *zài* ‘à / dans (un endroit)’	在 *zài* ‘se trouver à’
對 *duì* ‘vis-à-vis de’	對 *duì* ‘être en face de’
朝 *cháo* ‘vers (une orientation)’	朝 *cháo* ‘aller à’
用 *yòng* ‘à l’aide de’	用 *yòng* ‘employer’
給 *gěi* ‘à’	給 *gěi* ‘donner quelque chose à quelqu’un’
跟 *gēn* ‘en compagnie de’	跟 *gēn* ‘poursuivre’

Les prépositions ne peuvent être utilisées seules, ni prendre la forme répétitive, ni être associées à des particules aspectuelles comme 了 *le*, 著 *zhe* et 過 *guò*. Elles doivent toujours être accompagnées d’un complément.

● 連接詞 *liánjiēcí* ‘Conjonction’ (C)

On emploie les conjonctions pour relier des mots, des syntagmes, des phrases ou des groupes de phrases. Il existe quatre types de conjonctions :

1) Les conjonctions qui relient des mots ou des syntagmes, comme 同 *tóng* ‘et’, 和 *hé / hàn* ‘et’ et 跟 *gēn* ‘avec’.

2) Les conjonctions qui expriment une raison. Ce sont, par exemple, 即使 *jíshǐ* ‘bien que’, 儘管 *jǐnguǎn* ‘bien que’ et 雖然 *suīrán* ‘bien que’.

3) Les conjonctions qui expriment un effet. Citons 因此 *yīncǐ* ‘par conséquent’, 所以 *suǒyǐ* ‘ainsi’ et 然而 *ránér* ‘cependant’.

4) Les conjonctions qui servent à combiner des mots, des syntagmes ou des phrases comme 因為 *yīnweì* ‘car’, 並且 *bìngqiě* ‘de plus’ et 或者 *huòzhě* ‘ou’.

● 助詞 *zhùcí* ‘Auxiliaire’ (U)

Les auxiliaires sont des mots grammaticaux. Bien qu’ils ne soient pas nombreux, ils servent à former des mots, des syntagmes ou des phrases et à modifier la structure des

phrases ou leur aspect. Selon leur fonction grammaticale, ils peuvent être rangés en cinq groupes [*cf.* Guo Rui, 2002 [2004 : 235-236]] :

1) Les auxiliaires structuraux regroupent le subordinateur nominal 的 *de*, le subordinateur préverbal 地 *de*, le subordinateur postverbal 得 *de* [*cf.* Zhitang Yang-Drocourt, 2007 : 298-299] ou la marque placée devant le verbe (emphase littéraire de la relative) 所 *suǒ*.

2) Les auxiliaires aspectuels dits particules aspectuelles comprennent la particule aspectuelle perfective 了 *le*, la particule aspectuelle durative 著 *zhe* ou la particule aspectuelle d'expérience 過 *guò* [*cf.* Zhitang Yang-Drocourt, 2007 : 299-306].

3) Les auxiliaires utilisés pour former des phrases comparatives : 一般 *yībān* 'semblable à' et 似的 *shìde* 'pour ainsi dire, comme'.

4) Les auxiliaires réservés aux syntagmes numéraux : 多 *duō* 'plus de', 來 *lái* 'environ', 餘 *yú* 'excédent' et 分之 *fēn zhī* 'à l'aide duquel on peut former des fractions, comme 二分之一 *èr fēn zhī yī* 'un demi''.

5) Les auxiliaires qui ne rentrent pas dans les catégories précédentes : 等 *děng* 'et cetera', 以來 *yǐlái* 'depuis' et 等等 *děngděng* 'et cetera'.

● 語氣詞 *yǔqìcí* 'Particule modale' (Y)

Les particules modales se trouvent souvent à la fin ou au milieu d'une phrase. Elles sont de trois genres :

1) Les particules modales introduisant l'inchoatif, un nouveau procès, un procès en cours, etc. : 了 *le*, 呢 *ne* ou 著呢 *zhe ne*.

2) Les particules modales interrogatives : 吧 *ba*, 呢 *ne* ou 嗎 *ma*.

3) Les particules modales exclamatives : 啊 *a*.

Notons qu'il est possible de combiner deux ou plusieurs particules modales, par exemple, 了吧 *le ba*, 了嗎 *le ma* ou 著呢吧 *zhe ne ba*.

● 嘆詞 *tàncí* Exclamation (E)

Les exclamations sont, dans les phrases, utilisées indépendamment et séparées par des ponctuations. Elles peuvent à elles seules constituer une phrase. Elles ne possèdent ni signification, ni fonction syntaxique. Elles n'ont pour objet que l'expression d'un sentiment à l'aide d'intonations. Exemples : 呣 *móu*, 呦 *yōu*, 哼 *hēng*, 嗯 *ēn* ou 嘖 *zé.*

5.7 Développement des dictionnaires électroniques

5.7.1 Informations utilisées dans le dictionnaire

Des informations diverses peuvent être associées à chaque entrée des dictionnaires. Elles seront utilisées dans les différentes requêtes ou analyses, telles que recherche de groupes nominaux ou d'expressions. Les informations associées à une entrée de dictionnaire chinois concernent :

1) sa catégorie, (par exemple, nom, verbe, adjectif, adverbe …) ;

2) ses propriétés sémantiques, ainsi, l'étiquette <+Hum> désignera la classe humaine ;

3) le paradigme morphologique, par exemple, de réduplication, indiqué par l'étiquette <+FLX=>.

Ces informations seront associées à chaque entrée de dictionnaire. Voici comment se présentent quelques entrées de notre dictionnaire principal ***ChDic*** :

一,M
一一,D
一二,M
一眨眼,D
一帆風順,I[28]

[28] 一,M *yī* 'un'
一一,D *yīyī* 'un par un'
一二,M *yī'èr* 'un ou deux : quelques'
一眨眼,D *yīzhǎyǎn* 'en un clin d'œil : en un instant'
一帆風順,I *yīfān / fán fēngshùn* 'vent favorable plein la voile : tout va comme sur des roulettes'

Chaque entrée est suivie de sa catégorie. À l'adverbe « 一一 », par exemple, est associée l'étiquette <D>.

Si une unité lexicale relève de plus d'une catégorie, il est important de les distinguer. Pour cette étude, nous suggérons deux traitements différents selon la nature des Unités Linguistiques Atomiques considérées :

1) Si une unité lexicale appartient à plusieurs catégories qui sont bien déterminées, nous les distinguons par différentes entrées liées à la même graphie. Chaque entrée est suivie par une catégorie. On aura ainsi :

持續,V
持續,D
特別,A
特別,D
端正,A
端正,V
耐心,A
耐心,N
負責,V
負責,A[29]

2) Si une unité lexicale appartient syntaxiquement à plusieurs catégories, nous traitons l'unité comme si elle ne possédait qu'une catégorie originelle. Ceci concerne surtout les verbes nominalisés [*cf.* Guo Rui, 2002 [2004 : 185-188]]. Dans le dictionnaire ***ChDic***, les unités lexicales de ce type sont présentées selon leur catégorie originelle et étiquetées en conséquence. Ainsi :

分析,V
改造,V
管理,V
解決,V

[29] 持續,V *chíxù* 'continuer'
持續,D *chíxù* 'continuellement'
特別,A *tèbié* 'particulier'
特別,D *tèbié* 'particulièrement'
端正,A *duānzhèng* 'correct'
端正,V *duānzhèng* 'corriger'
耐心,A *nàixīn* 'patient'
耐心,N *nàixīn* 'patience'
負責,V *fùzé* 'être responsable de'
負責,A *fùzé* 'responsable'

調查,V
貢獻,V[30]

5.7.2 Dictionnaires

Pour mener à bien cette étude, nous avons construit six dictionnaires. Le dictionnaire principal est le ***ChDic*** (Chinese Dictionary) qui contient 79 046 entrées. Ses étiquettes sont présentées de la manière suivante [*cf.* Zhou Qiang et Yu Shiwen, 1996 : 1-11] :

Catégories	Abréviations	Exemples
Mots		
Nom	N	世代 *shìdài* ‘génération’
Nom d’entreprises, D’organismes, etc.	NT	中國銀行 *Zhōngguó yínháng* ‘Banque de Chine’
Nom propre	NZ	世界盃 *shìjièbēi* ‘coupe du monde’
Nom de lieu	S	城郊 *chéngjiāo* ‘faubourg’
Locatif	F	上 *shàng* ‘dessus’
Nom de temps	T	上午 *shàngwǔ* ‘matin’
Adjectif à valeur distinctive	B	世界級 *shìjièjí* ‘niveau mondial’
Numéral	M	一 *yī* ‘un’
Quantifieur	Q	棵 *kē* ‘quantifieur réservé aux arbres’
Pronom (ou démonstratif)	R	他 *tā* ‘il’
Adjectif à valeur descriptive	Z	乾巴巴 *gānbābā* ‘desséché’
Verbe	V	主持 *zhǔchí* ‘diriger’
Adjectif à valeur qualificative	A	安靜 *ānjìng* ‘silencieux’
Adverbe	D	事先 *shìxiān* ‘préalablement’
Onomatopée	O	叮咚 *dīngdōng* ‘cliquetis’
Préposition	P	向 *xiàng* ‘vers’
Conjonction	C	因為 *yīnwèi* ‘parce que’
Auxiliaire	U	分之 *fēn zhī* ‘auxiliaire pour former une fraction comme 二分之一 *èr fēn zhī yī* ‘un demi’’
Particule modale	Y	呢 *ne* ‘particule modale interrogative’

[30] 分析,V *fēnxī* ‘analyser’
改造,V *gǎizào* ‘restaurer’
管理,V *guǎnlǐ* ‘gérer’
解決,V *jiějué* ‘résoudre’
調查,V *diàochá* ‘enquêter’
貢獻,V *gòngxiàn* ‘offrir’

Exclamation	E	哎 *aī* ‘exclamation qui exprime l’étonnement ou la pitié’
Formule quadrisyllabique	I	羊腸小徑 *yángcháng-xiǎojìng* ‘sentier sinueux’
Locution	L	做東道 *zuòdōngdào* ‘inviter à goûter’
Abréviation	J	世博會 *shìbóhuì* ‘exposition universelle’

Affixations		
Préfixe	H	老 *lǎo*
Suffixe	K	子 *zi*
Semi-préfixe	SH	半 *bàn*
Semi-suffixe	SK	學 *xué*
Morphèmes non autonomes de différentes propriétés sémantiques Pour distinguer les morphèmes non autonomes, nous les classons selon leur propriété sémantique : en effet, ils peuvent être de nature nominale, adjectivale, verbale, prépositionnelle, etc. Ainsi, le nom 丞相 *chéngxiàng* <premier ministre-administrateur> ‘premier ministre’ est composé de deux morphèmes 丞 *chéng* et 相 *xiàng*. Le morphème 丞 *chéng* est étiqueté <NG>. Il en est de même pour 相 *xiàng*. Ces morphèmes sont donc, d’après leur sens, de nature nominale. Le classement de morphèmes non autonomes se fait comme suit :		
Nom	NG	程 *chéng* ‘mesure’
Locatif	FG	朔 *shuò* ‘nord’
Nom du temps	TG	晨 *chén* ‘matin’
Adjectif à valeur distinctive	BG	翠 *cuì* ‘bleu vert’
Numéral	MG	數 *shù* ‘nombre’
Pronom (ou démonstratif)	RG	汝 *rǔ* ‘tu ; vous’
Adjectif à valeur descriptive	ZG	赳 *jiū* ‘courageux’
Verbe	VG	趨 *qū* ‘se diriger vers’
Adjectif à valeur qualificative	AG	吉 *jí* ‘propice’
Adverbe	DG	孜 *zī* ‘diligemment’
Préposition	PG	在 *zài* ‘à’
Conjonction	CG	啻 *chì* ‘seulement’
Particule modale	YG	爾 *ěr* ‘particule finale’
Exclamation	EG	得 *dé* ‘morphème non autonome qui exprime une réaction face à une situation formidable ou terrible’
Caractères utilisés dans la constitution des onomatopées	OG	淙 *cóng*
Caractères servant à transposer les syllabes de mots étrangers	PHON	葡 *pú*
Caractères utilisés dans la constitution des prénoms	NPRG	東 *dōng*

Le dictionnaire ***DicBkTitl*** (Dictionary of Book Titles) contient 93 titres de livres assortis de l'étiquette <N+bk>. Par exemple,

三國演義,N+bk
水滸傳,N+bk
紅樓夢,N+bk[31]

Le dictionnaire ***DicChSurn*** (Dictionary of Chinese Surnames) contient 1 156 noms de famille chinois, auxquels on a donné l'étiquette <NPL>. Par exemple,

司,NPL
司空,NPL
司馬,NPL
司徒,NPL
司寇,NPL[32]

Le dictionnaire ***GeoDic*** (Geographical Dictionary) contient 2 418 noms géographiques suivis de l'étiquette <S>. Par exemple,

三峽,S
三藩市,S
上海,S
亞利桑那州,S
巴黎,S[33]

Le dictionnaire ***DicProNam*** (Dictionary of Proper Names) contient 171 noms propres personnels qui ont reçu l'étiquette <NPR>. Par exemple,

哥白尼,NPR
巴金,NPR
巴爾札克,NPR

[31] 三國演義,N+bk *sānguó yǎnyì* 'Le roman des trois royaumes'
水滸傳,N+bk *shuǐhǔzhuàn* 'Au bord de l'eau'
紅樓夢,N+bk *hónglóumèng* 'Le rêve dans le pavillon rouge'

[32] 司,NPL *Sī*
司空,NPL *Sīkōng*
司馬,NPL *Sīmǎ*
司徒,NPL *Sītú*
司寇,NPL *Sīkòu*

[33] 三峽,S *Sānxiá* 'Sanxia'
三藩市,S *Sānfānshì* 'San Francisco'
上海,S *Shànghǎi* 'Shanghai'
亞利桑那州,S *Yǎlìsāngnàzhōu* 'Arizona'
巴黎,S *Bālí* 'Paris'

畢卡索,NPR
墨翟,NPR[34]

Le dictionnaire ***DicExpres*** (Dictionary of Expressions) contient les expressions figées de deux types telles que les proverbes et les expressions à double volet. Ce dictionnaire regroupe 591 expressions chinoises étiquetées <EXPRES>. Par exemple,

一江春水向東流,EXPRES
人生如白駒過隙, EXPRES
前事不忘後事之師,EXPRES[35]

Les différentes structures phrastiques des expressions sont également mises en valeur dans ce dictionnaire :

(158) 十年樹木百年樹人,EXPRES
(159) 十年樹木，百年樹人/,EXPRES[36]

On voit qu'il est possible d'insérer une virgule pour diviser l'expression (159) en deux phrases [*cf*. (160)]. Pour bien distinguer les ponctuations appartenant aux expressions de celles qui sont utilisées pour indiquer le code, une barre oblique (/) insérée devant la virgule, précède l'étiquette <EXPRES> [*cf*. (160)].

[34] 哥白尼,NPR *Gēbáinǐ* 'Copernicus'
巴金,NPR *Bā Jīn* 'Ba Jin'
巴爾札克,NPR *Bāěrzhákè* 'Balzac'
畢卡索,NPR *Bìkǎsuǒ* 'Picasso'
墨翟,NPR *Mò Zhái* 'Mo Zhai'

[35] 一江春水向東流,EXPRES *yī jiāng chūnshuǐ xiàng dōng liú*
<un-fleuve-printemps-eau-vers-est-couler>
'au printemps le fleuve coule vers l'est'
人生如白駒過隙, EXPRES *rénshēng rú báijū guò xì*
<vie-comme-blanc-cheval-traverser-crevasse>
'la vie est comme un blanc coursier sautant un ravin'
前事不忘後事之師,EXPRES *qiánshì bù wàng hòushì zhī shī*
<précédent-affaire-ne pas-oublier-suivant-affaire-U-maître>
'n'oublie pas les événements passés, ce sont des maîtres pour l'avenir'

[36] 十年樹木百年樹人,EXPRES *shínián shùmù bǎinián shùrén*
<dix-an-planter-arbre-cent-an-cultiver-homme>
'dix ans pour cultiver un arbre, cent ans pour bien former un être humain : la formation d'hommes de talent demande plusieurs générations, de longs efforts'
十年樹木，百年樹人/,EXPRES *shínián shùmù, bǎinián shùrén*
<dix-an-planter-arbre, cent-an-cultiver-homme>
'dix ans pour cultiver un arbre, cent ans pour bien former un être humain : la formation d'hommes de talent demande plusieurs générations, de longs efforts'

5.8 Traitement de la réduplication[37]

La réduplication en chinois consiste à former des mots en répétant un ou plusieurs éléments. Ces éléments peuvent être une syllabe, un morphème ou bien un mot. Ces mots réduplicatifs sont aussi traités en bloc. Nous avons choisi deux méthodes pour les représenter :

1) Certains mots réduplicatifs sont lemmatisés dans le dictionnaire. Ils existent sous trois formes :

1.1) Mots réduplicatifs créés par redoublement de leurs syllabes ou de leurs morphèmes :

(160) 奶奶 *nǎinai* <grand-mère paternelle-grand-mère paternelle>
‘grand-mère paternelle’
星星 *xīngxing* <étoile-étoile> ‘étoile’
剛剛 *gānggāng* <tout à l’heure-tout à l’heure> ‘tout à l’heure, venir de’

1.2) Adjectifs à valeur descriptive constitués selon la structure **ABB**. **A** est un morphème et constitue fréquemment un adjectif. **B** est une syllabe ou un morphème. La réduplication **BB**, seule, n’a pas de sens. Exemples :

(161) 烏溜溜 *wūliūliū* <noire-*liūliū*> ‘noir brillant’
圓滾滾 *yuángǔngǔn* <rond-*gǔngǔn*> ‘rondelet ; tout rond’
傻呼呼 *shǎhūhū* <stupide-*hūhū*> ‘niais’

1.3) Mots réduplicatifs qui ne sont pas formés à partir de mots déjà existants. Plus précisément, au départ, l’association de **A** et de **B** ne produit pas un. Seule la réduplication **AABB** pourra lui en donner un. Cette réduplication **AABB** est fondée sur deux morphèmes différents et isolés (**A** et **B**) :

(162) 瓶瓶罐罐 *píngpíng-guànguàn* <bouteille-bouteille-poterie-poterie>
‘une quantité de bouteilles’
熙熙攘攘 *xīxī-rǎngrǎng* <immense-en désordre>
‘grouillement de la foule’
轟轟烈烈 *hōnghōng-lièliè* <grandiose-majestueux> ‘grandiose’

Ces trois sortes de mots réduplicatifs ont une entrée particulière dans le dictionnaire.

[37] *cf.* R. Kabore, 1998 : 359-376.

2) Si les mots réduplicatifs sont dérivés de mots attestés, leur traitement est différent de ceux sus-mentionnés. D'une part, on associe les étiquettes des diverses réduplications aux entrées concernées dans le dictionnaire. D'autre part, les règles de réduplication sont décrites à l'aide d'une grammaire ***NooJ*** de type flexionnel (fichier .nof).

Nous allons maintenant exposer de façon plus détaillée et au cas par cas comment fonctionne ce traitement en deux phases.

5.8.1 La réduplication AA

La réduplication **AA** concerne des adjectifs, des noms, des quantifieurs monosyllabiques, etc. Leurs formes réduplicatives sont constituées par leur propre répétition. Le nom 人 *rén* 'personne' se transforme en 人人 *rénrén* 'chaque personne' par redoublement. Dans le dictionnaire ***ChDic***, nous associons la propriété **FLX=Ren** aux entrées de ce type :

人,N+FLX=Ren
家,N+FLX=Ren
白,A+FLX=Ren
紅,A+FLX=Ren
朵,Q+FLX=Ren
隻,Q+FLX=Ren[38]

Dans la grammaire flexionnelle de ***NooJ Reduplicative forms.nof***, nous décrivons ainsi la règle de la réduplication **AA** :

Ren = <E>/Std + <D>/AA;

Le premier terme représente la forme originale des entrées. Le second terme qui contient la commande <D>, redouble ces entrées. Par conséquent, les formes réduplicatives seront liées automatiquement aux entrées lexicales. Grâce à cette règle on reconstitue automatiquement les formes réduplicatives des exemples mentionnés :

[38] 人,N+FLX=Ren *rén* 'personne'
家,N+FLX=Ren *jiā* 'famille'
白,A+FLX=Ren *bái* 'blanc'
紅,A+FLX=Ren *hóng* 'rouge'
朵,Q+FLX=Ren *duǒ* 'quantifieur réservé aux fleurs'
隻,Q+FLX=Ren *zhī* 'quantifieur réservé aux animaux'

人人
家家
白白
紅紅
朵朵
隻隻[39]

5.8.2 La réduplication ABB

Lorsqu'un adjectif ou un adjectif à valeur descriptive est composé de deux morphèmes **A** et **B**, il peut prendre la forme **ABB** par la répétition de son deuxième composant. La forme réduplicative appartient à la catégorie des adjectifs à valeur descriptive, même si son mot initial est un adjectif simple[40]. Par exemple, l'adjectif 冷清 *lěngqīng* <froid-tranquille> 'solitaire' deviendra 冷清清 *lěngqīngqīng* <froid-tranquille> 'solitaire' par le redoublement de son dernier composant morphologique. Pour décrire la réduplication **ABB**, nous avons, d'abord, associé la propriété **FLX=Lengqing** aux entrées lexicales concernées dans le dictionnaire ***ChDic*** :

冷清,A+FLX=Lengqing
孤單,A+FLX=Lengqing
火熱,Z+FLX=Lengqing
墨黑,Z+FLX=Lengqing[41]

Ensuite, nous avons décrit la règle de la réduplication **ABB** dans la grammaire flexionnelle de ***NooJ Reduplicative forms.nof*** :

Lengqing = <E>/Std + <D>/ABB;

[39] 人人 *rénrén* 'chaque personne'
家家 *jiājiā* 'chaque famille'
白白 *báibái* 'blanc'
紅紅 *hónghóng* 'rouge'
朵朵 *duǒduǒ* 'quantifieur réservé aux fleurs'
隻隻 *zhīzhī* 'quantifieur réservé aux animaux'

[40] Dans cette étude, nous ne traitons pas le changement de catégories lors de la constitution automatique des formes réduplicatives.

[41] 冷清,A+FLX=Lengqing *lěngqīng* <froid-tranquille> 'solitaire'
孤單,A+FLX=Lengqing *gūdān* <solitaire-seul> 'isolé'
火熱,Z+FLX=Lengqing *huǒrè* <feu-chaud> 'chaud comme le feu : brûlant'
墨黑,Z+FLX=Lengqing *mòhēi* <encre-noir> 'noir comme l'encre'

Le premier terme affiche la forme originale des entrées. Le second terme permet de doubler leur deuxième composant. La réduplication **ABB** est effectuée grâce à cette description formelle. Les formes réduplicatives dérivées des mots mentionnés ci-dessus sont présentées de la façon suivante :

冷清清
孤單單
火熱熱
墨黑黑[42]

5.8.3 La réduplication AABB

Si la forme **AB** est un mot dissyllabique, elle peut être redoublée en **AABB**. Ce mot dissyllabique peut être un nom, un adjectif, un adverbe, un numéral ou un locatif. Chaque composant (**A** ou **B**) est redoublé. Les formes réduplicatives **AABB** n'appartiennent plus à leurs catégories originales ; elles deviennent des adjectifs à valeur descriptive[43]. Bien que la forme **AABB** change la catégorie syntaxique du mot, elle conserve le sens propre du mot de départ, mais l'intensifie. Par exemple, l'adjectif 乾脆 *gāncuì* <sec-cassant> 'net' devient 乾乾脆脆 *gāngān-cuìcuì* 'très net'. Dans le dictionnaire ***ChDic***, nous avons associé la propriété **FLX=Gancui** aux entrées concernées :

恩怨,N+FLX=Gancui
風雨,N+FLX=Gancui
安靜,A+FLX=Gancui
乾脆,A+FLX=Gancui
切實,D+FLX=Gancui
原本,D+FLX=Gancui
上下,F+FLX=Gancui
前後,F+FLX=Gancui
千萬,M+FLX=Gancui
許多,M+FLX=Gancui[44]

[42] 冷清清 *lěngqīngqīng* <froid-tranquille> 'solitaire'
孤單單 *gūdāndān* <seul-simplement> 'isolé'
火熱熱 *huǒrèrè* <feu-brûlant> 'tout brûlant'
墨黑黑 *mòhēihēi* <encre-noirâtre> 'tout noir'

[43] *cf.* n. 40.

[44] 恩怨,N+FLX=Gancui *ēnyuàn* <bienveillance-haine> 'bienveillance et haine'
風雨,N+FLX=Gancui *fēngyǔ* <vent-pluie> 'vent et pluie : mauvais temps ; difficulté'
安靜,A+FLX=Gancui *ānjìng* <paisible-calme> 'silencieux'
乾脆,A+FLX=Gancui *gāncuì* <sec-cassant> 'net'

Nous avons décrit la réduplication **AABB** dans la grammaire flexionnelle de ***NooJ Reduplicative forms.nof***. Sa description formelle se présente ainsi :

Gancui = <E>/Std + <L><D><R><D>/AABB;

Le premier terme permet d'afficher les formes originales des entrées. Le second terme permet de produire de façon automatique leurs formes réduplicatives :

Au départ, le curseur se trouve après le mot dissyllabique **AB** associé à la propriété **FLX=Gancui**. La commande <L> fait reculer le curseur d'un caractère, et le place donc après **A**. La commande <D> sert à doubler **A**. La commande <R> fait alors avancer le curseur d'un caractère, ce qui le positionne après **B**. La commande <D> sert alors à doubler **B**. Cette description formelle vise à réaliser la forme réduplicative **AABB**, à partir du mot initial **AB**. Les formes réduplicatives des exemples ci-dessus sont les suivantes :

恩恩怨怨
風風雨雨
安安靜靜
乾乾脆脆
切切實實
原原本本
上上下下
前前後後
千千萬萬
許許多多[45]

切實,D+FLX=Gancui *qièshí* <juste-réel> 'réel'
原本,D+FLX=Gancui *yuánběn* <source-racine> 'origine'
上下,F+FLX=Gancui *shàngxià* <supérieur-inférieur> 'supérieurs et inférieurs'
前後,F+FLX=Gancui *qiánhòu* <avant-après> 'avant et après : autour de'
千萬,M+FLX=Gancui *qiānwàn* <mille-dix mille> 'des milliers et des milliers'
許多,M+FLX=Gancui *xǔduō* <environ-nombreux> 'nombreux'

[45] 恩恩怨怨 *ēn'ēn-yuànyuàn* 'beaucoup de bienveillance et de haine'
風風雨雨 *fēngfēng-yǔyǔ* 'pluies et vents : beaucoup de difficultés'
安安靜靜 *ān'ān-jìngjìng* 'très silencieux'
乾乾脆脆 *gāngān-cuìcuì* 'très net'
切切實實 *qièqiè-shíshí* 'absolument certain'
原原本本 *yuányuán-běnběn* 'du commencement à la fin ; depuis A jusqu'à Z'
上上下下 *shàngshàng-xiàxià* 'supérieurs et inférieurs : membres d'une communauté'
前前後後 *qiánqián-hòuhòu* 'depuis le commencement jusqu'à la fin : les uns à la suite des autres'
千千萬萬 *qiānqiān-wànwàn* 'un nombre incalculable'
許許多多 *xǔxǔ-duōduō* 'un grand nombre'

5.8.4 La réduplication ABAB

Si un adjectif, un numéral ou un adjectif à valeur descriptive est composé de deux morphèmes **A** et **B**, il peut adopter la forme réduplicative **ABAB**. Par exemple, la forme réduplicative 充分充分 *chōngfèn chōngfèn* 'très complet' est dérivée de l'adjectif 充分 *chōngfèn* <rempli-partie> 'complet'. Dans le dictionnaire ***ChDic***, nous avons associé la propriété **FLX=Chongfen** aux entrées concernées :

充分,A+FLX=Chongfen
熱鬧,A+FLX=Chongfen
千萬,M+FLX=Chongfen
許多,M+FLX=Chongfen
通紅,Z+FLX=Chongfen
火熱,Z+FLX=Chongfen[46]

La règle de la réduplication **ABAB** est décrite formellement dans la grammaire flexionnelle de ***NooJ Reduplicative forms.nof*** :

Chongfen = <E>/Std + <D2>/ABAB;

Le premier terme permet d'afficher les formes originales des entrées. Dans le second terme, la commande <D2> permet de redoubler ces mots dissyllabiques dont les composants morphologiques sont **A** et **B.** En appliquant cette règle aux entrées suivies de la propriété **FLX=Chongfen**, on obtient automatiquement la forme réduplicative de ces mots :

充分充分
熱鬧熱鬧
千萬千萬
許多許多
通紅通紅
火熱火熱[47]

[46] 充分,A+FLX=Chongfen *chōngfèn* <rempli-partie> 'complet'
熱鬧,A+FLX=Chongfen *rènào* <chaud-bruyant> 'animé ; mouvementé'
千萬,M+FLX=Chongfen *qiānwàn* <mille-dix mille> 'des milliers et des milliers'
許多,M+FLX=Chongfen *xǔduō* <environ-nombreux> 'nombreux'
通紅,Z+FLX=Chongfen *tōnghóng* <uni-rouge> 'tout rouge'
火熱,Z+FLX=Chongfen *huǒrè* <feu-chaud> 'chaud come le feu : brûlant'

[47] 充分充分 *chōngfèn chōngfèn* 'très complet'
熱鬧熱鬧 *rè'nào rè'nào* 'rendre animé'

Les formes réduplicatives peuvent appartenir ou non à la catégorie des mots initiaux. Par exemple, la forme réduplicative 通紅通紅 *tōnghóng tōnghóng* reste un adjectif à valeur descriptive, comme l'était le mot de départ 通紅 *tōnghóng*, ce qui n'est pas le cas pour le mot 熱鬧 *rènào* qui est, au départ, un adjectif, mais dont la forme réduplicative 熱鬧熱鬧 *rènào rènào* devient un verbe[48]. De plus, la réduplication **ABAB**, comme la réduplication **AABB**, conserve également le sens propre du mot de départ, mais en le rendant plus fort en l'intensifiant.

5.9 Conclusion

La formalisation du chinois moderne commence par le recensement et la description de ses Unités Linguistiques Atomiques. Ces unités sont intégrées dans des dictionnaires dans lesquels chaque entrée est associée à des propriétés linguistiques (morphologiques, syntaxiques ou sémantiques). Ces propriétés sont utilisées lors du développement des grammaires, que nous décrirons dans le septième chapitre. À l'occasion de cette étude, nous avons construit six dictionnaires électroniques en appliquant les critères exploités dans les sections 5.4 et 5.5. Ces six dictionnaires sont ***ChDic*** (Chinese Dictionary), ***DicBkTitl*** (Dictionary of Book Titles), ***DicChSurn*** (Dictionary of Chinese Surnames), ***DicExpres*** (Dictionary of Expressions), ***GeoDic*** (Geographical Dictionary) et ***DicProNam*** (Dictionary of Proper Names). Ils sont utilisés pour étiqueter notre corpus et pour développer des grammaires susceptibles de décrire la structure des syntagmes ou des phrases. Nous avons également développé des descriptions formelles pour produire de façon automatique les quatre sortes de réduplications.

千萬千萬 *qiānwàn qiānwàn* 'un nombre incalculable'
許多許多 *xǔduō xǔduō* 'un grand nombre'
通紅通紅 *tōnghóng tōnghóng* 'tout rouge'
火熱火熱 *huǒrè huǒrè* 'tout brûlant'

[48] *cf.* n. 40.

Chapitre 6

NON-CORRESPONDANCE ENTRE LES MORPHEMES ET LEURS GRAPHIES

Nous préciserons d'abord l'instabilité qui existe entre les éléments linguistiques tels que syllabes, morphèmes ou mots et leurs graphies. Par exemple, un morphème pouvant s'écrire selon plusieurs graphies, ou bien, inversement, deux ou plusieurs morphèmes sont représentés par une seule forme graphique. La relation entre deux éléments linguistiques n'est donc pas bi-univoque [*cf.* Zhitang Yang-Drocourt, 2007 : 345-362]. Nous aborderons, dans la section 6.1, les cas de non-correspondance entre les morphèmes et les graphies en les distinguant en trois catégories selon leur nature morphologique.

Dans le cadre du Traitement Automatique des Langues Naturelles, on peut distinguer facilement les homonymes grâce à l'indication de leur catégorie lexicale telle qu'elle apparaît dans le dictionnaire électronique. Par exemple, 花 *huā* est un homonyme homographe et homophone. Il renvoie au moins à trois mots. Lorsqu'il a le sens de 'fleur', c'est un nom. Lorsqu'il a le sens de 'dépenser', c'est un verbe. Lorsqu'il a le sens de 'bigarré', c'est un adjectif[49]. On ne les distingue pas graphiquement puisqu'ils ont la même graphie. Pour différencier ces trois mots de même graphie, on signale la catégorie spécifique à chacun d'eux. Chaque entrée représente un mot :

花,N
花,V
花,A

Néanmoins, il arrive que des graphies différentes soient traitées comme des unités distinctes, bien qu'elles ne représentent qu'un seul morphème. Par exemple, les graphies 僞 et 僞 représentent toutes deux le morphème qui a le sens de 'faux' et se prononcent *wèi* dans les deux cas. De ce fait, les différentes formes graphiques représentant un seul morphème, multiplient le nombre d'entrées dans les dictionnaires. Pour éviter cette multiplication, il est nécessaire de faire correspondre les différentes formes graphiques à une seule unité morphologique. Autrement dit, le programme informatique doit

[49] En tant qu'adjectif, le mot 花 *huā* peut avoir plusieurs sens. Nous ne tenons pas compte, ici, de ces problèmes de polysémie et nous avons choisi le sens de 'bigarré' pour notre exemple.

comprendre qu'un morphème peut s'écrire avec une ou plusieurs graphies et il doit être capable d'"unifier" ces différentes graphies en les reliant au seul morphème.

L'établissement de cette correspondance dépend d'une référence, c'est-à-dire, d'une forme graphique qui sera choisie comme forme "standard". Le choix de la graphie standard ne se fait pas au hasard. Il s'effectue selon les critères que nous décrirons dans la section 6.2. Dans la section 6.3, nous présenterons donc les applications qui permettent à ***NooJ*** de relier les différentes graphies à une seule unité morphologique.

6.1 Non-correspondance entre les morphèmes et leurs graphies

6.1.1 Morphèmes monosyllabiques

6.1.1.1 Homonymes

6.1.1.1.1 Homonymes homophones

Une syllabe chinoise peut représenter différents morphèmes. Par exemple, la syllabe *máo* représente en mandarin au moins sept morphèmes, qui chacun possède sa propre graphie, c'est-à-dire, un caractère unique :

(163) 毛 *máo* 'poil'
矛 *máo* 'lance'
茅 *máo* 'nom de diverses plantes herbacées, en particulier des *Imperata*'
旄 *máo* 'queue de yack fixée au sommet d'une hampe'
髦 *máo* 'touffe de cheveux'
蟊 *máo* 'ver rongeur de céréales'
錨 *máo* 'ancre'

Ces sept morphèmes, qui se prononcent de la même façon en mandarin mais qui ont sept significations différentes et se distinguent par leurs formes graphiques sont appelés des homonymes homophones.

6.1.1.1.2 Homonymes homographes et homophones

En chinois, certains morphèmes ont la même prononciation, la même graphie, mais des sens différents. Ainsi 花 *huā* qui, selon le contexte, signifie :

● fleur,
● bigarré,
● dépenser ;

ou 明 *míng* qui représente les cinq morphèmes suivants :

● briller,
● point du jour,
● distinguer clairement,
● comprendre,
● la dynastie des Ming (1368 — 1644).

Ce sont des homonymes homophones et homographes.

6.1.1.1.3 Homonymes homographes

Il existe des morphèmes qui se différencient par leur prononciation mais qui ont la même forme graphique. Ce sont des homonymes dits homographes. En voici quelques exemples regroupés en trois types selon leur prononciation :

1) Homonymes homographes se prononçant avec le même son en modifiant leur ton ou leur consonne :

(164) a. 使 *shǐ* ‘ordonner’
b. 使 *shǐ / shì* ‘ambassadeur’

(165) a. 少 *shǎo* ‘peu’
b. 少 *shào* ‘jeune’

(166) a. 悶 *mèn* ‘mélancolique’
b. 悶 *mēn* ‘étouffant ; s’enfermer’

(167) a. 相 *xiāng* ‘mutuel’
b. 相 *xiàng* ‘apparence physique ; observer’

(168) a. 空 *kōng* ‘vide’
b. 空 *kòng* ‘intervalle’

(169) a. 背 *bèi* ‘dos’
b. 背 *bēi* ‘porter sur le dos’

(170) a. 長 *cháng* ‘long’
b. 長 *zhǎng* ‘agrandir ; naître’

2) Homonymes homographes ayant en commun leur consonne mais se différenciant par leur voyelle :

(171) a. 和 *hé* ‘somme ; s’accorder’
b. 和 *hè* ‘répondre à un chant par un autre chant’
c. 和 *huo* ‘température douce’
d. 和 *hé / hàn* ‘avec’
e. 和 *huò* ‘mélanger’

f. 和 *huó* ‘ajouter de l’eau à de la farine ou à de la glaise pour les travailler’
g. 和 *hú* ‘mot lancé quand on gagne une partie, en particulier au mah-jong’

(172) a. 還 *hái* ‘de nouveau’
b. 還 *huán* ‘retourner ; rendre’

3) Homonymes homographes : se prononcent différemment :

(173) a. 塞 *sāi* ‘bouchon’
b. 塞 *sài* ‘région frontière’
c. 塞 *sè* ‘bloquer’

(174) a. 差 *chā* ‘erreur’
b. 差 *chà* ‘s’écarter de’
c. 差 *chāi* ‘service commandé par l’autorité’

(175) a. 更 *gēng* ‘changer’
b. 更 *gèng* ‘encore’
c. 更 *jīng* ‘veille (chacune des cinq veilles de la nuit, de deux heures chacune, de dix-neuf heures à cinq heures)’

6.1.1.2 Variantes graphiques

Dans le système d’écriture chinois, certains morphèmes peuvent être représentés par deux formes graphiques au moins, tout en gardant la même prononciation et la même signification. Ci-dessous, nous nous intéresserons à trois sortes de variantes graphiques de morphèmes monosyllabiques.

6.1.1.2.1 Variantes typographiques

Certains morphèmes monosyllabiques offrent des variantes de formes graphiques qui sont typographiques. Il peut s’agir d’une orientation différente des traits. Par exemple, le morphème *duì* ‘échanger’ peut être représenté par les deux variantes graphiques : 兌 et 兑. Leur différence, minime, réside dans l’orientation des traits supérieurs de la graphie, qui sont représentés soit comme ceci : 八, soit comme cela : 丷.

Ces deux graphies se retrouvent en tant que composants dans la construction de morphèmes monosyllabiques plus complexes. Par exemple, les morphèmes *yùe* ‘joyeux’ [*cf*. (177a-b)] et *shuō* ‘parler’ [*cf*. (178a-b)] s’écrivent chacun avec deux formes graphiques différentes dont certains composants renvoient aux deux graphies du morphème *duì*.

(176) a. 悅
b. 悦
(177) a. 說
b. 説

6.1.1.2.2 Variantes dispositionnelles

Les composants graphiques d'un morphème peuvent être disposés de manière verticale ou horizontale. En d'autres termes, un morphème chinois peut avoir différentes graphies qui se différencient par leur disposition compositionnelle. Cette différence dispositionnelle ne modifie ni la prononciation ni la signification du morphème concerné. Examinons les deux graphies possibles du morphème *qún* 'groupe'.

Ce morphème monosyllabique *qún* 'groupe' est formé d'un composant phonologique 君 *jūn* 'souverain' et d'un composant sémantique 羊 *yáng* 'mouton'. Ces deux composants graphiques peuvent s'organiser horizontalement afin de former la graphie 群 du morphème *qún* [*cf*. (179a)]. Celui-ci peut aussi se constituer selon un axe vertical par superposition du composant phonologique sur le composant sémantique : 羣 [*cf*. (179b)].

(178) a. 君羊 = 群

b. 群 →  →  = 羣

6.1.1.2.3 Variantes compositionnelles

Certains morphèmes possèdent des variantes graphiques, qui diffèrent par leur composition. Elles peuvent se classer en deux types.

1) Les différentes formes graphiques d'un morphème monosyllabique se définissent comme des allographes qui ont la même prononciation et la même signification [*cf*. Li Xingjian et Yu Zhihong, 2005 : 89-97]. C'est ainsi que :

(179) le morphème *chuáng* 'lit' s'écrit 床 ou 牀 ;
le morphème *pào* 'canon' s'écrit 砲 ou 炮 ;
le morphème *lèi* 'larme' s'écrit 淚 ou 泪 ;
le morphème *qiáng* 'mur' s'écrit 牆 ou 墻 ;
le morphème *tán* 'jarre' s'écrit 罈, 壜, 罈, 罎 ou 醰.

2) Comme nous venons de le voir, certaines graphies chinoises, comme celle du morphème *qún* 'groupe', sont constituées d'un composant phonologique et d'un composant sémantique. Un morphème qui s'écrit avec une graphie de ce type peut avoir différents composants sémantiques. Par exemple, le morphème monosyllabique *biē* 'trionyx' peut se présenter sous trois formes différentes :

(180) 

Elles se différencient seulement par leur composant sémantique. Leur composant phonologique 敝 *bì* demeure identique. Leurs composants sémantiques sont les suivants :

魚 *yú* ‘poisson’,
黽 *mǐn* ‘grenouille’
龜 *guī* ‘tortue’

Nous appelons ce type de variantes graphiques : variante de composition sémantique, car elles sont formées de différents composants sémantiques, ce qui permet d’interpréter de différentes manières le sens de morphèmes dans l’écriture.

6.1.2 Morphèmes polysyllabiques

6.1.2.1 Mots formés par deux syllabes liées

On sert des caractères pour écrire les mots formés de deux syllabes liées (en chinois 聯綿詞 *liánmiáncí*). L’objectif étant de noter leur prononciation à l’écrit, le choix de leurs formes graphiques est resté libre, si bien qu’ils possèdent plusieurs graphies. Exemples :

(181) *fǎngfú* ‘semblable’ possède trois graphies 仿佛, 彷彿 ou 髣髴.

cānghuáng ‘précipité’ peut s’écrire avec au moins cinq formes graphiques : 倉惶, 倉皇, 倉遑, 倉黃 ou 蒼黃.

gēdā ‘tertre’ possède au moins quatre graphies : 圪塔, 圪墶, 圪瘩 ou 屹塔.

6.1.2.2 Mots réduplicatifs

En chinois, certains morphèmes sont constitués par la réduplication en redoublant leur forme simple. Ils peuvent s’écrire avec des formes graphiques différentes. Exemples :

(182) Le morphème *yángyáng* signifiant ‘satisfait’ s’écrit selon les deux graphies 洋洋 ou 揚揚.

Le morphème *mòmò* signifiant ‘en grand nombre’ s’écrit selon les deux graphies 脉脉 ou 脈脈.

Le morphème *shēnshēn* signifiant ‘cœur débordant d’une affection muette’ s’écrit selon les deux graphies 莘莘 ou 甡甡.

6.1.2.3 Mots résultant de la transposition phonologique de mots étrangers

En chinois moderne, certains mots sont issus de la transposition phonologique de mots étrangers [*cf.* 5.1.1.1.2]. De ce fait, ils s’écrivent selon différentes formes graphiques, chaque caractère ayant été choisi pour correspondre phonologiquement à une syllabe du mot traduit. Tel est le cas de certains noms de pays empruntés à l’anglais. Exemples :

(183) « *Canada* » peut s’écrire selon les trois graphies 佳拿大, 加拿大 ou 家拿大. Ces trois graphies trisyllabiques se prononcent identiquement *Jiānádà*.

« *Italy* » qui se prononce en chinois *Yìdàlì* peut s’écrire 義大利 ou 意大利.

« *Singapore* » qui se prononce en chinois *Xīngjiāpō* peut être transposé de deux manières différentes : 新加坡 ou 星加坡.

6.1.2.4 Onomatopées polysyllabiques

Les variantes graphiques concernent aussi certaines onomatopées. Exemples :

(184) L’onomatopée *pūchī*, qui imite le ‘son produit par une personne qui pouffe de rire’ s’écrit selon les quatre formes graphiques 撲哧, 噗哧, 撲嗤 et 噗嗤.

L’onomatopée *pūtōng* ‘plouf imitant le bruit d’un objet qui tombe dans l’eau’ s’écrit selon les graphies 噗通 ou 噗咚.

6.1.3 Mots composés

Il arrive qu’un mot composé ait plusieurs graphies. Par exemple le mot dissyllabique *bǎomǔ* ‘nourrice’ possède les six formes graphiques suivantes :

(185) a. 保母
b. 保姆
c. 媬母

d. 媬姆
e. 褓母
f. 褓姆

Ce mot se compose de deux morphèmes *bǎo* et *mǔ*. Le premier morphème *bǎo* peut s'écrire avec trois caractères différents : 保, 媬 et 褓, qui se prononcent de la même manière et ont un sens similaire. La graphie 保 signifie 'protéger'. La graphie 媬 signifie 'femme qui instruit des enfants aristocratiques'. La graphie 褓 a le sens de 'langes' et évoque aussi la notion de protection. En raison de leur similitude sémantique ils peuvent devenir les composants graphiques du mot *bǎomǔ*.

Le deuxième morphème *mǔ* peut être écrit 母 ou 姆, qui ont la même prononciation et un sens similaire. La graphie 母 fait référence à 'une personne de sexe féminin qui allaite'. La graphie 姆 désigne 'une quinquagénaire qui dispense des soins de femme à d'autres personnes de sexe féminin'. Du point de vue sémantique, ces deux composants sont très proches, et peuvent en conséquence être intervertis dans les graphies du mot *bǎomǔ* 'nourrice'.

Il arrive aussi qu'un seul des composants du mot dissyllabique diffère. Ainsi *méiyǔ*, s'écrit selon deux graphies :

(186) a. 梅雨 <prune-pluie>
b. 霉雨 <moisissure-pluie>

Le mot *méiyǔ* renvoie à la saison des pluies, de juin à juillet, durant laquelle les prunes mûrissent, mais aussi durant laquelle la moisissure attaque certains objets. La graphie sera choisie en fonction du contexte à privilégier, soit avec la graphie 梅雨, avec l'idée que les prunes mûrissent, soit avec la graphie 霉雨 pour insister sur l'apparition de la moisissure.

La composition de certaines graphies peut obéir à des raisons plus abstraites. Considérons les deux variantes graphiques du mot *gǔdǒng* 'antiquité' :

(187) a. 古董
b. 骨董

Ce mot composé de deux morphèmes *gǔ* et *dǒng*. Le deuxième morphème *dǒng* s'écrit dans les deux cas avec la même graphie 董. Mais la graphie du premier composant morphologique *gǔ* est tantôt 古, tantôt 骨 qui ont la même prononciation mais un sens

différent. La graphie 古 signifie ‘ancien’. La graphie 骨 ‘os’. Pour former graphiquement le mot *gǔdǒng*, leurs significations deviennent éventuellement une référence sémantique. Ainsi, la forme graphique 骨董 est relativement plus abstraite que celle 古董 mais toutes deux font référence à la notion d’‘antiquité’.

6.2 Critères de standardisation

Dans cette section, nous présenterons les trois critères que nous avons appliqués pour choisir telle forme graphique comme standard, et lui faire correspondre ses variantes. Ces trois critères ont été définis par le Comité national des langues du Ministère de l’Éducation de la République populaire de Chine quand il a publié la liste des graphies standard, sous le titre « *The First Series of Standardized Forms of Words with Non-standardized Variant Forms* »[50]. L’application pratique de ce texte date du 31 mars 2002. [*cf. China Education and Research Network* : http://www.edu.cn/20011228/3015609.shtml].

6.2.1 Critère de fréquence

Comment choisir un standard parmi les diverses formes graphiques ? Les standards changent avec le temps et selon les utilisateurs. Le critère de fréquence est décrit ainsi :

> *“En se fondant sur la statistique scientifique de la fréquence des occurrences et sur l’investigation sociale, on sélectionne les formes lexicales qui sont les plus régulièrement employées et on les choisit comme standard. Le principe de l’usage le plus fréquent est considéré comme le critère le plus important, puisque la règle linguistique est établie par l’habitude conventionnelle langagière d’une société. D’après certaines recherches, il existe une différence notoire dans la fréquence des occurrences pour 90 % des variantes lexicales. Les formes lexicales sélectionnées selon le critère de la fréquence de l’usage s’accordent parfaitement avec le critère de rationalité, etc. Bien que peu de formes lexicales de fréquence élevée ne soient pas complètement soumises à leurs règles étymologiques ou linguistiques, elles deviennent des formes lexicales standard par convention. De ce fait, on doit respecter le choix de cette société.”*[51]

[50] « *The First Series of Standardized Forms of Words with Non-standardized Variant Forms* » est le titre anglais traduit du titre chinois 第一批异形词整理表 *dì-yī pī yìxíngcí zhěnglǐbiǎo*.

[51] 根据科学的词频统计和社会调查，选取公众目前普遍使用的词形作为推荐词形。把通用性原则作为整理异形词的首要原则，这是由语言的约定俗成的社会属性所决定的。据多方考察，90%以上的常见异形词在使用中词频逐渐出现显著性差异，符合通用性原则的词形绝大多数与理据性等

Les résultats obtenus à partir des enquêtes statistiques[52] suggèrent donc le choix d'une forme standard. Par exemple nous avons déterminé la forme standard du mot *biăndòu* 'haricot' qui se compose de deux morphèmes dont le premier possède quatre formes graphiques.

(188)

	Graphie standard
a.	扁豆
	Variantes graphiques
b.	萹豆
c.	藊豆
d.	稨豆

Selon l'enquête statistique, la forme graphique 扁豆 est plus souvent utilisée que les trois autres. De ce fait, elle est considérée comme la forme standard du mot *biăndòu* 'haricot'.

6.2.2 Principe de rationalité

Quand l'étude statistique ne permet pas d'aboutir à des conclusions claires, on doit recourir à d'autres critères, dont celui de la rationalité. Le Comité national des langues présente ainsi ce critère :

> *"Certains mots qui s'écrivent avec des graphies différentes sont moins utilisées, et la fréquence de leurs occurrences ne présente pas de différence notable. Il est naturellement difficile de décider des formes lexicales standard à travers le principe de la fréquence de l'usage. Ainsi, c'est par l'étude linguistique qu'on décidera de l'adoption de telle ou telle forme lexicale. Ce*

原则是一致的。即使少数词频高的词形与语源或理据不完全一致，但一旦约定俗成，也应尊重社会的选择。

gēnjù kēxué de cípín tǒngjì hé / hàn shèhuì diàochá, xuǎnqǔ gōngzhòng mùqián pǔbiàn shǐyòng de cíxíng zuòwéi tuījiàn cíxíng. bǎ tōngyòngxìng yuánzé zuòwéi zhěnglǐ yìxíngcí de shǒuyào yuánzé, zhè shì yóu yǔyán de yuēdìng-súchéng de shèhuì shǔxìng suǒ juédìng de. jù duōfāng kǎochá, 90% yǐshàng de chángjiàn yìxíngcí zài shǐyòng zhōng cípín zhújiàn chūxiàn xiǎnzhùxìng chāyì, fúhé tōngyòngxìng yuánzé de cíxíng jué dàduōshù yǔ lǐjùxìng děng yuánzé shì yīzhì de. jíshǐ shǎoshù cípín gāo de cíxíng yǔ yǔyuán huò lǐjù bù wánquán yīzhì, dàn yīdàn yuēdìng-súchéng, yě yīng zūnzhòng shèhuì de xuǎnzé.

[52] Nous avons réalisé l'étude statistique d'après nos données littéraires et journalistiques et en utilisant le moteur de recherche Google. Nous avons aussi eu recours aux dictionnaires 康熙字典 *Kāngxī zìdiăn* '*Dictionnaire de Kangxi*' [1996], 說文解字注 *Shuōwén jiězì zhù* '*Shuowen jiezi et son interprétation*' [2005], 漢語大詞典 *Hànyǔ dàcídiăn* '*Grand dictionnaire chinois*' [1994], 辭海 *Cíhăi* '*Lexique chinois*' [1986] (éditions taïwanaises) et 现代汉语规范字典 *Xiàndài hànyǔ guīfàn zìdiăn* '*Dictionnaire de normalisation du chinois moderne*' [1998].

choix de la forme standard permet de mieux comprendre la signification de ces standards et de faciliter leur application dans les textes."[53]

Prenons l'exemple du mot *fùxí* 'réviser' qui s'écrit avec deux graphies :

(189) a. 複習
b. 復習

L'étude statistique montre que la forme graphique 復習 a une fréquence bien plus élevée que 複習. Pourtant, elle n'a pas été choisie comme forme standard. La raison en est que la forme graphique 複習 est "rationnellement plus correcte" que la forme 復習. La graphie morphologique 複 *fù* signifie 'répéter', tandis que la graphie morphologique 復 *fù* a le sens de 'retourner'. Le mot *fùxí*, au sens de 'réviser', illustre l'idée de lecture répétitive. La graphie morphologique 複 convient donc mieux que 復 pour représenter, à l'écrit, cette idée. Ainsi, le facteur qui a déterminé le choix du standard est ici la signification morphologique, et la mesure statistique devient un facteur secondaire.

6.2.3 Symétrie systématisée entre les morphèmes et les mots

Il arrive qu'une graphie soit un composant morphologique commun à des séries de mots. La fréquence de ces mots diffère probablement d'une série à l'autre. Le choix des formes standard nécessite donc une symétrie harmonique dans chaque série de mots. Le critère de symétrie est ainsi présenté :

"Il existe une harmonie solide systématique à l'intérieur du vocabulaire. Quand on normalise les différentes formes graphiques de mots, la symétrie de formes graphiques dans les séries de mots, qui partagent les mêmes morphèmes, doit également être prise en compte. "[54]

Observons trois groupes de mots [*cf.* (191)-(193)] qui ont en commun deux graphies morphologiques : 畫 *huà* 'dessiner' et 劃 *huà* 'délimiter'.

[53] 某些异形词目前较少使用，或词频无显著性差异，难以依据通用性原则确定取舍，则从词语发展的理据性角度推荐一种较为合理的词形，以便于理解词义和方便使用。
mǒuxiē yìxíngcí mùqián jiào shǎo shǐyòng, huò cípín wú xiǎnzhùxìng chāyì, nányǐ yījù tōngyòngxìng yuánzé quèdìng qǔshě, zé cóng cíyǔ fāzhǎn de lǐjùxìng jiǎodù tuījiàn yī zhǒng jiào wéi hélǐ de cíxíng, yǐ biàn yú lǐjiě cíyì hé / hàn fāngbiàn shǐyòng.

[54] 词汇内部有较强的系统性，在整理异形词时要考虑同语素系列词用字的一致性。
cíhuì nèibù yǒu jiào qiáng de xìtǒngxìng, zài zhěnglǐ yìxíngcí shí yào kǎolǜ tóng yǔsù xìliècí yòngzì de yīzhìxìng.

(190) a. 筆畫 *bǐhuà* <crayon-dessiner> ‘trait’
b. 筆劃 *bǐhuà* <crayon-délimiter> ‘trait’
(191) a. 籌畫 *chóuhuà* <projeter-dessiner> ‘projeter’
b. 籌劃 *chóuhuà* <projeter-délimiter> ‘projeter’
(192) a. 計畫 *jìhuà* <compter-dessiner> ‘programmer’
b. 計劃 *jìhuà* <compter-délimiter> ‘programmer’

Les graphies 畫 et 劃 constituent les morphèmes communs à ces trois groupes de mots. Les différences de fréquence de chacun des trois groupes apparaissent clairement dans le tableau ci-desous :

		Fréquence élevée (graphies standard)		Fréquence basse (variantes graphiques)
(193)	a.	筆畫	b.	筆劃
(194)	a.	籌劃	b.	籌畫
(195)	a.	計劃	b.	計畫

Le mot *bǐhuà* a deux formes graphiques : 筆畫 et 筆劃. La première forme 筆畫 qui comporte la graphie 畫 présente une fréquence plus élevée que la deuxième 筆劃. Mais, en ce qui concerne *chóuhuà* et *jìhuà*, les statistiques prouvent l’inverse : les occurrences de 籌劃 et de 計劃 sont bien plus fréquentes que celles de 籌畫 et de 計畫. Dans le cadre de la systématisation des formes graphiques standard, les formes standard peuvent avoir différentes graphies morphologiques, comme en témoignent les exemples mentionnés ci-dessus. La forme graphique 筆畫 est considérée comme la forme standard du mot *bǐhuà*. La forme standard du mot *chóuhuà* est 籌劃. Celle du mot *jìhuà* est 計劃 et tous les mots qui contiennent ces deux graphies morphologiques doivent respecter cette symétrie de la standardisation de leurs formes graphiques :

		Graphies standard		Variantes graphiques	Prononciations et significations
(196)	a.	計劃書	b.	計畫書	*jìhuàshū* <programmer-livre> ‘projet’
(197)	a.	計劃案	b.	計畫案	*jìhuààn* <programmer-dossier> ‘programme’
(198)	a.	計劃圖	b.	計畫圖	*jìhuàtú* <programmer-dessin> ‘plan local d’urbanisme’

6.3 Solutions dans *NooJ*

6.3.1 Traitement des variantes de caractères[55]

Quand la non-correspondance entre un morphème et sa graphie concerne un monosyllabe écrit avec un seul caractère, la solution que nous avons apportée a consisté à établir une liste de variantes des formes graphiques. Dans cette liste, nous avons associé les variantes graphiques aux standards à travers la formule **A**:**B**. La première graphie **A** représente la forme graphique variante. La deuxième graphie **B** est considérée comme la graphie standard employée dans nos dictionnaires et nos grammaires :

Variante graphique : Graphie standard	Prononciations et significations
冊:册	*cè* ‘registre’
喑:哑	*yǎ* ‘muet’
壠:壟	*lǒng* ‘talus’
朵:朵	*duǒ* ‘quantifieur réservé aux fleurs’
泪:泪	*lèi* ‘larme’
煑:煮	*zhǔ* ‘cuire à l’eau’
牀:床	*chuáng* ‘lit’
眞:真	*zhēn* ‘réel’
説:說	*shuō* ‘parler’
鵝:鹅	*é* ‘oie domestique’

Figure 28 : Extrait de la liste de variantes des caractères

Dans nos dictionnaires et grammaires, nous n’avons entré que les graphies standard, ainsi que les mots se composant de ces graphies standard :

一分為二,I
以為,V
為,VG[56]

Figure 29 : Extrait du dictionnaire ***ChDic***

Quelquefois, plusieurs variantes graphiques correspondent à une graphie standard. Les variantes de graphies monosyllabiques sont listées et se réfèrent à la même graphie standard :

[55] Dans ce traitement, un caractère chinois peut représenter graphiquement, une syllabe, un morphème ou un mot.

[56] 一分為二,I *yìfēnwéi'èr* ‘trancher ou diviser en deux’
以為,V *yǐwéi* ‘croire’
為,VG *wéi* ‘diriger’

Variante graphique : Graphie standard	Prononciations et significations
鬥:鬥	*dòu* ‘lutter’
鬪:鬥	*dòu* ‘lutter’
鬭:鬥	*dòu* ‘lutter’

Figure 30 : Exemple de trois variantes de caractère correspondant à un seul caractère standard

Les variantes de caractères de la première colonne seront remplacées par la graphie standard de la deuxième colonne. Ce remplacement se fait automatiquement, lorsqu’on importe le fichier sous ***NooJ*** :

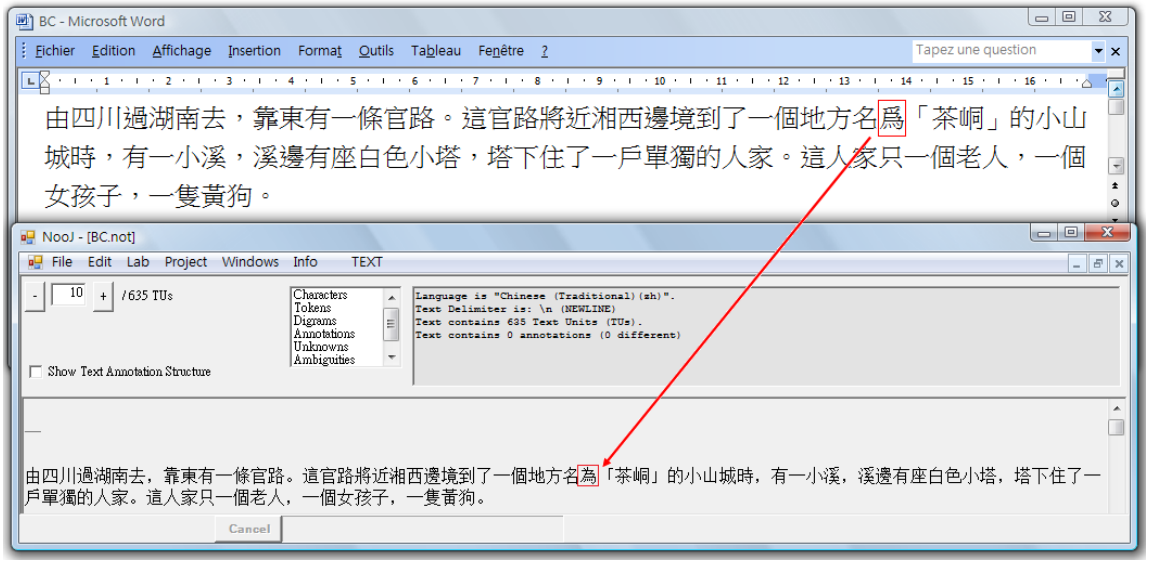


Figure 31 : Importation d’un texte dans ***NooJ***

Lorsqu’on analyse un texte dans lequel se présentent des variantes de graphies monosyllabiques, ***NooJ*** remplace chaque variante graphique par sa graphie standard. Cette démarche s’effectue de façon transparente, avant la consultation des dictionnaires et l’application des grammaires.

6.3.2 Standardisation et formalisation des unités lexicales[57]

Si des unités lexicales s’écrivent avec plusieurs graphies, nous les traitons directement dans nos six dictionnaires construits. Les différentes variantes graphiques de ces mots y sont intégrées. Chaque forme graphique est représentée par une entrée lexicale, et toutes les variantes sont explicitement reliées à la forme standard. Par exemple, le mot *bǎomǔ*

[57] Dans ce traitement, les unités lexicales peuvent être des polysyllabes monomorphémiques, des mots composés, etc.

‘nourrice’ s’écrit avec six formes graphiques définies 保母, 保姆, 媬母, 媬姆, 褓姆 et 褓母. Il ne peut s’écrire pas avec les graphies *葆母 ou *葆姆. Chaque entrée de variante est automatiquement suivie de sa forme standard, comme il apparaît ci-dessous :

保姆,N+Hum
保母,保姆,N
媬母,保姆,N
媬姆,保姆,N
褓母,保姆,N
褓姆,保姆,N

NooJ relie automatiquement les variantes à une entrée lexicale donnée. Les variantes du mot *bǎomǔ* ‘nourrice’ 保母, 媬母, 媬姆, 褓母 et 褓姆 constituent également les entrées de dictionnaire, elles sont associées à la forme standard 保姆.

La graphie 保姆 est choisie comme forme standard dans le dictionnaire ***ChDic***. ***NooJ*** va, tout d’abord, faire correspondre 保姆 à ses cinq variantes : 保母, 媬母, 媬姆, 褓母 et 褓姆. Puis l’annotation lexicale verra ses formes variantes remplacées par la forme standard, comme l’annotation dans le journal du 26 décembre 2007 :

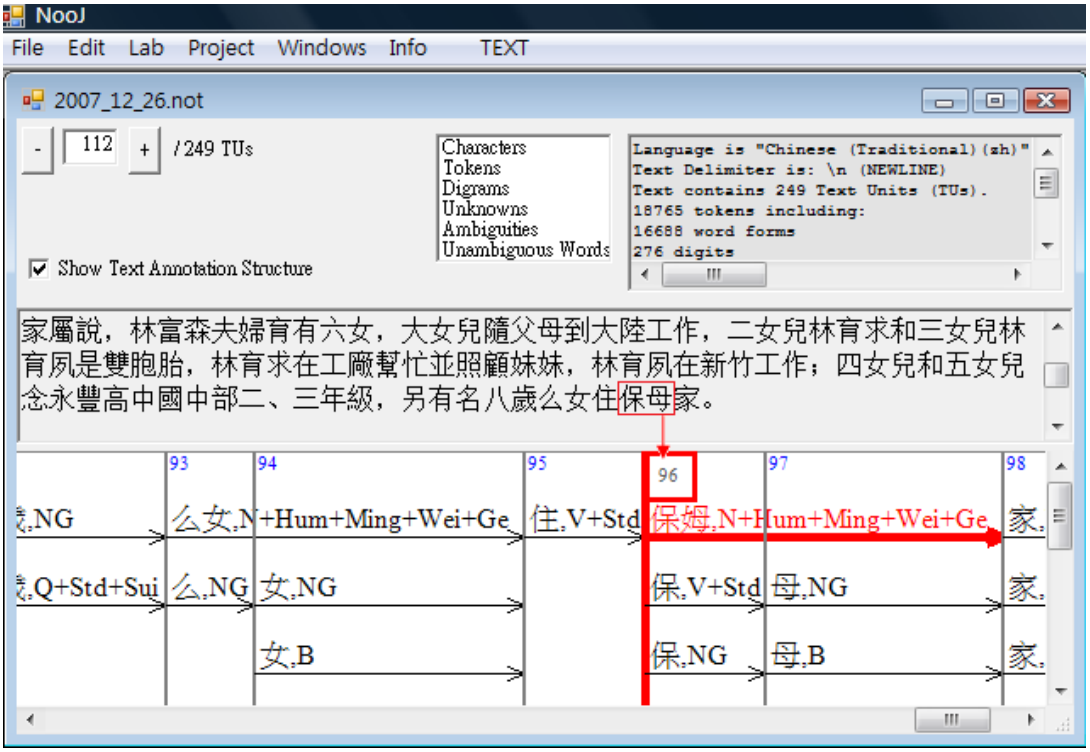


Figure 32 : Analyse du mot 保母 *bǎomǔ* ‘nourrice’

6.4 Conclusion

Dans ce chapitre, nous avons d'abord étudié la non-correspondance entre les morphèmes et leurs formes graphiques. Puis, nous avons retenu, concrètement, trois critères pour la sélection des formes standard parmi les variantes : 1) critère de fréquence ; 2) principe de rationalité ; 3) symétrie systématisée entre les morphèmes et les mots.

Lorsque nous avons appliqué ces trois critères pour déterminer les formes standard, nous avons établi une liste de variantes des graphies morphologiques monosyllabiques. À partir de cette liste de variantes, ***NooJ*** peut associer automatiquement toutes les variantes à leur forme standard, avant même de consulter nos dictionnaires. Dans ces derniers, les variantes graphiques des unités lexicales polysyllabiques doivent être associées à leur forme standard. Grâce à cette association, ***NooJ*** peut reconnaître et relier les variantes lexicales à leur forme standard.

Chapitre 7

TRAITEMENT SYNTAXIQUE

Nous présenterons d'abord les six types principaux de syntagmes chinois. Ensuite, nous décrirons des grammaires pour la désambiguïsation syntaxique. Ces grammaires se fondent sur des règles non contextuelles (hors contexte) pour éliminer les ambiguïtés. On les appelle des grammaires locales, car elles décrivent de façon formelle des structures employées localement. Dans cette étude, les grammaires locales sont utilisées pour formaliser des structures syntagmatiques, telles que la composition numérique, les expressions temporelles et les appellations personnelles.

Outre les grammaires locales, il en est d'autres qui permettent de représenter des syntagmes. Nous avons ainsi élaboré des grammaires susceptibles de décrire la structure des cinq types de groupes nominaux noyaux. Nous les préciserons dans la deuxième partie de ce chapitre. Nous montrerons aussi les résultats obtenus lorsqu'on applique les grammaires syntaxiques à la reconnaissance des groupes nominaux noyaux d'un texte donné, soit le roman de Lao She *Quatre générations sous un même toit*.

7.1 Syntagmes

7.1.1 Sujet-Prédicat

Le sujet constitue un agent sur lequel le locuteur veut apporter des informations. Le prédicat décrit des faits que ce sujet met en œuvre. Le sujet peut être un nom, un nom de temps, un nom de lieu, un pronom, un verbe, un adjectif, un adjectif à valeur descriptive, etc. Le prédicat peut aussi être un nom, un nom de temps, un nom de lieu, un pronom, un verbe, un adjectif, un adjectif à valeur descriptive, etc. Si un syntagme est de forme Sujet-Prédicat, le sujet précède toujours le prédicat. Exemples :

(199) 玫瑰漂亮。 *méiguī piàoliàng.* <rose-beau.>
'La rose est belle.'
她洗衣服。 *tā xǐ yīfu / fú.* <elle-laver-vêtement.>
'Elle lave des vêtements.'
後天要來兩位朋友。 *hòutiān yào lái liǎng wèi péngyǒu.*
<après-demain-vouloir-venir-deux-Q-ami.>
'Après-demain deux amis viendront.'

小明北京人。 *Xiǎo Míng Běijing / jīng rén.* <Xiao Ming-Pékinois.>
'Xiao Ming est Pékinois.'
喝水有益健康。 *hēshuǐ yǒuyì jiànkāng.* <boire-eau-avantage-santé.>
'Boire de l'eau est bon pour la santé.'

7.1.2 Verbe-Complément d'objet

La structure relève une dominance du verbe sur le complément d'objet. Le verbe, qui peut être un syntagme verbal, exprime un mouvement ou un comportement. Le complément d'objet répond au verbe. Il est constitué d'éléments substantifs ou prédicatifs. Les éléments substantifs sont des noms, des noms de temps, des noms de lieu, des pronoms, etc. Les éléments prédicatifs sont des verbes, des adjectifs, etc. Exemples :

(200) 吃蘋果 *chī píngguǒ* <manger-pomme> 'manger la pomme'
整理花園 *zhěnglǐ huāyuán* <ranger-jardin> 'ranger le jardin'
參觀巴黎 *cānguān Bālí* <visiter-Paris> 'visiter Paris'
介紹唐朝 *jièshào Tángcháo* <présenter-dynastie des Tang> 'présenter la dynastie des Tang (618 — 907)'
進來一位婦人 *jìnlái yī wèi fùrén* <entrer-un-Q-femme> 'une femme entre'

7.1.3 Verbe-Complément / Adjectif-Complément

Dans un syntagme de forme Verbe-Complément ou Adjectif-Complément, le complément se place après le verbe. Les syntagmes de ce type sont structurés de la même façon que les syntagmes de structure Verbe-Complément d'objet. Mais on remarque que complément et complément d'objet de verbe ont des fonctions différentes. Un complément d'objet complète le procès exprimé par le verbe. Un complément constitue une précision ou un développement qui décrivent le procès que le verbe met en œuvre. Nous présentons, ci-dessous, cinq types de syntagmes Verbe-Complément et le syntagme Adjectif-Complément.

7.1.3.1 Verbe–Complément résultatif

Un complément résultatif a pour but de décrire le résultat produit par un verbe. Il peut être un verbe ou un adjectif suivant immédiatement le verbe principal. Exemples :

(201) Structure Verbe–Verbe
看見 *kànjiàn* <regarder-apercevoir> 'apercevoir'
跌倒 *diédǎo* <trébucher-tomber> 'tomber ; faire une chute'

(202) Strucutre Verbe–Adjectif
曬乾 *shàigān* <exposer au soleil-sec> ‘sécher au soleil’
煮熟 *zhǔshóu* <cuire-(bien) cuit> ‘cuire jusqu’à être cuit’
洗乾淨 *xǐ gānjìng* <nettoyer-propre> ‘nettoyer proprement’

7.1.3.2 Verbe–Complément de direction

Les verbes de direction, tels que 來 *lái* ‘venir’, 去 *qù* ‘aller’, 進 *jìn* ‘entrer’, 出 *chū* ‘sortir’, 上 *shàng* ‘monter’, 下 *xià* ‘descendre’, 回 *huí* ‘rentrer’, 過 *guò* ‘passer’, et 起 *qǐ* ‘lever’, peuvent aussi être les compléments d’un verbe. Par exemple,

(203) 爬上 *páshàng* <grimper-monter> ‘grimper’

Des verbes formés par l’association de deux verbes de direction, comme ceux présentés dans le tableau ci-dessous, peuvent, eux aussi, constituer le complément d’un verbe.

上來 *shànglái* ‘monter’	下來 *xiàlái* ‘descendre’	進來 *jìnlái* ‘entrer’	出來 *chūlái* ‘sortir’	回來 *huílái* ‘rentrer’	過來 *guòlái* ‘passer’	起來 *qǐlái* ‘se lever’	開來 *kāilái* ‘s’échapper’
上去 *shàngqù* ‘monter’	下去 *xiàqù* ‘descendre’	進去 *jìnqù* ‘entrer’	出去 *chūqù* ‘sortir’	回去 *huíqù* ‘rentrer’	過去 *guòqù* ‘passer’		

Les verbes ci-dessus, formés par l’association de deux verbes de direction se définissent comme des compléments complexes de direction. Ils peuvent suivre un autre verbe, pour former des syntagmes de type Verbe-Complément :

(204) 爬上來 *pá shànglái* <grimper-monter> ‘grimper vers le haut’
跑上去 *pǎo shàngqù* <courir-monter> ‘courir vers le haut’
飛起來 *fēi qǐlái* <voler-se lever> ‘voler vers le haut : s’envoler’

Les verbes, 來 *lái* ou 去 *qù*, associés avec un autre verbe, peuvent eux aussi indiquer la direction dans laquelle se fait un mouvement, mais, cette fois, du point de vue du locuteur. Ainsi, on peut exprimer ‘Il entre.’ par les deux phrases suivantes :

(205) a. 他走進來。 *tā zǒu jìnlái.* <il-marcher-entrer.>
b. 他走進去。 *tā zǒu jìnqù.* <il-marcher-entrer.>

Dans le premier cas (206a), le complément complexe de direction, *jìnlái* 進來, indique que le mouvement du verbe *zǒu* 走 se fait vers le locuteur. Autrement dit, le sujet *tā* 他 s’approche du locuteur (où, dans un texte littéraire, vers le narrateur) qui se trouve à

l'intérieur de la pièce dans laquelle le sujet va entrer. Dans le second cas (206b), au contraire, le complément complexe de direction, *jìnqù* 進去, indique que le sujet *tā* 他, s'éloigne du locuteur qui se trouve à l'extérieur de la pièce dans laquelle le sujet va entrer.

7.1.3.3 Verbe–Complément d'aboutissement

Les compléments d'aboutissement sont formés de syntagmes prépositionnels. Ces syntagmes sont introduits par des prépositions telles que 到 *dào* 'jusqu'à', 向 *xiàng* 'vers', 在 *zài* 'à' ou 往 *wǎng* 'vers' suivies d'un complément d'objet, qui peut être un nom, un nom de temps, ou un nom de lieu. La structure Verbe-Complément d'aboutissement met l'accent sur un temps ou un lieu. Exemples :

(206) 跑向花園 *pǎo xiàng huāyuán* <courir-P-jardin> 'courir vers le jardin'
開往高雄 *kāi wǎng Gāoxióng* <conduire-P-Kaohsiung> 'partir vers Kaohsiung'
看到清晨 *kàn dào qīngchén* <regarder-P-petit matin> 'lire jusqu'au petit matin'
放在桌子上 *fàng zài zhuōzi shàng* <déposer-P-table-dessus> 'déposer sur la table'

7.1.3.4 Verbe–Complément potentiel

Le complément de potentialité exprime les possibilités de résultat, qu'elles soient affirmatives ou négatives. Ce complément de potentialité est formé de 得 *de* ou de 不 *bù* suivi par un adjectif ou un verbe. 得 *de* exprime la potentialité affirmative ; 不 *bù* la potentialité négative. Exemples :

(207) a. 講得清 *jiǎng de qīng* <parler-De-claire> 'arriver à parler des choses clairement'
b. 講不清 *jiǎng bù qīng* <parler-Bù-clair> 'ne pas réussir à parler des choses clairement'

7.1.3.5 Verbe–Complément introduit par le subordinateur postverbal 得 *de*

Le complément introduit par le subordinateur postverbal 得 *de*, peut être une unité lexicale simple ou un syntagme, voire une proposition complète. Exemples :

(208) 排得整齊 *pái de zhěngqí* <ranger-De-(bien) aligné> 'ranger en ordre'
煩惱得睡不著覺 *fánnǎo de shuì bù zháo jiào*
<s'ennuyer-De-dormir-Bù-atteindre-période de sommeil>
's'ennuyer au point de ne pas arriver à dormir'

7.1.3.6 Adjectif–Complément de degré

La structure Adjectif-Complément de degré présente le niveau de réalisation d'une situation ou d'un mouvement. L'adjectif est suivi d'un complément de degré qui est généralement un adverbe, par exemple, 極 *jí* 'extrêmement', 多 *duō* 'excessivement' ou 透 *tòu* 'complètement'. Éventuellement, la particule modale finale 了 *le* est ajoutée à la fin des syntagmes de forme Adjectif-Complément. L'ensemble, adverbe et particule finale, vise à compléter la description donnée par l'adjectif.

(209) 棒極了 *bàng jí le* <excellent-extrêmement-Le> 'excellentissime'
舒服多了 *shūfú duō le* <confortable-excessivement-Le>
'excessivement confortable'
刺激透了 *cìjī tòu le* <excitant-complètement-Le> 'tout à fait excitant'

7.1.4 Modifieur-Tête

Dans les syntagmes de type Modifieur-Tête, le modifieur détermine ou qualifie la tête qu'il précède, celle-ci étant constituée d'un nom ou d'un syntagme nominal. Le modifieur peut être un adjectif, un verbe, un autre nom, un adjectif à valeur descriptive, un syntagme de nature verbale ou un syntagme de type Sujet-Prédicat.

(210) Modifieur – Tête : A – N
聰明人 *cōngmíng rén* <intelligent-personne> 'personne intelligente'

(211) Modifieur – Tête : V – N
討論議題 *tǎolùn yìtí* <discuter-thème> 'thème de discussion'

(212) Modifieur – Tête : N – N
大理石桌子 *dàlǐshí zhuōzi* <marbre-table> 'table en marbre'

(213) Modifieur – Tête : Z – Syntagme nominal
糊裡糊塗一個人 *húlǐhútú yī gè rén*
<étourdi-un-Q-personne> '(une) personne étourdie'

(214) Modifieur – Tête : Syntagme de nature verbale (N V) – N
a. 香煙走私集團 *xiāngyān zǒusī jítuán* <tabac-faire de la contrebande-groupe>
'groupe de contrebande de tabac'
b. 資源回收場 *zīyuán huíshōu chǎng* <ressource-recycler-place>
'local de recyclage des déchets'

(215) Modifieur – Tête : Modifieur constitué d'un Sujet-Prédicat – Tête constituée d'un syntagme nominal
他們上課那教室。 *tāmen shàngkè nà jiàoshì.* <ils-suivre des cours-ce-salle.>
'Cette salle-là où ils suivent des cours.'

7.1.4.1 Utilisation du subordinateur nominal 的 *de*[58]

7.1.4.1.1 Présence du subordinateur nominal 的 de

Il existe trois cas dans lesquelles le subordinateur nominal 的 *de* est obligatoire.

1) Quand le modifieur est un nom, un nom de lieu, un nom de temps ou un pronom et qu'il exprime l'appartenance, le subordinateur nominal 的 *de* permet de cerner les limites de cette appartenance au nom-tête.

(216) 朋友的小孩 *péngyǒu de xiǎohái* <ami-De-enfant>
'enfant d'un ami'
高雄的景色 *Gāoxióng de jǐngsè* <Kaohsiung-De-paysage>
'paysage de Kaohsiung'
今天的午餐 *jīntiān de wǔcān* <aujourd'hui-De-déjeuner>
'déjeuner d'aujourd'hui'
她的項鍊 *tā de xiàngliàn* <elle-De-collier>
'son collier'

2) Lorsque le modifieur est formé d'un syntagme de forme Numéral-Quantifieur, le modifieur de cette structure représente une qualification quantitative portée sur la tête nominale. Par ailleurs, on constate une hiérarchie sémantique dans ce syntagme de Numéral-Quantifieur. Le subordinateur nominal 的 *de* joue ce rôle comme on le voit dans l'exemple ci-dessous :

(217) a. 一顆五公斤的西瓜 *yī kē wǔ gōngjīn de xīguā*
<un-Q-cinq-kilo-De-pastèque> 'une pastèque de cinq kilos'
b. *一顆五公斤西瓜 *yī kē wǔ gōngjīn xīguā*

Dans l'exemple (218a), le modifieur est formé du syntagme de structure Numéral-Quantifieur *yī kē wǔ gōngjīn* 一顆五公斤. *wǔ gōngjīn* 五公斤 décrit le poids de cette pastèque, tandis que *yī kē* 一顆 présente la totalité de cette pastèque. Ainsi, *yī kē* 一顆 est, sémantiquement en ordre hiérarchique, supérieur à *wǔ gōngjīn* 五公斤. Dans ce cas-là, il est obligatoire de faire précéder la tête nominale du subordinateur nominal 的 *de*.

[58] *cf.* Zhu Dexi, 1982 [2004 : 140-149].

3) Lorsqu'un syntagme de nature prédicative occupe la fonction de modifieur, le subordinateur nominal 的 *de* est obligatoire et précède la tête. Exemples :

(218) a. 喝酒的杯子 *hējiǔ de bēizi* <boire du vin-De-verre> 'verre à vin'
b. 請來的客人 *qǐnglái de kèrén* <inviter-venir-De-hôte> 'hôte invité'

7.1.4.1.2 Absence du subordinateur nominal 的 de

On ne doit pas utiliser le subordinateur nominal 的 *de* pour former des syntagmes de structure Modifieur-Tête :

1) lorsque le modifieur comporte une expression de quantité sur la tête :

(219) 一顆奇異果 *yī kē qíyìguǒ* <un-Q-kiwi> 'un kiwi'

2) lorsque le modifieur peut être une des catégories que la tête possède :

(220) 薪資問題 *xīnzī wèntí* <salaire-problème> 'problème de salaire'

3) lorsque le modifieur se présente comme un des types particuliers que peut prendre la tête :

(221) 燕麥粥 *yànmài zhōu* <bouillie claire-avoine> 'bouillie claire d'avoine'

7.1.4.1.3 Emploi optionnel du subordinateur nominal 的 de

L'utilisation du subordinateur nominal 的 *de* peut être optionnelle, dans deux cas :

1) Quand sa présence ou son absence ne change pas la signification d'un syntagme :

(222) a. 大理石傢俱 *dàlǐshí jiājù* <marbre-meuble> 'meuble en marbre'
b. 大理石的傢俱 *dàlǐshí de jiājù* <marbre-De-meuble> 'meuble en marbre'
(223) a. 漂亮衣服 *piàoliàng yīfu / fú* <joli-vêtement> 'joli vêtement'
b. 漂亮的衣服 *piàoliàng de yīfu / fú* <joli-De-vêtement> 'joli vêtement'

2) En revanche, il est des cas où sa présence ou son absence risque de modifier le sens d'un syntagme :

(224) a. 孩子脾氣 *háizi píqì* <enfant-caractère> 'enfantillage'
b. 孩子的脾氣 *háizi de píqì* <enfant-De-caractère> 'caractères de l'enfant'
(225) a.狐狸尾巴 *húlí wěiba* <renard-queue> 'ruse'
b. 狐狸的尾巴 *húlí de wěiba* <renard-De-queue> 'queue de renard'

7.1.4.2 Modifieurs complexes suivis d'une seule Tête

Il est possible d'employer plusieurs modifieurs pour qualifier ou déterminer une seule tête nominale. Dans quel ordre présenter ces modifieurs ? C'est ce que nous allons voir en analysant les syntagmes suivants qui signifient tous 'cette paire d'yeux brillants de la fille'.

(226) appartenance — désignation — quantité — description
女孩那一雙閃亮亮的眼睛
nǚhái nà yī shuāng shǎnliàngliàng de yǎnjing / jīng
<fille-ce-un-Q-brillant-De-œil>

Les quatre modifieurs servent à qualifier la tête nominale. Ils lui apportent chacun une qualification. En fonction de leur propriété sémantique, on peut les décrire ainsi :

● Appartenance : Le nom 女孩 *nǚhái* 'fille' indique que les yeux appartiennent à la fille.
● Désignation : Le démonstratif 那 *nà* 'ce ~ là' désigne précisément les yeux de cette fille.
● Quantité : Le syntagme composé d'un numéral et d'un quantifieur 一雙 *yī shuāng* <un-Q> 'une paire' précise le nombre d'yeux concernés.
● Description : L'adjectif à valeur descriptive 閃亮亮 *shǎnliàngliàng* 'brillant' décrit ces yeux.

Notons, cependant, que ces quatre modifieurs peuvent se présenter dans des ordres différents :

(227) appartenance — désignation — description — quantité
女孩那閃亮亮的一雙眼睛
nǚhái nà shǎnliàngliàng de yī shuāng yǎnjing / jīng
<fille-ce-brillant-De-un-Q-œil>

(228) appartenance — description — désignation — quantité
女孩閃亮亮的那一雙眼睛
nǚhái shǎnliàngliàng de nà yī shuāng yǎnjing / jīng
<fille-brillant-De-ce-un-Q-œil>

7.1.4.3 Compositions complexes de Modifieur-Tête

Il existe aussi des compositions complexes dans lesquelles un ou plusieurs modifieurs peuvent qualifier une ou plusieurs têtes. Zhu Dexi [1982] relevait dans son *Cours de grammaire* cinq combinaisons fréquentes [2004 : 150] :

1) A 的 N_1 + N_2 + N_3 + ... + N_n

(229) 蔚藍的天空和海洋
wèilán de tiānkōng hé / hàn hǎiyáng
<bleu-De-ciel-et-mer>
'ciel et mer bleus'

2) A_1 + A_2 + A_3 + ... + A_n 的 N

(230) 健康、快樂、活潑的孩子
jiànkāng, kuàilè, huópō de háizi
<santé, joyeux, dynamique-De-enfant>
'enfant en bonne santé, joyeux et dynamique'

3) A_1 的 + A_2 的 + A_3 的 + ... + A_n 的 N

(231) 健康的、快樂的、活潑的孩子
jiànkāng de, kuàilè de, huópō de háizi
<santé-De, joyeux-De, dynamique-De-enfant>
'enfant en bonne santé, joyeux et dynamique'

4) A_1 + A_2 + A_3 + ... + A_n 的 N_1 + N_2 + N_3 + ... + N_n

(232) 親切、和藹的父親和母親
qīnqiè, hé'ǎi de fùqīn hé / hàn mǔqīn
<aimable, gentil-De-père-et-mère>
'père et mère aimables et gentils'

5) A_1 的 + A_2 的 + A_3 的 + ... + A_n 的 N_1 + N_2 + N_3 + ... + N_n

(233) 健康的、快樂的、活潑的父親、母親和孩子
jiànkāng de, kuàilè de, huópō de fùqīn, mǔqīn hé / hàn háizi
<santé-De, joyeux-De, dynamique-De-père, mère-et-enfant>
'père, mère et enfant en bonne santé, joyeux et dynamiques'

7.1.5 Circonstant-Verbe / Circonstant-Adjectif[59]

La structure Circonstant-Verbe / Circonstant-Adjectif sert à décrire des faits ou à exposer des situations. Le verbe ou l'adjectif révèle l'état ou la nature de la situation. Le circonstant, suivi du verbe ou de l'adjectif, précise ou qualifie la description proposée par le verbe ou l'adjectif. Dans les syntagmes formés selon la structure Circonstant-Verbe ou Circonstant-Adjectif, les circonstants peuvent être classés en différents types selon leur signification. Exemples :

[59] Pour traduire le terme terminologique chinois 狀中 *zhuàng zhōng*, nous avons adapté le terme utilisé par Zhitang Yang-Drocourt [2007 : 322].

(234) Circonstant – Verbe

a. D – V
剛剛知道 *gānggāng zhīdào* <récemment-savoir>
‘être depuis peu au courant de’
b. A – V
安靜聆聽 *ānjìng língtīng* <silencieux-écouter>
‘écouter en silence’

c. Z – V
悠悠閒閒散步 *yōuyōu-xiánxián sànbù* <très tranquille-se promener>
‘se promener tranquillement’

d. O – V
哇哇大叫 *wāwā dàjiào* <onomatopée de sanglots-crier>
‘éclater en sanglots’

e. Syntagme prépositionnel – V

Sont exceptionnelles quelques prépositions, par exemple, 到 *dào*, 在 *zài*, 把 *bǎ*, 被 *bèi*, 給 *gěi* et 跟 *gēn* avec lesquelles on peut former des syntagmes prépositionnels. Ce type de syntagmes prépositionnels peut se présenter comme un circonstant et précéder un verbe. Exemples :

跟朋友說 *gēn péngyǒu shuō* <P-ami-parler>
‘parler à un ami’
把書歸位 *bǎ shū guīwèi* <P-livre-retourner-place>
‘remettre le livre à sa place’

(235) Circonstant – Adjectif

D – A
非常漂亮 *fēicháng piàoliàng* <vraiment-joli>
‘vraiment joli’
很高興 *hěn gāoxìng* <très-content>
‘très content’

Le subordinateur préverbal, 地 *de* peut être intégré entre le circonstant et le verbe ou l’adjectif. Exemples :

(236) Circonstant – 地 – Verbe

a. A – 地 – V
安靜地聆聽 *ānjìng de língtīng* <silencieux-De-écouter>
‘écouter en silence’

b. Z – 地 – V
悠悠閒閒地散步 *yōuyōu-xiánxián de sànbù*
<très tranquille-De-se promener>
‘se promener tranquillement’

c. O – 地– V
哇哇地大叫 *wāwā de dàjiào* <onomatopée de sanglots-De-crier>
‘éclater en sanglots’

(237) Circonstant – 地 – Adjectif

D – 地 – A
非常地漂亮 *fēicháng de piàoliàng* <vraiment-De-joli>
‘vraiment joli’

7.1.6 Coordination

La structure Coordination sert à énumérer des objets, des mouvements, des idées, etc. Dans le cas d’une énumération, cette structure peut être formée avec ou sans des conjonctions de coordination. Les syntagmes constitués de deux ou de plusieurs unités lexicales sont appelés syntagmes de type Coordination, bien que certains d’entre eux soient structurés sans l’aide des conjonctions [*cf.* Zhu Dexi, 1982 [2004 : 156-159]].

Les unités coordonnées peuvent être substantives :

(238) 蔬菜水果 *shūcài shuǐguǒ* <fruit-légume> ‘légume (et) fruit’
柴鹽油米醬醋茶 *chái yán yóu mǐ jiàng cù chá*
<combustible-sel-huile-riz-sauce-vinaigre-thé>
‘combustible (,) sel (,) huile (,) riz (,) sauce (,) vinaigre (et) thé’

Elles peuvent être prédicatives :

(239) 唱歌跳舞 *chànggē tiàowǔ* <chanter-danser> ‘chanter (et) danser’
滑雪游泳 *huáxuě yóuyǒng* <faire du ski-nager> ‘faire du ski (et) nager’

Lorsqu’on constitue un syntagme de forme Coordination, il est possible d’intégrer des ponctuations, des particules modales ou des conjonctions de coordination entre les substantifs, ou encore des adverbes entre les prédicatifs.

(240) a. 蔬菜、水果 *shūcài, shuǐguǒ* <légume, fruit> ‘légume, fruit’
b. 蔬菜啦、水果啦 *shūcài lā, shuǐguǒ lā* <légume-Lā, fruit-Lā> ‘légume, fruit’

(241) a. 柴、鹽、油、米、醬、醋、茶 *chái, yán, yóu, mǐ, jiàng, cù, chá*
<combustible, sel, huile, riz, sauce, vinaigre, thé>
‘combustible, sel, huile, riz, sauce, vinaigre, thé’

b. 柴啦、鹽啦、油啦、米啦、醬啦、醋啦、茶啦
chái lā, yán lā, yóu lā, mǐ lā, jiàng lā, cù lā, chá lā
<combustible-Lā, sel-Lā, huile-Lā, riz-Lā, sauce-Lā, vinaigre-Lā, thé-Lā>
'combustible, sel, huile, riz, sauce, vinaigre, thé'

(242) a. 又唱歌又跳舞 *yòu chànggē yòu tiàowǔ* <aussi-chanter-aussi-danser>
'chanter et aussi danser'
b. 唱歌並且跳舞 *chànggē bìngqiě tiàowǔ* <chanter-de plus-danser>
'chanter et danser'

(243) a. 滑雪、游泳 *huáxuě, yóuyǒng* <faire du ski, nager> 'faire du ski, nager'
b. 滑雪啦、游泳啦 *huáxuě lā, yóuyǒng lā* <faire du ski-Lā, nager-Lā> 'faire du ski, nager'

Les syntagmes conjonctifs servent à énumérer des éléments. Ceux-ci peuvent être des objets optionnels ou obligatoires selon la conjonction de coordination qui les associe. Exemples :

(244) 蔬菜與水果 *shūcài yǔ shuǐguǒ* <légume-et-fruit> 'légume et fruit'

Les objets énumérés étaient reliés par la conjonction 與 *yǔ* 'et', aucun des deux ne peut être omis.

(245) 蔬菜或水果 *shūcài huò shuǐguǒ* <fruit-ou-légume> 'légume ou fruit'

Le syntagme est formé à l'aide de la conjonction 或 *huò* 'ou', il propose un choix entre les deux noms cités. Ces deux derniers représentent des objets optionnels.

7.2 Description des structures locales

7.2.1 Grammaire de la composition numérique

Il existe quatre types de nombres numéraux en chinois [*cf.* 5.6.3]. Ils se composent, par exemple, de la manière suivante :

三百 *sānbǎi* <trois-cent> 'trois cents',
五千 *wǔqiān* <cinq-mille> 'cinq mille' ou
六千七百八十九 *liùqiān qībǎi bāshíjiǔ* <six-mille-sept-cent-quatre-vingt-neuf>
'six mille sept cent quatre-vingt neuf', etc.

Mais il est difficile, dans un dictionnaire, de décrire de façon exhaustive les compositions numériques. La grammaire ***ChineseNumerals*** décrit les combinaisons des numéraux cardinaux avec des numéraux (cardinaux) positionnels :

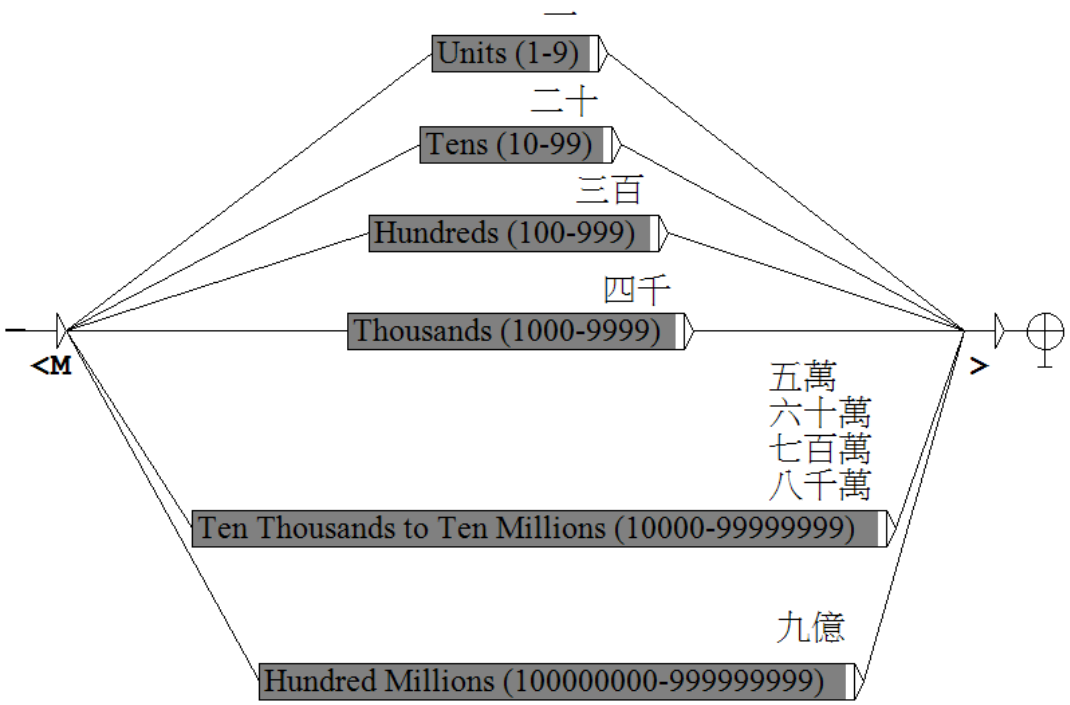


Figure 33 : Grammar ***ChineseNumerals***

Cette grammaire ***ChineseNumerals*** prend en compte les compositions allant de 一 *yī* 'un' à 九億九千九百九十九萬九千九百九十九 *jiǔyì jiǔqiān jiǔbǎi jiǔshíjiǔ wàn jiǔqiān jiǔbǎi jiǔshíjiǔ* 'neuf cent quatre-vingt dix-neuf millions neuf cent quatre-vingt dix-neuf mille neuf cent quatre-vingt dix-neuf'. Le graphe principal de la Figure 33 regroupe six sous-graphes. Par exemple, le sous-graphe ***Units (1-9)*** qui décrit les numéraux de 一 *yī* 'un' à 九 *jiǔ* 'neuf' est représenté comme suit :

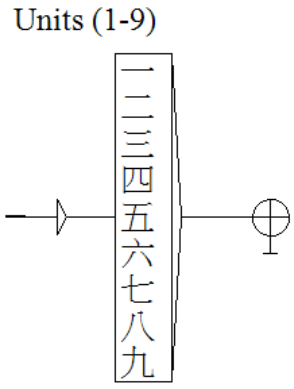


Figure 34 : Sous-graphe ***Units (1-9)***

Le numéral 'deux' ayant deux graphies, 二 *èr* et 兩 *liǎng*, qui diffèrent par l'usage, nous avons dû les distinguer lors du développement de la grammaire ***ChineseNumerals***. Nous

avons accordé une description particulière à chaque graphie, comme si ces deux graphies étaient deux composants morphologiques uniques. Une des distinctions se trouve dans le sous-graphe ***Hundreds (100-999)*** :

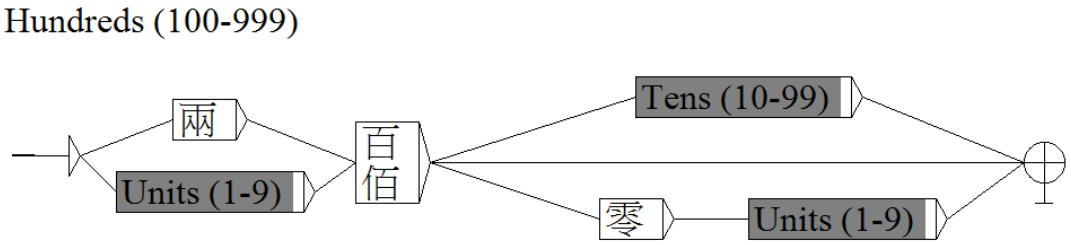


Figure 35 : Sous-graphe ***Hundreds (100-999)***

Ce dispositif permet, d'une part, de reconnaître les deux formes graphiques qui représentent le nombre numéral 'deux' ; d'autre part, de décrire précisément les compositions numériques liées à ces deux formes graphiques. Ce traitement permet de reconnaître les compositions numériques telles que

二十
兩百
二百[60]

En même temps qu'il élimine des compositions inacceptables comme :

*兩十
*三百零兩[61]

Par ailleurs, il est possible de combiner deux numéraux cardinaux suivis d'un numéral (cardinal) positionnel. Exemples :

五六百
七八千[62]

[60] 二十 *èrshí* <deux-dix> 'vingt'
兩百 *liǎngbǎi* <deux-cent> 'deux cents'
二百 *èrbǎi* <deux-cent> 'deux cents'

[61] *兩十 *liǎngshí* <deux-dix> '*liǎng* (deux) dix'
*三百零兩 *sānbǎi líng liǎng* <trois-cent-zéro-deux> 'trois cent *liǎng* (deux)'

[62] 五六百 *wǔ-liù bǎi* <cinq-six-cent> 'cinq ou six cents'
七八千 *qī-bā qiān* <sept-huit-mille> 'sept ou huit mille'

Cette composition numérique est développée de la façon suivante dans la grammaire ***CardPositNums*** :

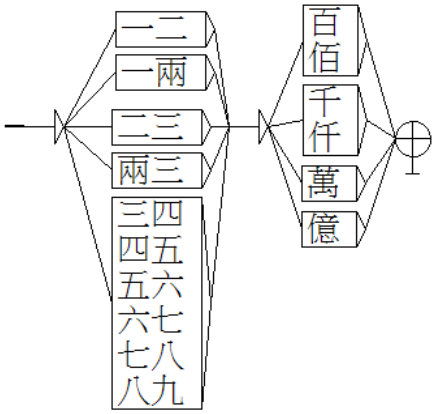


Figure 36 : Grammaire ***CardPositNums***

7.2.2 Grammaire des expressions temporelles

Pour formaliser les expressions de temps — durée, fréquence ou horaires —, nous avons élaboré la grammaire ***TimeExpression***. Un de ses trois sous-graphes ***During*** analyse les expressions qui renvoient à des durées[63] :

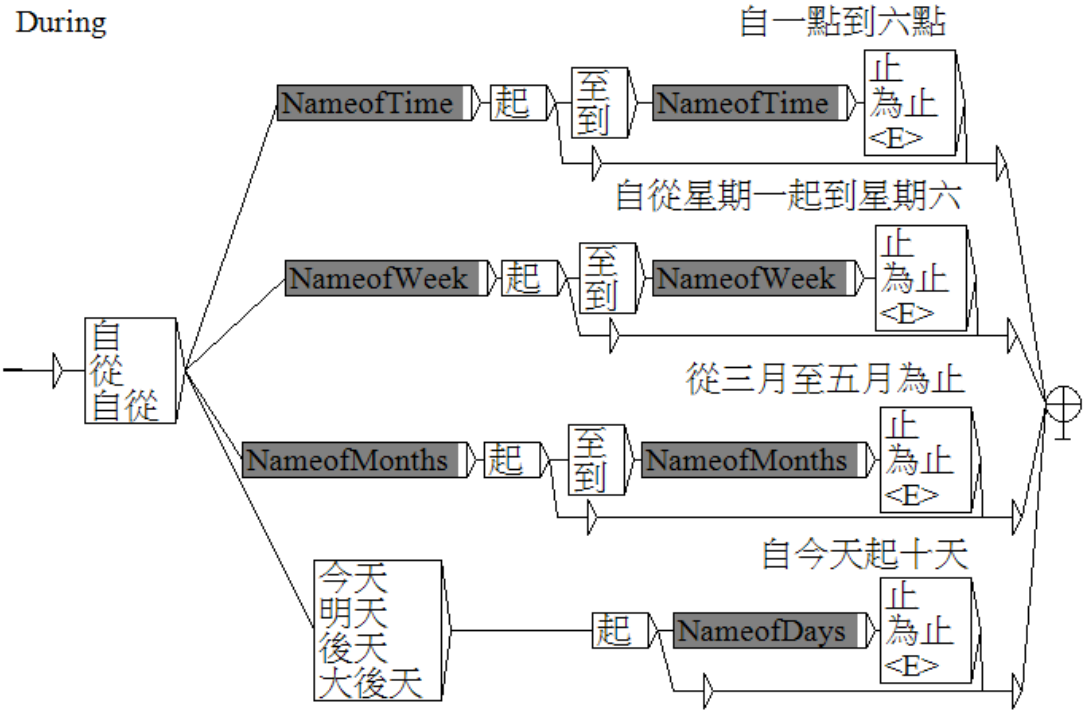


Figure 37 : Sous-graphe ***During***

[63] Le symbole spécifique <E> représente la séquence de caractères vide, c'est-à-dire, un élément neutre de la concaténation. Il sert à noter un élément optionnel ou un élément élidé.

Ce sous-graphe ***During*** représente des expressions de durée telles que :

從一點到六點
自從星期一起到星期六
從三月至五月為止
自今天起十天[64]

La grammaire ***TimeExpression*** permet aussi d'écarter des expressions incorrectes comme par exemple :

*二月三十日
*十一月三十一日[65]

7.2.3 Grammaire des appellatifs personnels

Nous avons construit une grammaire ***HumanTitles*** qui permet de reconnaître les différents types d'appellatifs personnels. Cette grammaire contient vingt-quatre sous-graphes imbriqués :

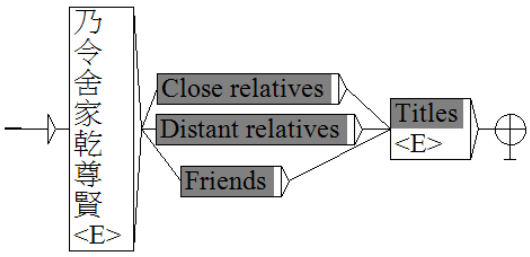


Figure 38 : Grammaire ***HumanTitles***

Les appellatifs sont nombreux et différents selon qu'ils se réfèrent au côté paternel ou au côté maternel de la famille. Il existe principalement cinq générations hiérarchiques du côté paternel. Le sous-graphe ***Paternal relatives*** précise les appellatifs adéquats pour ces cinq générations. Les appellatifs à l'intérieur de chacune de ces générations sont en outre présentés dans un sous-graphe :

[64] 從一點到六點 *cóng yī diǎn dào liù diǎn* <P-un-heure-P-six-heure> 'd'une heure à six heures'
自從星期一起到星期六 *zìcóng xīngqíyī qǐ dào xīngqíliù* <P-lundi-à partir de-P-samedi>
'du lundi au samedi'
從三月至五月為止 *cóng sānyuè zhì wǔyuè wéizhǐ* <P-trois-mois-P-cinq-mois-se faire arrêter>
'de mars jusqu'à mai'
自今天起十天 *zì jīntiān qǐ shí tiān* <P-aujourd'hui-à partir de-dix-jour> 'dix jours à partir d'aujourd'hui'

[65] *二月三十日 *èryuè sānshí rì* <deux-mois-trente-jour> 'le trente février'
*十一月三十一日 *shíyīyuè sānshíyī rì* <onze-mois-trente-un-jour> 'le trente et un novembre'

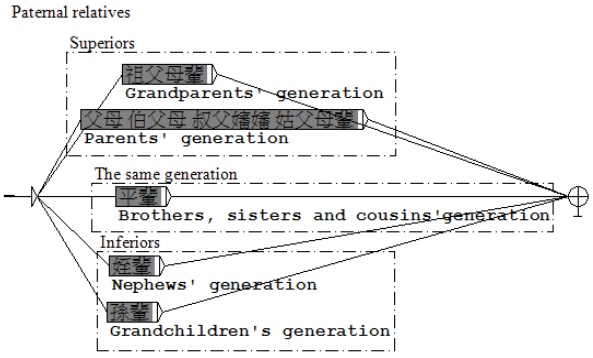


Figure 39 : Sous-graphe ***Paternal relatives***

Le sous-graphe *父母 伯父母 叔父嬸嬸 姑父母輩* ***(Parents' generation)*** contient les appellatifs employés pour nommer les membres de la famille, appartenant à la même génération que le père :

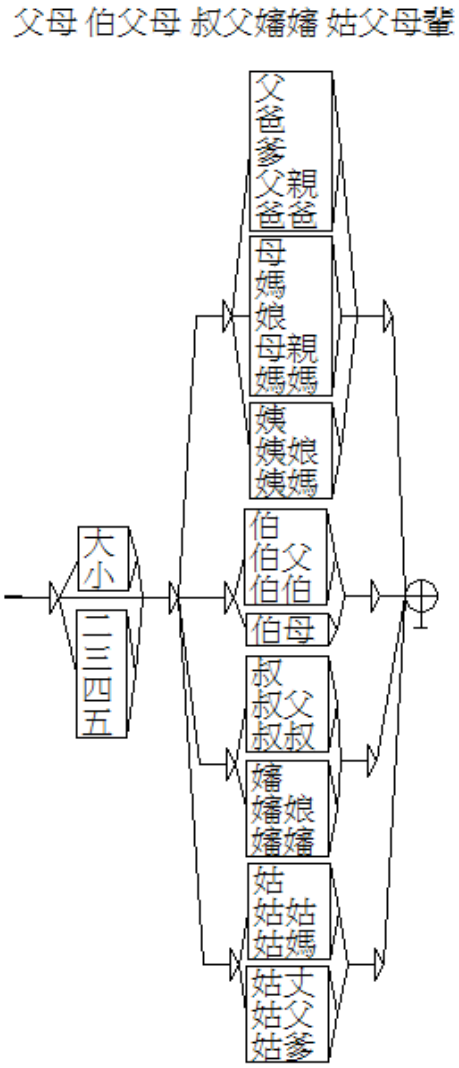


Figure 40 : Sous-graphe *父母 伯父母 叔父嬸嬸 姑父母輩* ***(Parents' generation)***

Le sous-graphe ***Maternal relatives*** présente les appellatifs utilisés pour nommer les membres de la famille du côté maternel :

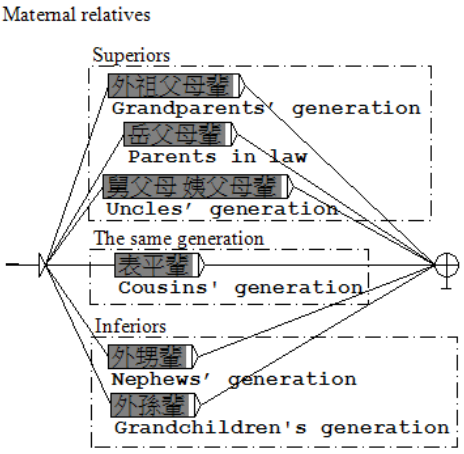


Figure 41 : Sous-graphe ***Maternal relatives***

Ce sous-graphe ***Maternal relatives*** regroupe les appellatifs hiérarchiques sur six générations. Les appellatifs correspondant aux beaux-parents sont définis ainsi :

岳父母輩

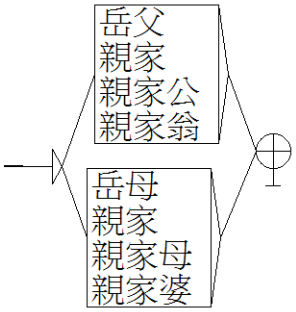


Figure 42 : Sous-graphe ***岳父母輩***

Certains titres ou certains appellatifs peuvent se combiner, par exemple,

總統先生
校長夫人
部長大人[66]

[66] 總統先生 *zǒngtǒng xiānshēng* <président-monsieur> 'Monsieur le Président de l'État'
校長夫人 *xiàozhǎng fūrén* <directeur d'école-madame> 'Madame la Directrice d'école'
部長大人 *bùzhǎng dàrén* <ministre-monsieur> 'Monsieur le Ministre'

La grammaire ***DoubleTitle*** décrit leur composition :

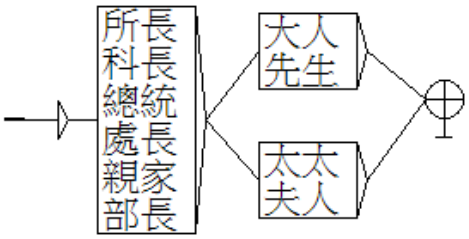


Figure 43 : Grammaire ***DoubleTitle***

7.3 Description des groupes nominaux

En chinois moderne, les groupes nominaux noyaux peuvent être classés en cinq types [*cf.* Huang Changning et Zhao Jun, 1998 : 40-46 et Zhan Weidong, 2000 : 39-58]. Le tableau ci-dessous présente, avec des exemples, les différentes structures qu'ils peuvent adopter :

Types de groupes nominaux	Structures de composition	Exemples
ProperName	小<M>	小三 *Xiǎo Sān* <H-San> 'Xiao San'
	老<NPL>	老陳 *Lǎo Chén* <H-Chen> 'Lao Chen'
Apposition	<NP><NP>	招弟他們 *Zhāodì tāmen* <Zhaodi-ils> 'les Zhaodi, ils'
	<NP><MP>	她們兩人 *tāmen liǎngrén* <elles-deux personnes> 'Toutes les deux'
ModifierHead	<QP><NP>	一張桌子 *yī zhāng zhuōzi* <un-Q-table> 'une table'
	<AP><的><NP>	美麗的衣服 *měilì de yīfu / fú* <beau-De-vêtement> 'beau vêtement'
Addition	<NP><式>	小家庭式的生活 *xiǎo jiātíngshì de shēnghuó* <petit-famille-SK-De-vie> 'mode de vie d'une petite famille'
	<NP><、><NP><C><NP><們>	先生、太太與小姐們 *xiānshēng, tàitài yǔ xiǎojiěmen* <monsieur, madame-et-mademoiselle-K> 'Messieurs, Mesdames et Mesdemoiselles'
Coordination	<NP><NP>	蔬菜水果 *shūcài shuǐguǒ* <fruit-légume> 'fruit (et) légume'
	<NP><C><NP>	鄉村與都市 *xiāngcūn yǔ dūshì* <campagne-et-ville> 'campagne et ville'

Nous allons présenter les grammaires qui formalisent ces cinq types de groupes nominaux.

7.3.1 NP_ProperName

7.3.1.1 Description de la grammaire

La grammaire ***NP_ProperName*** décrit la structure des groupes nominaux incluant des noms de famille, des prénoms, des surnoms et leurs compositions possibles [*cf*. Figure 44].

Cette grammaire permet de reconnaître les appellations composées des préfixes 老 *lǎo* ou 小 *xiǎo*, suivis d'un chiffre et éventuellement du suffixe 兒 *r*, comme par exemple :

老二
小三
老五
小六兒
老七兒
小八兒[67]

L'appellation 老大 *Lǎo Dà* <H-grand> 'premier' est aussi décrite dans cette grammaire. Signalons que l'expression incorrecte : *老一 *Lǎo Yī* <H-un> est exclue par cette grammaire.

[67] 老二 *Lǎo Èr* <H-deux> 'Lao Er : deuxième'
小三 *Xiǎo Sān* <H-trois> 'Xiao San : troisième'
老五 *Lǎo Wǔ* <H-cinq> 'Lao Wu : cinquième'
小六兒 *Xiǎo Liùr* <H-six-K> 'Xiao Liur : sixième'
老七兒 *Lǎo Qīr* <H-sept-K> 'Lao Qir : septième'
小八兒 *Xiǎo Bār* <H-huit-K> 'Xiao Bar : huitième'

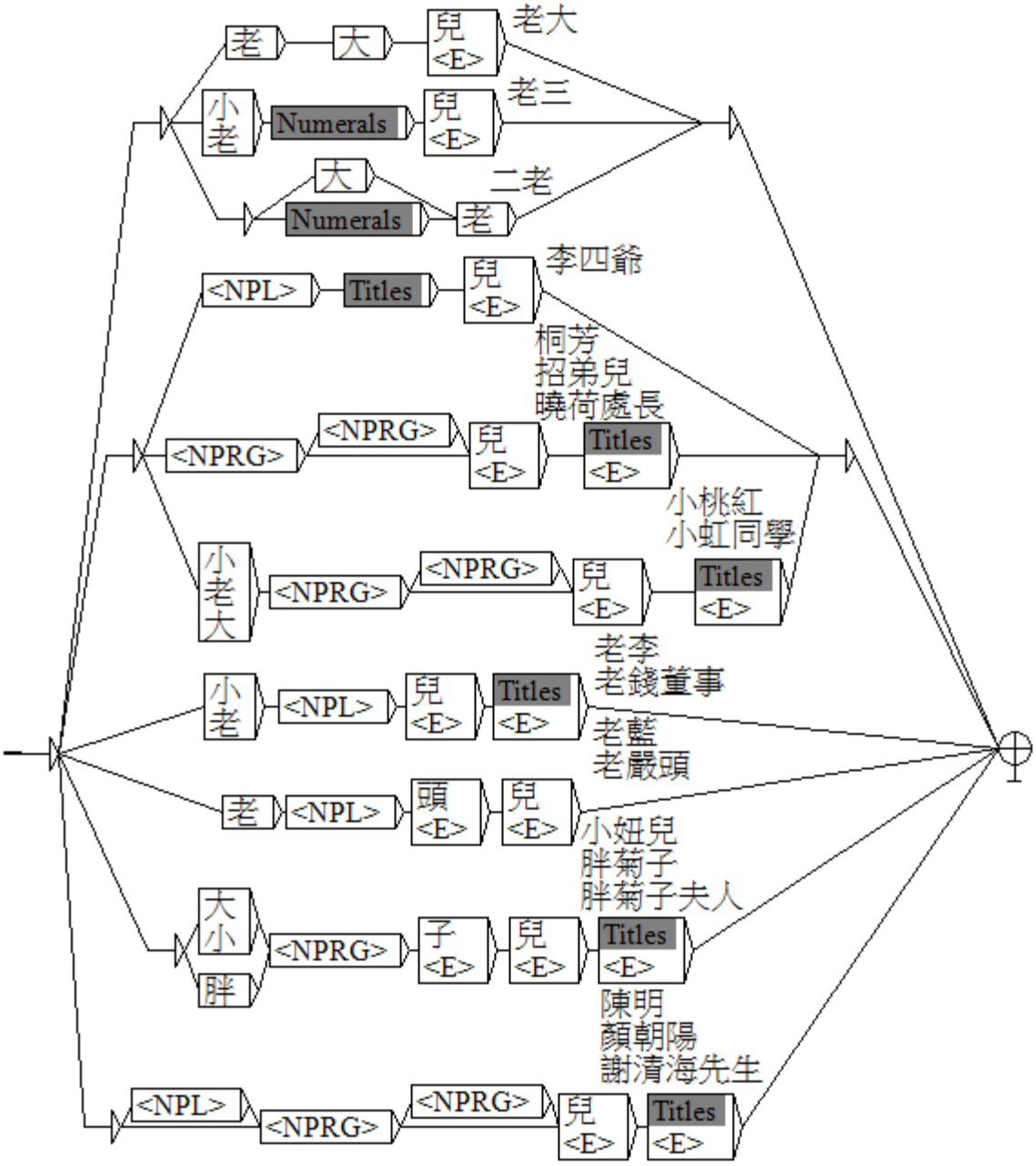


Figure 44 : Grammaire ***NP_ProperName***

Arbitrairement, un prénom chinois se compose d'un ou de deux caractères. Pour reconnaître les prénoms dans les données, nous avons associé l'étiquette <NPRG> aux caractères les plus fréquemment utilisés dans la formation des prénoms chinois [*cf.* Li Jianhua et Wang Xiaolong, 2000 : 46-49]. Dans le dictionnaire ***ChDic***, ces caractères sont suivis de l'étiquette mentionnée ci-dessus et présentés ainsi :

芳,NPRG
順,NPRG

燕,NPRG[68]

Lors du développement de la grammaire ***NP_ProperName***, les composants de prénoms sont limités à deux caractères. La grammaire décrit cette structuration des prénoms chinois, qui peuvent être suivis ou non du suffixe 兒 *r*. Exemples :

招順
桃花兒[69]

Les prénoms peuvent également être suivis de titres ou d'appellations. Cette structuration est aussi décrite. Exemples :

清河先生
桂櫻小姐[70]

Les noms de famille chinois, comme les prénoms français, sont bien déterminés. Nous avons donc construit le dictionnaire ***DicChSurn*** dans lequel tout nom de famille se voit attribué l'étiquette <NPL>, par exemple,

丁,NPL
上,NPL
上官,NPL[71]

L'étiquette <NPL> est réutilisée dans la grammaire ***NP_ProperName*** pour décrire des structures syntagmatiques incluant des noms de famille, par exemple :

● un nom de famille précédé du préfixe 老 *lǎo* : 老陳 *Lǎo Chén* 'Lao Chen', ou

● la précédente structure suivie d'un titre ou du suffixe 頭 *tou* :

老陳總經理 *Lǎo Chén zǒngjīnglǐ* <H-Chen-directeur général> 'Directeur général Lao Chen' ou
老陳頭 *Lǎo Chén tou* <H-Chen-K> 'Lao Chen Tou'.

[68] 芳,NPRG *fāng*
順,NPRG *shùn*
燕,NPRG *yàn*
[69] 招順 *Zhāoshùn* 'Zhaoshun'
桃花兒 *Táohuār* 'Taohuar'
[70] 清河先生 *Qīnghé xiānshēng* <Qinghe-monsieur> 'Monsieur Qinghe'
桂櫻小姐 *Guìyīng xiǎojiě* <Guiying-mademoiselle> 'Mademoiselle Guiying'
[71] 丁,NPL *Dīng*
上,NPL *Shàng*
上官,NPL *Shàngguān*

Par ailleurs, cette grammaire ***NP_ProperName*** permet de préciser les surnoms composés à l'aide du préfixe 小 *xiăo*, du morphème 大 *dà* avec le sens de 'grand', de 胖 *pàng* avec le sens de 'gros' ou de suffixes 子 *zi* ou 兒 *r*. Exemples :

小妞
大毛
胖桃子
小狗子
小順子兒
胖桃兒[72]

Cette grammaire reconnaît les noms propres formés d'un prénom, d'un nom de famille ou d'un titre. Elle formalise ainsi les noms propres tels que :

陳明
顏朝陽
謝清海先生[73]

7.3.1.2 Évaluation de la grammaire

En appliquant la grammaire ***NP_ProperName***, ***NooJ*** peut reconnaître ce type de groupes nominaux dans le roman de Lao She *Quatre générations sous un même toit*. Nous avons obtenu le résultat suivant :

[72] 小妞 *Xiăo Niū* 'Xiao Niu'
大毛 *Dà Máo* 'Da Mao'
胖桃子 *Pàng Táozi* 'Pang Taozi'
小狗子 *Xiăo Gŏuzi* 'Xiao Gouzi'
小順子兒 *Xiăo Shùnzir* 'Xiao Shunzir'
胖桃兒 *Pàng Táor* 'Pang Taor'

[73] 陳明 *Chén Míng* 'Chen Ming'
顏朝陽 *Yán Zhāoyáng* 'Yan Zhaoyang'
謝清海先生 *Xiè Qīnghăi xiānshēng* <Xie-Qinghai-monsieur> 'Monsieur Xie Qinghai'

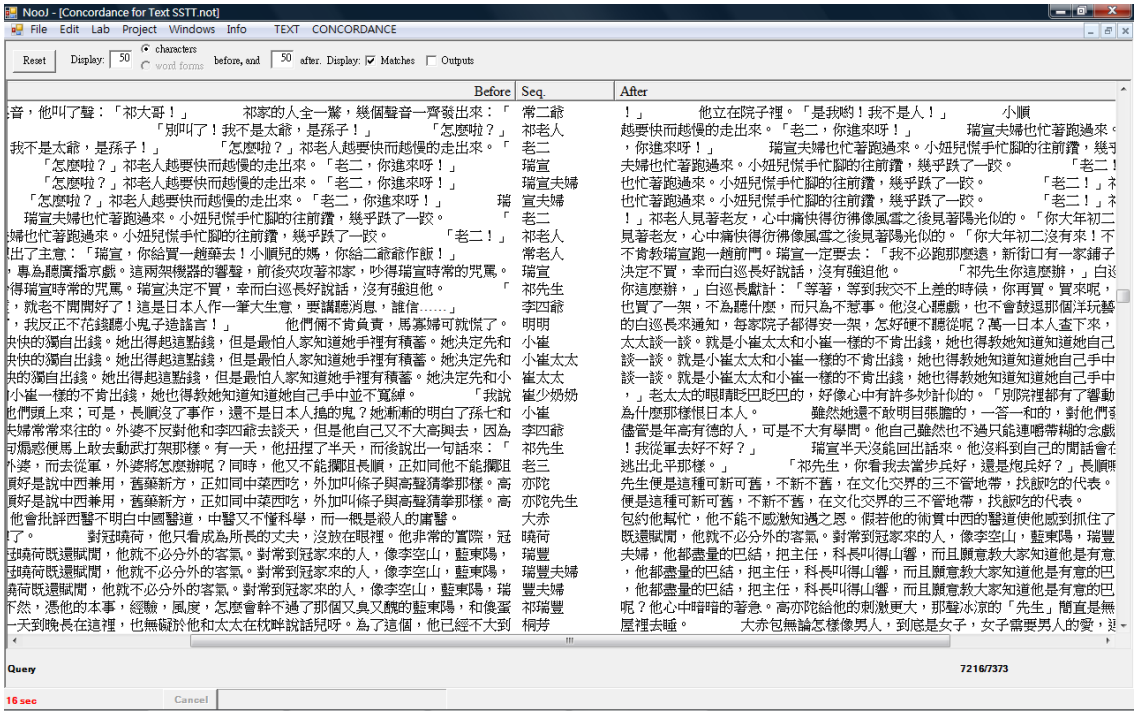


Figure 45 : Concordance des groupes nominaux de type ***ProperName***

Certaines entrées de la concordance ne sont pas correctes selon le contexte. On nomme « bruits » les entrées de la concordance non pertinentes. Ces bruits sont dus à des ambiguïtés lexicales. Exemples :

*明明
*大赤
*宣夫婦[74]

Dans le dictionnaire ***ChDic*** les caractères 明 *míng*, 大 *dà*, 赤 *chì* et 宣 *xuān*, sont étiquetés <NPRG>. Ils sont donc reconnus comme des prénoms. De plus, lorsqu'ils s'assemblent, leurs combinaisons sont également traitées comme des prénoms dissyllabiques et elles sont alors reconnues comme telles et deviennent des entrées de la concordance. Néanmoins, ces combinaisons ne sont pas toujours pertinentes en regard du contexte. Par exemple, 明明 *míngmíng* peut être un prénom chinois, mais selon le contexte, c'est ici un adverbe. On constate aussi que les deux séquences 大赤 *Dà Chì* et 宣夫婦 *Xuān fūfù* quoique sémantiquement correctes, ne sont pas pertinentes dans le texte de Lao She.

[74] *明明 *Míngmíng* 'Mingming'
*大赤 *Dà Chì* 'Da Chi'
*宣夫婦 *Xuān fūfù* <Xuan-couple> 'le couple Xuan'

Il existe des groupes nominaux de ce type qui ne sont pas reconnus par cette grammaire ***NP_ProperName***, car leur structure n'y est pas décrite. On appelle « silences » les unités pertinentes mais non affichées lors de la recherche automatique des textes. Exemples :

默翁
錢默翁
日本老爺[75]

En conclusion, en appliquant cette grammaire, nous avons obtenu 62 % de précision et 81 % de rappel.

7.3.2 NP_Apposition

7.3.2.1 Description de la grammaire

Les groupes nominaux de type ***Apposition*** sont composés d'unités lexicales appositives. Ils sont décrits dans la grammaire suivante [*cf.* Figure 46].

Les groupes nominaux de type ***Apposition*** sont composés essentiellement de noms humains, de pronoms personnels, de noms de famille ou de titres. Ces composants sont identifiés par différentes étiquettes <S>, <NPL> et <H+Hum>. Signalons au passage que, composés, eux aussi, par apposition, tous les noms de lieu sont codés par <S> dans le dictionnaire ***GeoDic***, par exemple,

中國,S
北京,S
日本,S
東京,S
法國,S
巴黎,S[76]

[75] 默翁 *Mò wēng* <Mo-monsieur> 'Monsieur Mo'
錢默翁 *Qián Mò wēng* <Qian-Mo-monsieur> 'Monsieur Qian Mo'
日本老爺 *Rìběn lǎoyé* <japonais-maître> 'Maître japonais'

[76] 中國,S *Zhōngguó* 'Chine'
北京,S *Běijing / jīng* 'Pékin'
日本,S *Rìběn* 'Japon'
東京,S *Dōngjīng* 'Tokyo'
法國,S *Fǎ / Fà guó* 'France'
巴黎,S *Bālí* 'Paris'

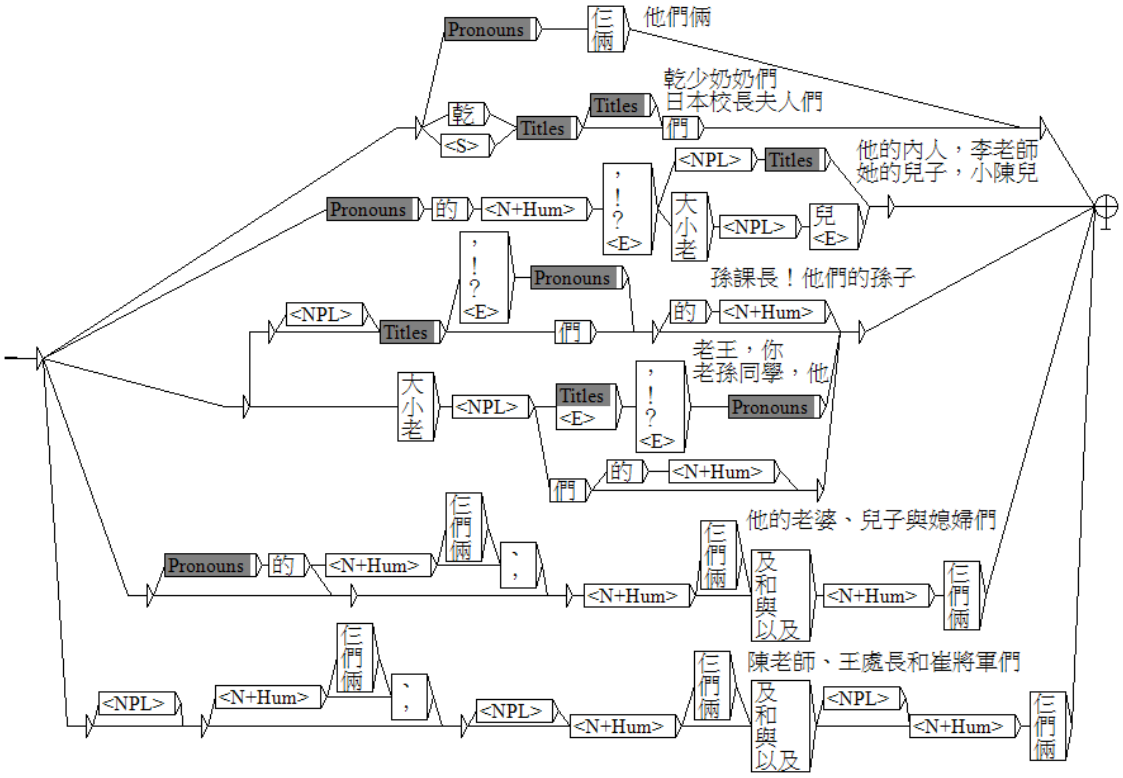


Figure 46 : Grammaire ***NP_Apposition***

L'étiquette <H+Hum> est associée aux noms humains, dans le dictionnaire ***ChDic***, exemples :

學生,N+Hum
老師,N+Hum
研究員,N+Hum
舞蹈家,N+Hum
鋼琴家,N+Hum[77]

Les entrées de ces trois types servent à former des groupes nominaux dits appositifs et peuvent éventuellement inclure des ponctuations. Exemples :

乾少奶奶們[78]
日本校長夫人們
他的內人，李老師

[77] 學生,N+Hum *xuéshēng* <étudier-homme instruit> 'étudiant'
老師,N+Hum *lǎoshī* <H-maître> 'professeur'
研究員,N+Hum *yánjiū / jiù yuán* <effectuer des recherches-K> 'chercheur'
舞蹈家,N+Hum *wǔdǎojiā* <danse-K> 'danseur'
鋼琴家,N+Hum *gāngqínjiā* <piano-K> 'pianiste'

[78] 乾 *gān* est l'appellation donnée à une personne adoptive. Mais l'adoption n'est pas toujours officielle. 少奶奶 *shàonǎinai* est un titre réservé aux épouses des fils de famille.

她的兒子，小陳兒
孫課長！他們的孫子
老王，你
陳老師、王處長和崔將軍們[79]

7.3.2.2 Évaluation de la grammaire

En appliquant la grammaire de la Figure 46, ***NooJ*** a pu reconnaître des groupes nominaux de type ***Apposition*** dans le roman de Lao She *Quatre générations sous un même toit*. Le résultat se présente comme suit :

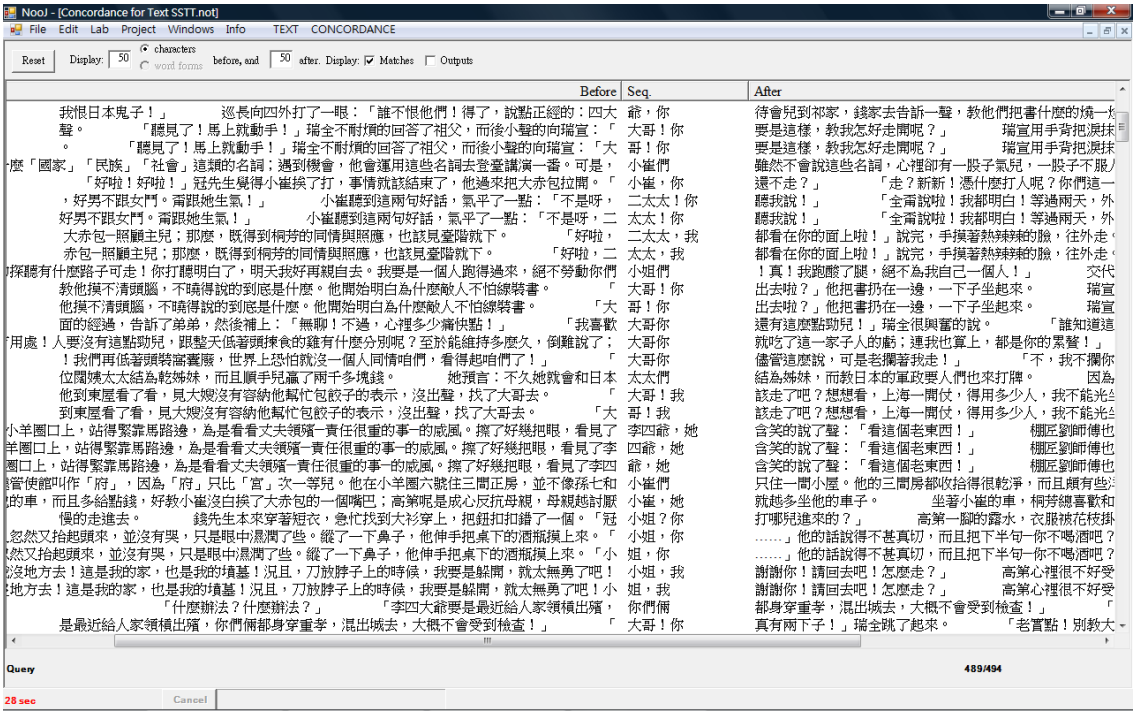


Figure 47 : Concordance des groupes nominaux de type ***Apposition***

Lors de la recherche des groupes nominaux de type ***Apposition***, on constate que les bruits les plus notables sont des séquences non pertinentes au contexte. En d'autres termes,

[79] 乾少奶奶們 *gān shàonǎinaimen* <adoptif-jeune-madame-K> 'les jeunes dames adoptives'
日本校長夫人們 *Rìběn xiàozhǎng fūrénmen* <Japon-directeur d'école-madame-K>
'Mesdames les Directrices d'école japonaises'
他的內人，李老師 *tā de nèirén, Lǐ lǎoshī* <il-De-femme, Li-professeur> 'sa femme, Professeur Li'
她的兒子，小陳兒 *tā de érzi, Xiǎo Chénr* <elle-De-fils, H-Chen-K> 'son fils, Xiao Chenr'
孫課長！他們的孫子 *Sūn kèzhǎng ! tāmen de sūnzi* <Sun-chef de service ! ils-De-petit enfant>
'Chef de service Sun ! Leur petit-fils'
老王，你 *Lǎo Wáng, nǐ* <H-Wang, toi> 'Lao Wang, toi'
陳老師、王處長和崔將軍們 *Chén lǎoshī, Wáng chùzhǎng hé / hàn Cuī jiāngjūnmen*
<Chen-professeur, Wang-chef de bureau-et-Cui-général-K>
'le Professeur Chen, le Chef de bureau Wang et le Général Cui'

ces séquences repérées sont grammaticalement correctes et respectent la description syntaxique. Pourtant, elles ne sont pas adaptées au contexte. Considérons, par exemple :

*哥！你
*爺，她
*小崔，她[80]

Selon le contexte, la séquence pertinente n'est pas 哥！你 *gē ! nǐ*, mais 大哥！你 *dàgē ! nǐ* <grand frère ! tu> 'grand frère ! tu'. Quant aux deux autres séquences 爺，她 *yé, tā* et 小崔，她 *Xiǎo Cuī, tā*, elles ne sont pas sémantiquement correctes, puisque des pronoms personnels féminins y sont appliqués à des personnes masculines. Une ambiguïté sémantique se présente entre le sujet et le pronom personnel, si bien que les deux séquences sont traitées comme des bruits dans cette recherche de groupes nominaux.

Dans le roman de Lao She, il y a des groupes nominaux de type ***Apposition*** qui sont formés selon d'autres structures que celles décrites par la grammaire. Ces groupes nominaux ne sont naturellement pas reconnus. Citons en quelques uns :

咱們姓祁的人
兩位小姐，高第與招弟
這個宇宙的主宰—冠曉荷[81]

En utilisant cette grammaire, nous avons donc obtenu 63 % de précision et 78 % de rappel.

7.3.3 NP_ModifierHead

7.3.3.1 Description de la grammaire

Les groupes nominaux formés à partir de la structure Modifieur-Tête sont appelés groupes nominaux de type ***ModifierHead***. Leur structure syntagmatique est décrite dans la grammaire ***NP_ModifierHead*** :

[80] *哥！你 *gē ! nǐ* <(grand) frère ! tu> '(grand) frère ! tu'
*爺，她 *yé, tā* <monsieur, elle> 'monsieur, elle'
*小崔，她 *Xiǎo Cuī, tā* <H-Cui, elle> 'Xiao Cui, elle'

[81] 咱們姓祁的人 *zánmen xìng Qí de rén* <nous-nommer-Qi-De-homme> 'nous les Qi'
兩位小姐，高第與招弟 *liǎng wèi xiǎojiě, Gāodì yǔ Zhāodì* <deux-Q-demoiselle, Gaodi-et-Zhaodi> 'les deux demoiselles, Gaodi et Zhaodi'
這個宇宙的主宰—冠曉荷 *zhè gè yǔzhòu de zhǔzǎi — Guàn Xiǎohé*
<ce-Q-univers-gouverneur—Guan Xiaohe> 'le gouverneur de cet univers — Guan Xiaohe'

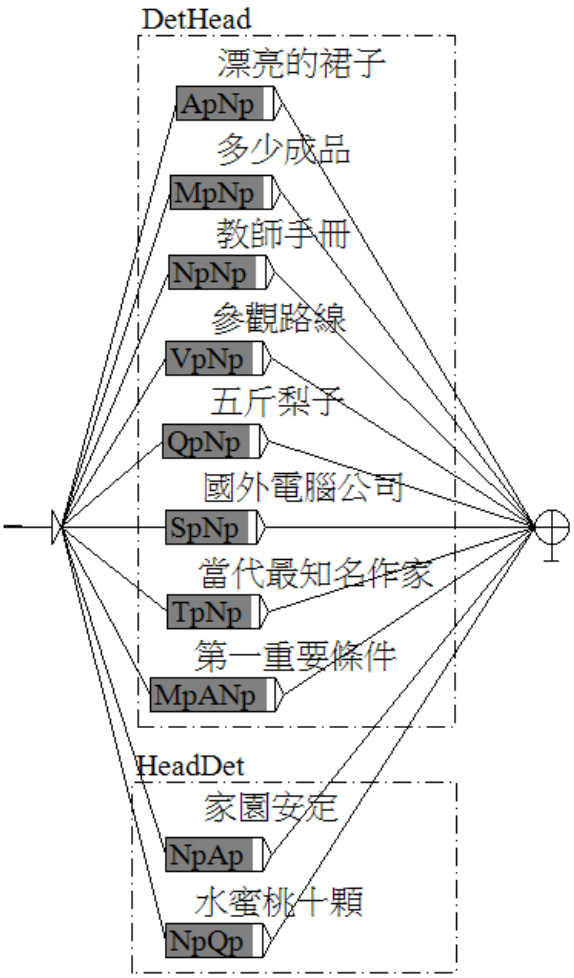


Figure 48 : Grammaire ***NP_ModifierHead***

Cette grammaire ***NP_ModifierHead*** formalise dix sous-catégories de groupes nominaux de type ***ModifierHead***, chacune étant décrite dans un sous-graphe. Les huit premiers présentent les groupes nominaux structurés selon le modèle **Modifieur** suivi de **Tête**, les deux derniers présentent, ceux qui sont construits sur le modèle **Tête** suivie de **Modifieur** [*cf.* Zhan Weidong, 2000 : 46-58].

Un des sous-graphes ***ApNp*** a été conçu pour formaliser les groupes nominaux dans lesquels les groupes adjectivaux constituent les modifieurs. Il se présente ainsi :

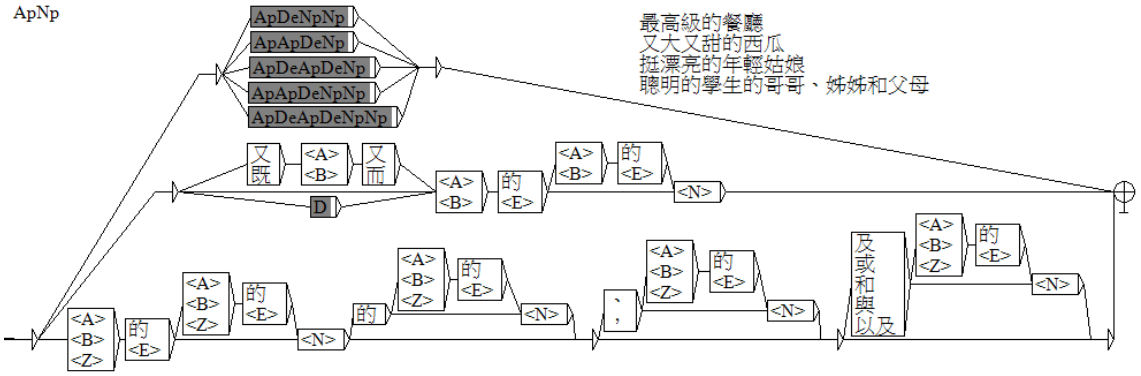


Figure 49 : Sous-graphe ***ApNp***

Il peut reconnaître des groupes nominaux tels que :

最高級的餐廳
又大又甜的西瓜
挺漂亮的年輕姑娘
聰明的學生的哥哥、姊姊和父母[82]

7.3.3.2 Évaluation de la grammaire

On peut utiliser la grammaire ***NP_ModifierHead*** pour reconnaître ce type de groupes nominaux dans le corpus. En appliquant cette grammaire à *Quatre générations sous un même toit*, nous avons obtenu le résultat suivant :

[82] 最高級的餐廳 *zuì gāojí de cāntīng* <le plus-meilleur-De-restaurant>
'restaurant le plus meilleur'
又大又甜的西瓜 *yòu dà yòu tián de xīguā* <aussi-grand-aussi-sucré-De-pastèque>
'grose pastèque sucrée'
挺漂亮的年輕姑娘 *tǐng piàoliàng de niánqīng gūniang* / *niáng* <très-beau-De-jeune-fille>
'jeune fille très belle'
聰明的學生的哥哥、姊姊和父母 *cōngmíng de xuésheng de gēgē, zǐzǐ / jiějiě hé / hàn fùmǔ*
<intelligent-De-étudiant-De-frère, sœur-et-parents>
'frères, sœurs et parents de l'étudiant intelligent'

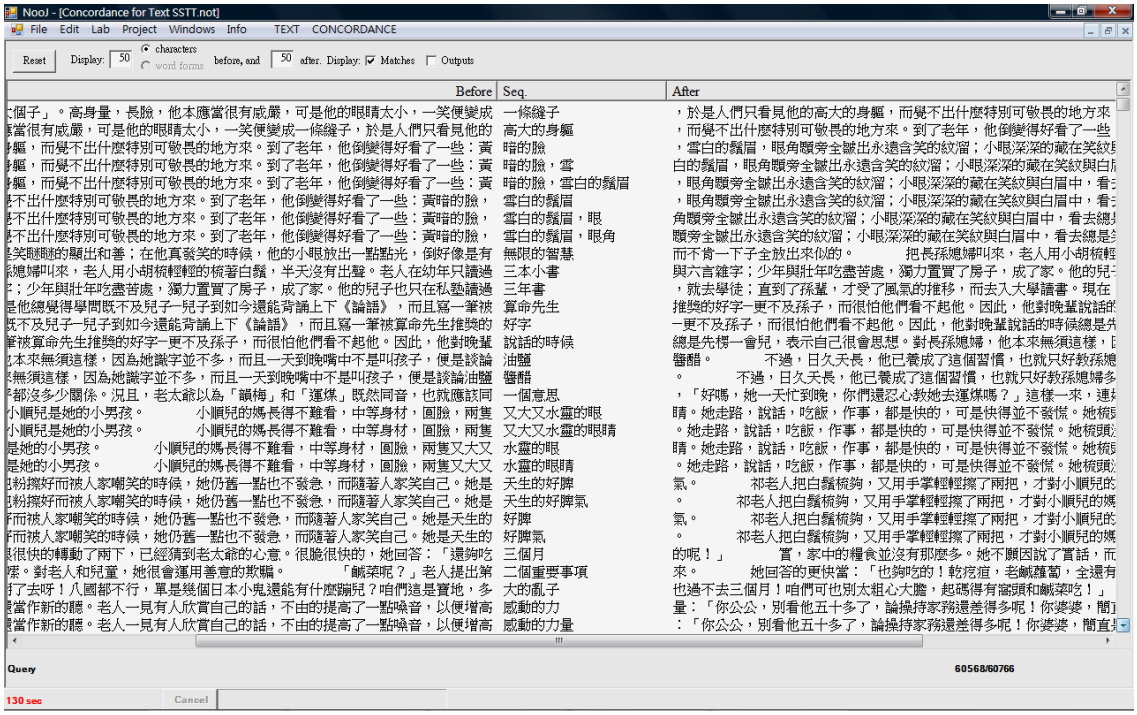


Figure 50 : Concordance des groupes nominaux de type ***ModifierHead***

En examinant le résultat de la recherche, on constate que certaines séquences ambigües apparaissent dans la concordance. Il est difficile de lever ce genre d'ambiguïtés, car on ne peut pas, lors de l'élaboration de la grammaire ***NP_ModifierHead***, décider d'une contrainte en ce qui concerne la longueur des unités lexicales. Mentionnons quelques exemples qui sont grammaticalement corrects, mais sémantiquement incorrects :

*暗的臉，雪
*又大有水靈的眼
*感動的力[83]

雪 *xuě* constitue un nom et il peut aussi être un composant morphologique de l'adjectif à valeur descriptive 雪白 *xuěbái* <neige-blanc>'blanc immaculé'. Ces deux unités lexicales sont donc conçues comme des entrées de dictionnaire. On associe à chacune une étiquette particulière : <N> pour 雪 *xuě* ; <Z> pour 雪白 *xuěbái*. De ce fait, ***NooJ*** peut reconnaître 雪 *xuě*, composant morphologique du 雪白 *xuěbái*, comme un nom avec lequel les groupes nominaux se forment.

[83] *暗的臉，雪 *àn de liǎn, xuě* <sombre-De-visage, neige> 'visage sombre, neige'
*又大又水靈的眼 *yòu dà yòu shuǐlíng de yǎn* <aussi-grand-aussi-vif-De-œil> 'grands yeux vifs'
*感動的力 *gǎndòng de lì* <émouvant-force> 'force émouvante'

Selon le contexte reconnu par la grammaire, le groupe nominal pertinent est 又大有水靈的眼睛 *yòu dà yòu shuǐlíng de yǎnjing / jīng*. Néanmoins, le deuxième exemple mentionné est sémantiquement tout aussi correct que celui-ci. Ces groupes nominaux apparaissent ensemble dans la concordance, car les deux unités lexicales 眼 *yǎn* et 眼睛 *yǎnjing / jīng* sont toutes deux des entrées nominales de dictionnaire et peuvent être reconnues par la grammaire ***NP_ModifierHead***.

Le troisième exemple s'explique de la même manière. Les deux noms 力 *lì* et 力量 *lìliàng* sont chacun des entrées de dictionnaire et signifient 'force'. Donc, ***NooJ*** peut les reconnaître quand on applique la grammaire ***NP_ModifierHead***, bien que selon le contexte l'unité pertinente soit 力量 *lìliàng*.

Néanmoins, les groupes nominaux de type ***ModifierHead*** ne sont pas tous reconnus par la grammaire, car leur structure est plus complexe que celle décrite dans la grammaire ***NP_ModifierHead***. Voici quelques exemples de groupes nominaux pertinents non reconnus :

年輕輕的公母倆
這早晚的年輕夫妻
西城護國寺附近的「小羊圈」[84]

En appliquant cette grammaire au roman de Lao She, nous avons donc obtenu 69 % de précision et 73 % de rappel dans la recherche de groupes nominaux de type ***ModifierHead***.

7.3.4 NP_Addition

7.3.4.1 Description de la grammaire

Les groupes nominaux de type ***Addition*** sont construits à l'aide du suffixe 們 *men*, du suffixe 性 *xìng*, du semi-suffixe 式 *shì*, des auxiliaires marqueurs de fin d'énumération comme 等 *děng* ou 等等 *děngděng*, etc. Exemples :

[84] 年輕輕的公母倆 *niánqīngqīng de gōngmǔliǎ* <jeune-De-homme-femme-tous deux> 'jeune couple'
這早晚的年輕夫妻 *zhè zǎowǎn de niánqīng fūqī* <ce-matin-soir-De-jeune-couple>
'le jeune couple actuel'
西城護國寺附近的「小羊圈」 *xīchéng hùguósì fùjìn de 「xiǎoyángjuàn」*
<ouest-ville-Huguo-monastère bouddhique-voisinage-De-« petit-enclos des moutons »>
'« petit espace » près du Huguo monastère bouddhique se situant à l'ouest de la ville'

魔鬼式的頭髮
宗教性的服務
對聯與題字等等
先生、太太與小姐們[85]

La grammaire ***NP_Addition*** a été conçue pour décrire de façon formelle quatre structures de groupes nominaux de type ***Addition***. Elle contient cinq sous-graphes :

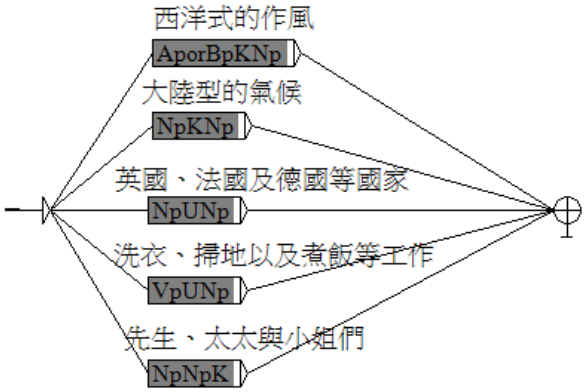


Figure 51 : Grammaire ***NP_Addition***

Un des quatre sous-graphes, ***NpUNp***, se présente comme suit :

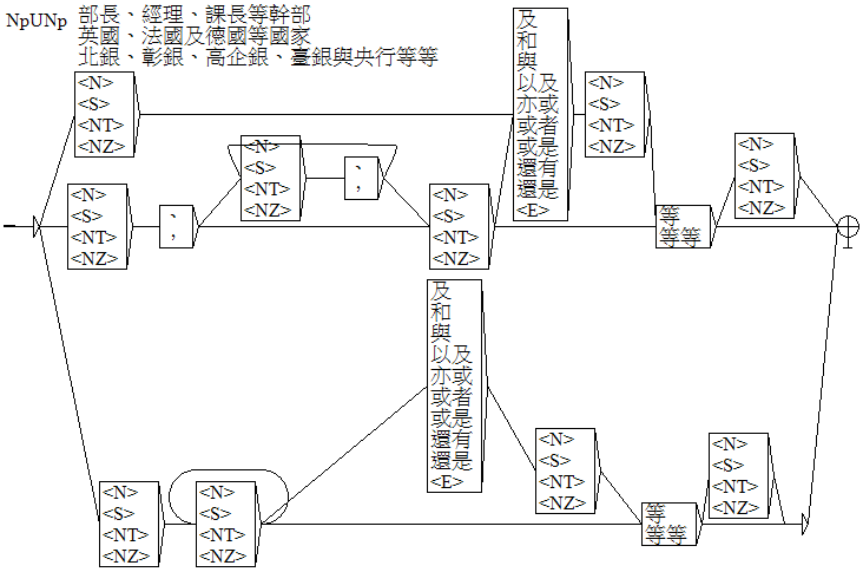


Figure 52 : Sous-graphe ***NpUNp***

[85] 魔鬼式的頭髮 *móguǐshì de tóufǎ* <démon-SK-De-cheveux> ‘les cheveux en forme de démon’
宗教性的服務 *zōngjiàoxìng de fúwù* <religion-K-De-service> ‘service religieux’
對聯與題字等等 *duìlián yǔ tízì děngděng* <sentence parallèle-et-inscription sur une œuvre d’art-etc.> ‘sentences parallèles, inscriptions sur une œuvre d’art, etc.’
先生、太太與小姐們 *xiānshēng, tàitài yǔ xiǎojiěmen* <monsieur, madame-et-mademoiselle-K> ‘Messieurs, Mesdames et Mesdemoiselles’

Il sert à préciser les énumérations composées de noms, de noms de lieu, de noms d'entreprises ou d'organismes et des noms propres. Exemples :

部長、經理、課長等幹部
英國、法國及德國等國家
北銀、彰銀、高企銀、臺銀與央行等等[86]

7.3.4.2 Évaluation de la grammaire

La grammaire ***NP_Addition*** permet de reconnaître les groupes nominaux de type ***Addition***. En l'appliquant au roman de Lao She, nous avons obtenu le résultat suivant :

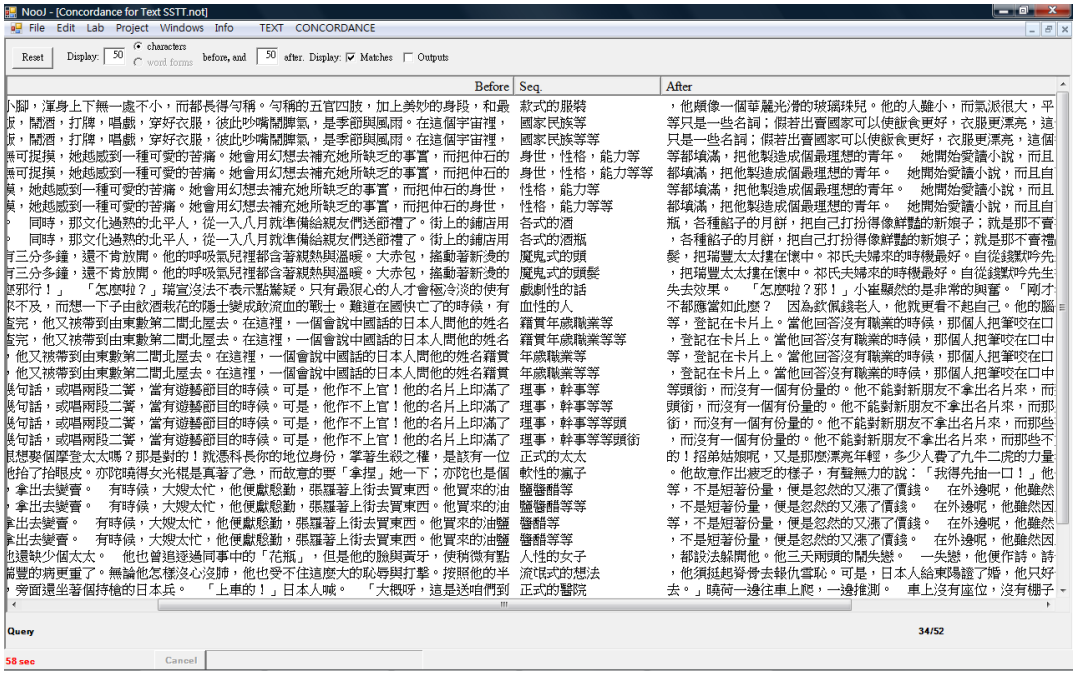


Figure 53 : Concordance des groupes nominaux de type ***Addition***

On constate toujours des ambiguïtés lexicales dans la concordance, ce qui produit certains bruits lors de la recherche automatique. Citons quelques exemples relevés, lors de la recherche des groupes nominaux dans le roman de Lao She :

[86] 部長、經理、課長等幹部 *bùzhǎng, jīnglǐ, kèzhǎng děng gànbù* <ministre, directeur, chef de servie-etc.-cadre> 'les cadres tels que ministres, directeurs, chefs de service, etc.'
英國、法國及德國等國家 *Yīngguó, Fǎ / Fà guó jí Déguó děng guójiā* <Angleterre, France-et-Allemagne-etc.-pays> 'les pays comme l'Angleterre, la France, l'Allemagne, etc.'
北銀、彰銀、高企銀、臺銀與央行等等 *běiyín, zhāngyín, gāoqìyín, táiyín yǔ yāngháng děngděng* <Taïpei-banque, Zhanghua-banque, Kaohsiung-affaire-banque, Taïwan-banque-et-nationale-banque-etc.> 'Banque de Taïpei, Banque de Zhanghua, Banque d'Affaires de Kaohsiung, Banque de Taïwan, Banque nationale, etc.'

*國家民族等
*魔鬼式的頭
*醬醋等[87]

Ces trois exemples sont sémantiquement corrects, et ils respectent bien la structure syntaxique décrite par la grammaire ***NP_Addition***. Pourtant, ils ne sont pas pertinents selon le contexte. Les syntagmes pertinents sont les suivants :

國家民族等等
魔鬼式的頭髮
油鹽醬醋等等[88]

Les ambiguïtés lexicales constatées dans la concordance sont dues aux faits suivants. On peut ajouter librement les deux unités lexicales 等 *děng* ou 等等 *děngděng* à la fin d'une énumération. Il n'y a aucune différence sémantique entre elles. Lors de la recherche automatique, un groupe nominal composé de l'unité lexicale 等等 *děngděng* peut correspondre à deux entrées de la concordance. C'est le cas de 國家民族等等 *guójiā mínzú děngděng*. Ce groupe nominal possède les deux entrées suivantes :

國家民族等
國家民族等等[89]

D'autre part, les deux unités lexicales 頭 *tóu* et 頭髮 *tóufǎ* sont des entrées nominales de dictionnaire. Nous avons dit que les composants de groupes nominaux de type ***Addition*** peuvent être assumés par les noms dans la grammaire de la Figure 51. Néanmoins, la longueur d'une unité lexicale ne peut pas être contrôlée par la grammaire ***NP_Addition***. Si les composants d'une séquence peuvent correspondre à deux ou plusieurs unités lexicales listées dans le dictionnaire, cette séquence aura deux entrées de la concordance. Donc, les deux entrées suivantes proviennent du même groupe nominal, mais c'est la deuxième qui se trouve dans le roman de Lao She :

[87] *國家民族等 *guójiā mínzú děng* <pays-peuple-etc.> 'le pays, les peuples, etc.'
*魔鬼式的頭 *móguǐshì de tóu* <démon-SK-De-tête> 'la tête en forme de démon'
*醬醋等 *jiàng cù děng* <sauce-vinaigre-etc.> 'des sauces, du vinaigre, etc.'

[88] 國家民族等等 *guójiā mínzú děngděng* <pays-peuple-etc.> 'le pays, les peuples, etc.'
魔鬼式的頭髮 *móguǐshì de tóufǎ* <démon-SK-De-cheveu> 'les cheveux en forme de démon'
油鹽醬醋等等 *yóu yán jiàng cù děngděng* <huile-sel-sauce-vinaigre-etc.> 'de l'huile, du sel, des sauces, du vinaigre, etc.'

[89] 國家民族等 *guójiā mínzú děng* <pays-peuple-etc.> 'le pays, les peuples, etc.'
國家民族等等 *guójiā mínzú děngděng* <pays-peuple-etc.> 'le pays, les peuples, etc.'

魔鬼式的頭
魔鬼式的頭髮[90]

Lorsque nous avons construit la grammaire ***NP_Addition***, nous n'avons pas limité le nombre d'éléments susceptibles d'être énumérés. Si l'énumération contient plusieurs éléments, elle aura plusieurs entrées. Chacune possède un nombre différent d'éléments énumérés. Ainsi en prenant en compte l'utilisation des unités lexicales 等 *děng* ou 等等 *děngděng*, le groupe nominal 油鹽醬醋等等 *yóu yán jiàng cù děngděng* peut répondre au moins aux quatre entrées suivantes :

醬醋等
醬醋等等
油鹽醬醋等
油鹽醬醋等等[91]

La grammaire ***NP_Addition*** ne reconnaît pas tous les groupes nominaux de type ***Addition***. Les groupes nominaux non reconnus sont formés de structures plus complexes ou plus longues. On ajoute des ponctuations ou utilise des plusieurs syntagmes pour constituer ce type de groupes nominaux. Mentionnons quelques exemples non reconnus par la grammaire de la Figure 51 :

中國式的「辨證法」
「大嫂」「媽媽」等應得的稱呼
小偷兒，私運煙土的，和嘎雜子們[92]

[90] 魔鬼式的頭 *móguǐshì de tóu* <démon-SK-De-tête> 'la tête en forme de démon'
魔鬼式的頭髮 *móguǐshì de tóufǎ* <démon-SK-De-cheveu> 'les cheveux en forme de démon'

[91] 醬醋等 *jiàng cù děng* <sauce-vinaigre-etc.> 'des sauces, du vinaigre, etc.'
醬醋等等 *jiàng cù děngděng* <sauce-vinaigre-etc.> 'des sauces, du vinaigre, etc.'
油鹽醬醋等 *yóu yán jiàng cù děng* <huile-sel-sauce-vinaigre-etc.> 'de l'huile, du sel, des sauces, du vinaigre, etc.'
油鹽醬醋等等 *yóu yán jiàng cù děngděng* <huile-sel-sauce-vinaigre-etc.> 'de l'huile, du sel, des sauces, du vinaigre, etc.'

[92] 中國式的「辯證法」 *Zhōngguóshì de 「biànzhèngfǎ」*
<Chinois-SK-De-« dialectique »>
'la « dialectique » à la chinoise'
「大嫂」「媽媽」等應得的稱呼 *「dàsǎo」「māma」děng yīngdé de chēnghū*
<« belle-sœur »-«maman»-etc.-mérité-De-appellation>
'les appellations méritées telles que « belle-sœur », «maman», etc.'
小偷兒，私運煙土的，和嘎雜子們 *xiǎotōur, sīyùn yāntǔ de, hé / hàn gāzázimen*
<voleur-K, faire de la contrebande-tabac-De, et-canaille-K>
'les voleurs, les contrebandiers en tabac et les canailles'

Dans la recherche de groupes nominaux de type ***Addition*** dans *Quatre générations sous un même toit*, nous avons donc obtenu 83 % de précision et 85 % de rappel.

7.3.5 NP_Coordination

7.3.5.1 Description de la grammaire

En chinois, il est possible de constituer des syntagmes de coordination, en s'appuyant ou non sur des conjonctions de coordination. Ainsi, des groupes nominaux peuvent être formés à l'aide ou sans l'aide de conjonctions de coordination. Pour décrire de façon formelle leurs structures, la grammaire ***Np_Coordination*** utilise la construction :

Figure 54 : Grammaire ***NP_Coordination***

Cette grammaire a été élaborée pour formaliser des groupes nominaux de type ***Coordination***, répartis en deux groupes : 1) les groupes nominaux n'utilisant pas de conjonctions de coordination [*cf.* sous-graphe ***NpNp***] ; 2) ceux utilisant des conjonctions [*cf.* sous-graphe ***NpCNp***]. Le sous-graphe ***NpCNp*** contient encore sept structures différentes. Une de ses structures nommée ***ApNpCApNp*** est représentée ci-dessous :

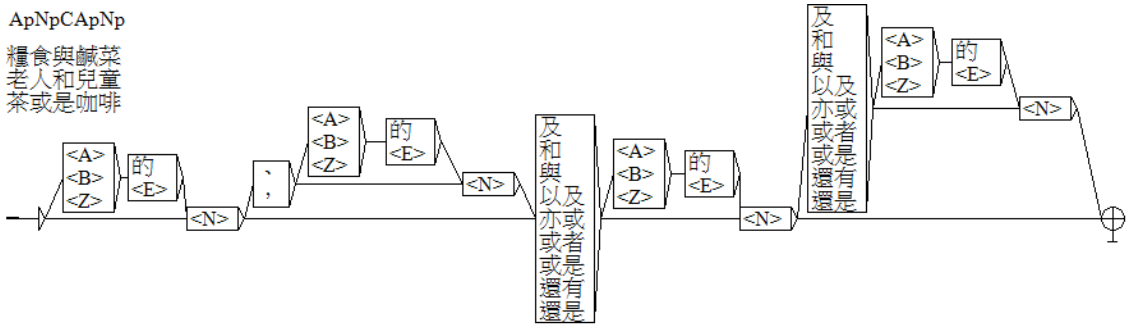


Figure 55 : Sous-groupe ***ApNpCApNp***

Lors de l'application de la grammaire ***NP_Coordination***, ***NooJ*** peut reconnaître des groupes nominaux basés sur les conjonctions. Mentionnons quelques exemples :

糧食與鹹菜
老人和兒童
茶或是咖啡[93]

7.3.5.2 Évaluation de la grammaire

Nous avons appliqué la grammaire ***NP_Coordination*** à *Quatre générations sous un même toit*, pour reconnaître les groupes nominaux formés par ***Coordination***. Le résultat de la recherche se présentait ainsi :

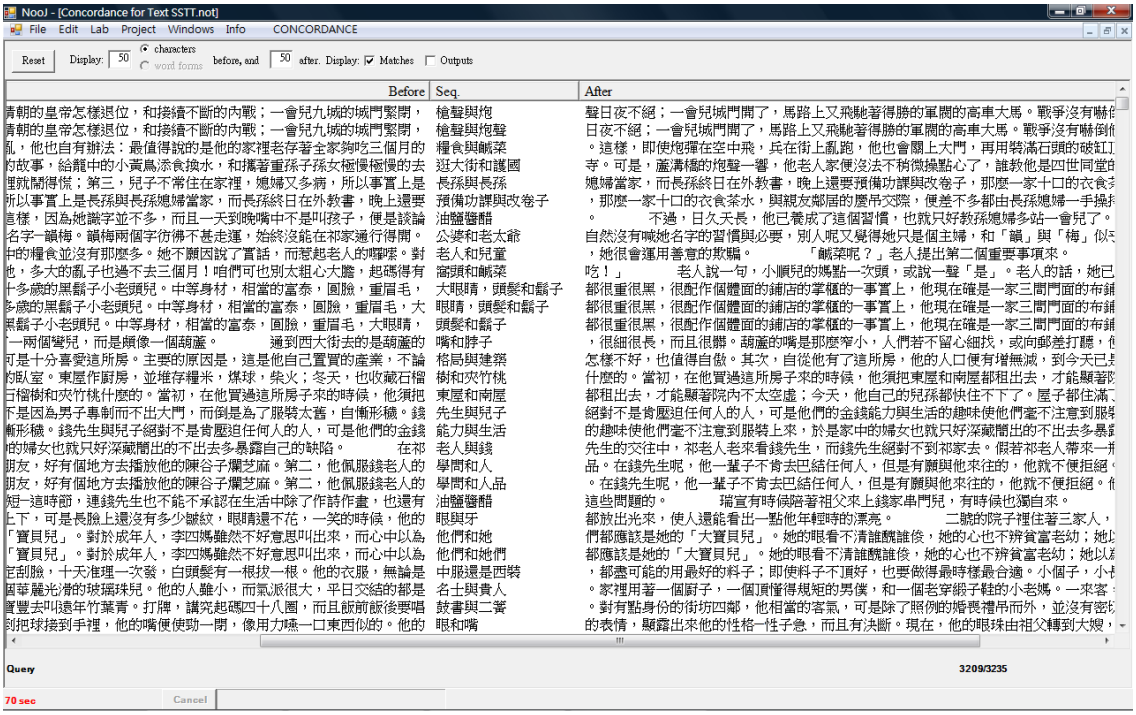


Figure 56 : Concordance de groupes nominaux de type ***Coordination***

Les exemples mentionnés ci-dessous sont pertinents selon le contexte du roman :

大眼睛，頭髮和鬍子
學問和人品
祁老人與錢先生[94]

[93] 糧食與鹹菜 *liángshí yǔ xiáncài* <nourriture-et-légume macéré dan la saumure>
'nourritures et légumes macérés dans la saumure'
老人和兒童 *lǎorén hé / hàn értóng* <vieillard-et-enfant>
'vieillards et enfants'
茶或是咖啡 *chá huòshì kāfēi* <thé-ou-café>
'thé ou café'

[94] 大眼睛，頭髮和鬍子 *dà yǎnjing / jīng, tóufǎ hé / hàn húzi* <grand-œil, cheveu-et-barbe>
'(ses) grands yeux, (ses) cheveux et (sa) barbe'
學問和人品 *xuéwèn hé / hàn rénpǐn* <connaissance-et-tenue morale>

Les composants de ces trois exemples peuvent être analysés selon différentes unités lexicales d'après leur longueur. Par ailleurs, dans un groupe nominal de type ***Coordination***, on peut ajouter ou non des adjectifs pour qualifier les noms. Ainsi, ces exemples peuvent correspondre éventuellement aux entrées de la concordance comme suit :

眼睛，頭髮和鬍子
學問和人
老人與錢[95]

Nous avons dit que l'adjectif est un élément potentiel dans la grammaire ***NP_Coordination***, ainsi que le groupe nominal 眼睛，頭髮和鬍子 *yǎnjing / jīng, tóufǎ hé / hàn húzi* est une des entrées de la concordance possibles pour analyser celui-ci 大眼睛，頭髮和鬍子 *dà yǎnjing / jīng, tóufǎ hé / hàn húzi*.

Les deux noms 人 *rén* et 人品 *rénpǐn* sont des entrées de dictionnaire. Puisque ces deux unités lexicales sont nominales, elles peuvent toutes deux être reconnues par la grammaire malgré leur différence de longueur. De ce fait, le groupe nominal 學問和人品 *xuéwèn hé / hàn rénpǐn* peut répondre aux entrées de la concordance suivantes, si l'on ne tient pas compte de leur sens :

學問和人
學問和人品

Le résultat s'explique par le fait que la signification contextuelle ne peut pas être décrite dans la grammaire, qui ne permet de reconnaître les groupes nominaux que d'après leur structure syntaxique.

Parfois, certaines séquences correspondent parfaitement à l'objet de la recherche, bien que leur sens ne soit pas tout à fait pertinent selon le contexte. C'est le cas du groupe nominal 祁老人與錢先生 *Qí lǎorén yǔ Qián xiānshēng*. Une partie de ce groupe nominal

'connaissances et tenue morale'
祁老人與錢先生 *Qí lǎorén yǔ Qián xiānshēng* <Qi-vieillard-et-Qian-monsieur>
'le vieillard Qi et Monsieur Qian'

[95] *眼睛，頭髮和鬍子 *yǎnjing / jīng, tóufǎ hé / hàn húzi* <œil, cheveu-et-barbe>
'(ses) yeux, (ses) cheveux et (sa) barbe'
*學問和人 *xuéwèn hé / hàn rén* <connaissance-et-homme>
'(les) connaissances et (les) êtres humains'
*老人與錢 *lǎorén yǔ qián* <vieillard-et-argent>
'(le) vieillard et (l')argent'

correspond bien à la structure de groupes nominaux de type ***Coordination***. Donc, elle apparaît dans la concordance et se présente comme suit :

老人與錢

On peut traiter cette séquence comme un groupe nominal de type ***Coordination***, à condition d'ignorer son contexte. C'est ainsi qu'elle est reconnue par la grammaire de la Figure 54.

Cette grammaire ne reconnaît pas tous les groupes nominaux de ce type, car les autres ont une structure plus complexe. Mentionnons ci-dessous quelques exemples non reconnus trouvés dans *Quatre générations sous un même toit* :

北海、東安市場和一什麼電影園
又是搓拳，又是磨掌的
一點黃油，咖啡，或真正的牛津橙子醬[96]

En utilisant cette grammaire, nous avons donc obtenu 84 % de précision et 79 % de rappel pour reconnaître les groupes nominaux de type ***Coordination*** dans le roman de Lao She.

7.4 Conclusion

Nous avons construit une série de grammaires à partir des six dictionnaires électroniques. Ces grammaires visent deux cibles :

- La reconnaissance de certains syntagmes locaux tels que les expressions temporelles, les expressions numériques, les appellations, etc.

  Les grammaires développées pour reconnaître des structures locales sont nommées grammaires locales. Elles servent à lever des ambiguïtés syntaxiques en décrivant les structures locales.

[96] 北海、東安市場和一什麼電影園 *Běihǎi, Dōngān shìchǎng hé / hàn — shénme diànyǐngyuán*
<Beihai, Dongan-marché-et — quelconque-cinéma>
'Beihai, le marché Dongan et un quelconque cinéma'
又是搓拳，又是磨掌的 *yòushì cuōquán, yòushì mózhǎng de*
<aussi-se frotter les poings, aussi-se frotter les paumes-De>
'se frotter les poings aussi bien que les paumes'
一點黃油，咖啡，或真正的牛津橙子醬 *yīdiǎn huángyóu, kāfēi, huò zhēnzhèng de Niújīn chéngzijiàng*
<un peu-beurre, café, ou-vrai-De-Oxford-orange-confiture>
'un peu de beurre, de café, ou des vraies confitures d'orange d'Oxford'

- La reconnaissance des groupes nominaux noyaux.

  Les grammaires syntaxiques servent à décrire de façon formelle les cinq types de groupes nominaux noyaux. Elles permettent de reconnaître des différents types de groupes nominaux dans le roman de Lao She.

  Comme les prénoms et les noms de famille peuvent s'écrire avec des caractères qui peuvent être des graphies de morphèmes, de mots ou de caractères non signifiants pour transposer les syllabes de mots étrangers, une ambiguïté lexicale se présente lors de l'analyse de *Quatre générations sous un même toit*. Certaines graphies sont analysées comme des prénoms ou des noms de famille, bien qu'elles soient, en réalité selon le contexte, des composants morphologiques. Or l'application des deux grammaires ***NP_Apposition*** et ***NP_PropreName*** ne procure que peu de précision, mais un bon rappel : la première atteint à 62 % de précision et 81 % de rappel ; et la deuxième 63 % de précision et 78 % de rappel.

  Par ailleurs, les groupes nominaux de type ***ModifierHead*** sont très nombreux et leurs structures sont aussi très variées. En appliquant la grammaire ***NP_ModifierHead***, on n'obtient que 69 % de précision et 73 % de rappel.

  Par contre, lors de l'application des grammaires ***NP_Addition*** et ***NP_Coordination***, ***NooJ*** montre qu'environ 80 % des groupes nominaux sont apparentés aux deux types que formalisent ces grammaires. On obtient 83 % de précision et 85 % de rappel pour la recherche des groupes nominaux de type ***Addition*** ; et 84 % de précision et 79 % de rappel pour la recherche des groupes nominaux de type ***Coordination***.

Chapitre 8

APPLICATION : UNE ETUDE SUR L'EVOLUTION DES THEMES DANS LES ŒUVRES LITTERAIRES

L'analyse thématique permet de représenter et de caractériser les informations contenues dans un corpus. Cette recherche orientée permet d'identifier et de différencier des thèmes, ainsi que de montrer leur évolution dans les données textuelles pour une période choisie. C'est là l'intérêt de l'analyse thématique, tant pour l'indexation que pour l'exploitation des connaissances textuelles. Nons allons étudier l'évolution des thèmes dans quelques textes littéraires en nous appliquant le module chinois développé dans ***NooJ***.

Nous expliquerons d'abord la façon dont sont choisis, dans les textes sélectionnés, les termes représentatifs, puis comment ils sont classés par thèmes. Ensuite, nous présenterons des études sur l'évolution chronologique des thèmes dans notre corpus limité à trente-neuf textes littéraires publiés de 1919 à 1993 [*cf.* Annexe 1], contenant donc des thèmes traités dans des ouvrages écrits au XX^{e} siècle. Cet espace temporel permet de retracer l'évolution des thèmes dans le cadre d'une période donnée. Enfin, nous montrerons, sous forme graphique, les résultats obtenus.

8.1 Identification thématique

Un thème est illustré par un ou plusieurs termes représentatifs. L'identification des thèmes doit se faire en deux phases :

1) Prétraitement des données ;

2) Extraction des connaissances.

Nous présenterons, ci-dessous, les analyses de ces deux phases qui permettent de reconnaître les termes représentatifs et les thèmes dans les trente-neuf textes littéraires sélectionnés.

8.1.1 Prétraitement des données

Nous avons considéré chaque texte comme une unité de base d'où sont extraites des unités d'information, c'est-à-dire, les unités linguistiques correspondant aux mots-clefs. Il s'agit de mots pleins lemmatisés. Ce sont des noms, des verbes ou des adjectifs.

Nous avons tout d'abord étiqueté les textes en établissant une liste de mots-clefs. À partir de cette liste, nous avons procédé à l'extraction des termes représentatifs, tandis que les mots jugés non signifiants ont été éliminés. Cette extraction est fondée sur une analyse littéraire et linguistique [*cf*. Zhu Donglin, Ding Fan et Zhu Xiaojin, 2000 et Hong Zicheng, 2007]. Nous avons ensuite construit un petit dictionnaire contenant cent mots-clefs en tant que termes représentatifs. Exemple de présentation de ce dictionnaire :

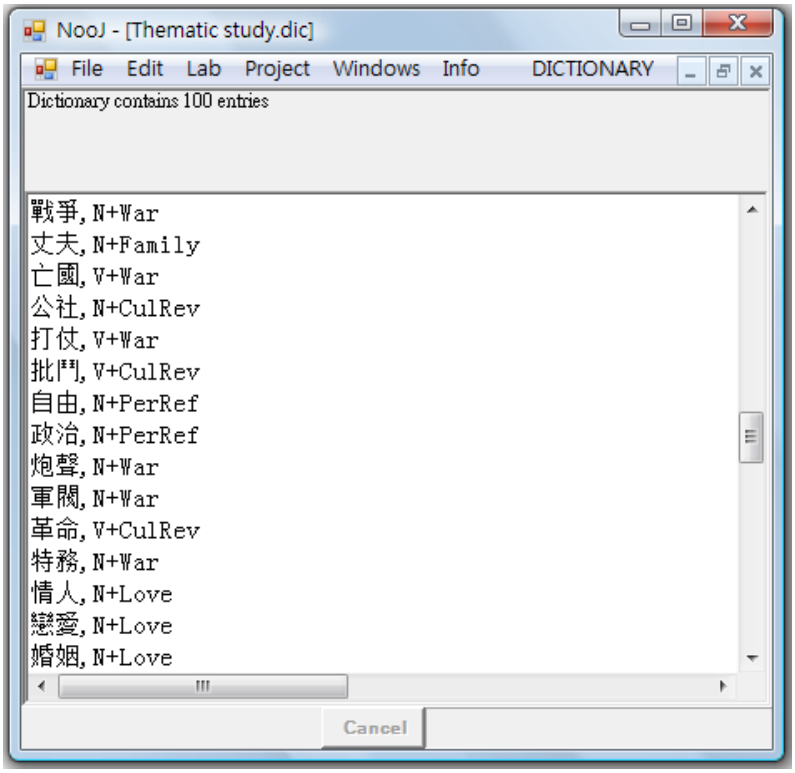


Figure 57 : Extrait du dictionnaire ***Thematic study***

Les occurrences de ces termes représentatifs ont ensuite été représentées par une matrice, qui exprime leur présence ou leur absence dans chaque texte. Nous présentons ci-dessous un extrait de cette matrice :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		**Titre des œuvres**	SRDQC	CL	NH	NHZ	DZZ	JIA	ZY	BC	CT	JHYY	HLHZ	CQ	SSTT
2		**Date de leur parution**	1919	1921	1923	1923	1928	1932	1932	1934	1934	1939	1940	1944	1944/1945/1948
3	**Termes représentatifs**	人民	0	0	0	5	3	1	1	3	0	9	0	0	32
4		女人	0	9	51	6	47	41	60	9	2	486	38	68	105
5		女兒	0	8	5	3	18	13	62	7	1	231	14	49	115
6		女朋友	0	0	0	0	0	0	6	0	0	1	0	3	3
7		父親	19	3	11	37	14	74	39	9	3	508	19	62	177
8		司令	0	0	0	0	0	4	1	0	0	76	0	0	15
9		民主	0	0	0	0	0	0	7	0	0	12	0	0	2
10		平等	0	0	1	3	0	2	1	0	0	7	0	0	6
11		母親	0	0	53	59	47	111	16	20	0	784	63	83	65
12		同志	0	0	1	17	0	0	6	0	0	6	0	0	2

Figure 58 : Extrait de la matrice unité d'information / unité de base

Cette matrice met en jeu l'usage des termes représentatifs. La figure ci-dessous permet de représenter la fréquence des termes représentatifs tels que 抗戰 *kàngzhàn* 'guerre de résistance', 愛情 *àiqíng* 'amour' et 解放 *jiěfàng* 'libérer' :

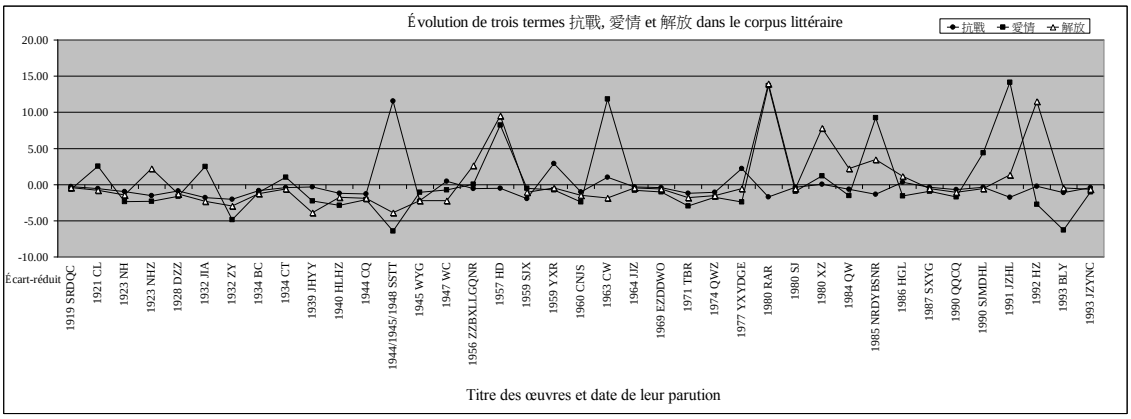


Figure 59 : La fréquence des trois termes 抗戰 *kàngzhàn*, 愛情 *àiqíng* et 解放 *jiěfàng*

8.1.2 Extraction des termes représentatifs

Les termes représentatifs constituent le vocabulaire caractéristique des textes. Leur classement, opéré selon leurs propriétés sémantiques, fait apparaître des catégories thématiques. Cinq thèmes sont définis : l'amour, la guerre, la famille, les réflexions personnelles et la révolution culturelle. Ce sont les thèmes les plus élaborés dans les œuvres littéraires du XX^{e} siècle [*cf*. Gong Hong, 1998 ; Zhu Donglin, Ding Fan et Zhu Xiaojin, 2000 et Hong Zicheng, 2007]. On peut catégoriser chaque texte par ces thèmes.

8.2 Évolution thématique

Lorsque les cent termes représentatifs sont classés par thèmes, la proportion des thèmes est mise en jeu et elle est traduite par l'écart-réduit.

8.2.1 Thème de l'amour

Les termes représentant le thème « amour » sont étiquetés <Love> dans le dictionnaire ***Thematic study***. Ce thème est central dans la littérature chinoise. L'extraction de dictionnaire se présente comme suit :

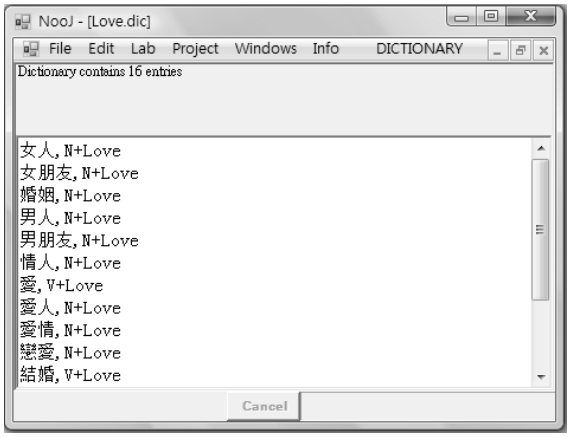


Figure 60 : Extraction de termes représentatifs de l'amour

En utilisant ce dictionnaire, on peut annoter ces termes représentatifs et les colorier dans un texte tel que *The Orange Is Red* de Qi Jun [1991] :

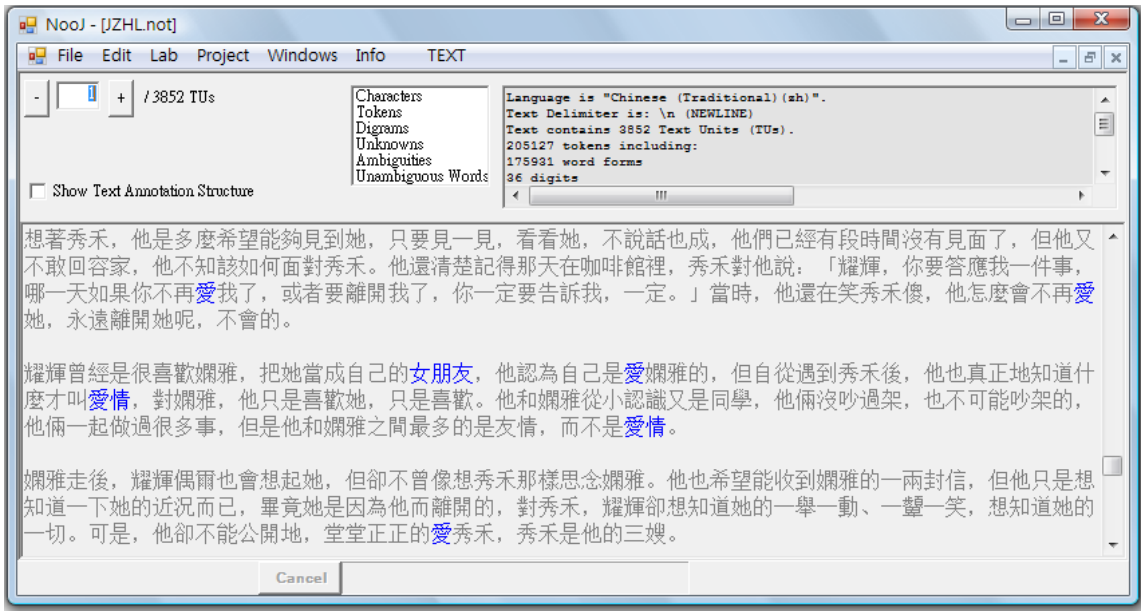


Figure 61 : Coloriage de termes représentant l'amour dans *The Orange Is Red*

Nous avons appliqué le dictionnaire pour retrouver les termes employés dans chaque œuvre littéraire. Ensuite, en prenant en compte les statistiques des termes utilisés, nous avons pu constater l'importance de ce thème chez chaque auteur. Le résultat est représenté par le graphique suivant :

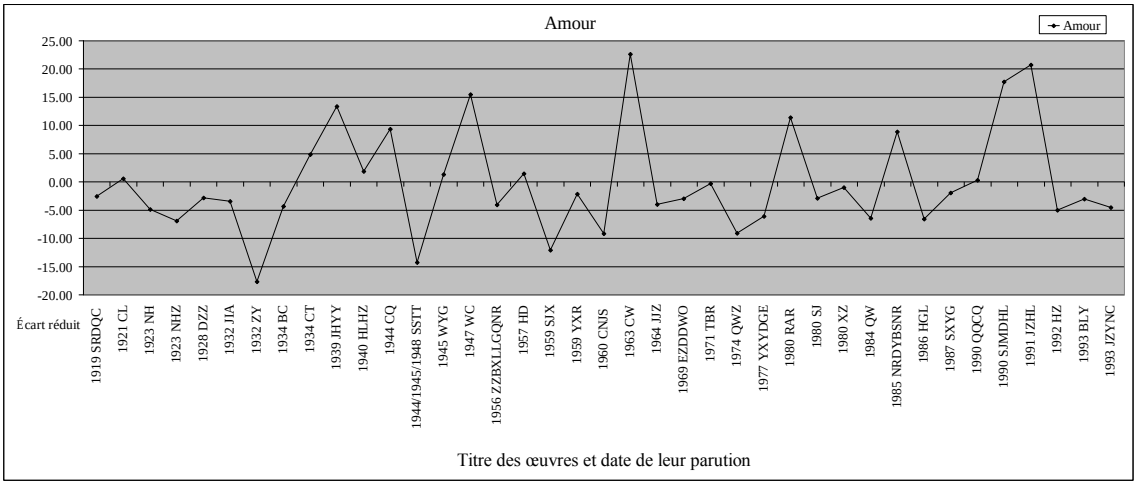


Figure 62 : Étude du thème « amour » dans les trente-neuf œuvres littéraires

On remarque que certains livres n'abordent guère le thème de l'amour. C'est le cas de *Minuit* de Mao Dun [1932], de *Quatre générations sous un même toit* de Lao She [1944/1945/1948] ou de *Three Family Lane* de Ouyang Shan [1959].

8.2.2 Thème de la famille

Dans les œuvres littéraires, les auteurs décrivent souvent la famille, les relations qu'entretiennent ses membres et les sentiments qu'ils éprouvent les uns entre les autres. Nous avons étiqueté <Family> les termes représentant ce thème dans le dictionnaire ***Thematic study***. L'extraction de ce dictionnaire se présente ainsi :

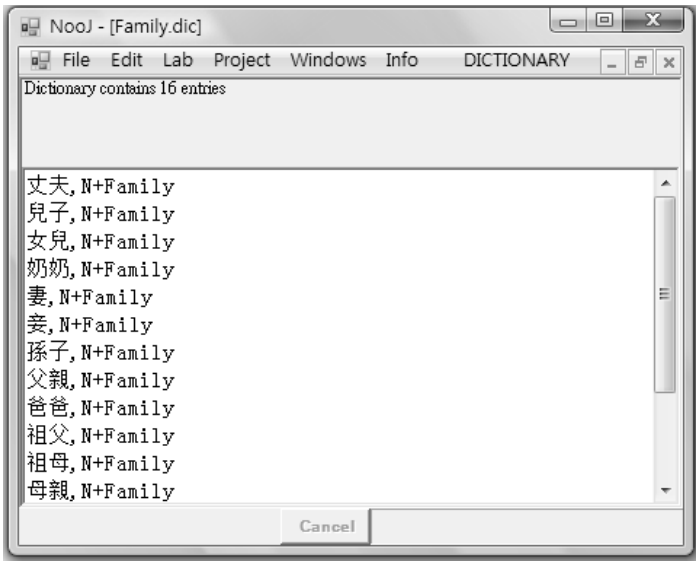


Figure 63 : Extraction de termes représentatifs de la famille

Nous avons utilisé ce dictionnaire pour reconnaître les termes représentatifs du thème « famille » dans *Plateau du cerf blanc* de Chen Zhongshi [1993]. Le texte colorié se présente comme suit :

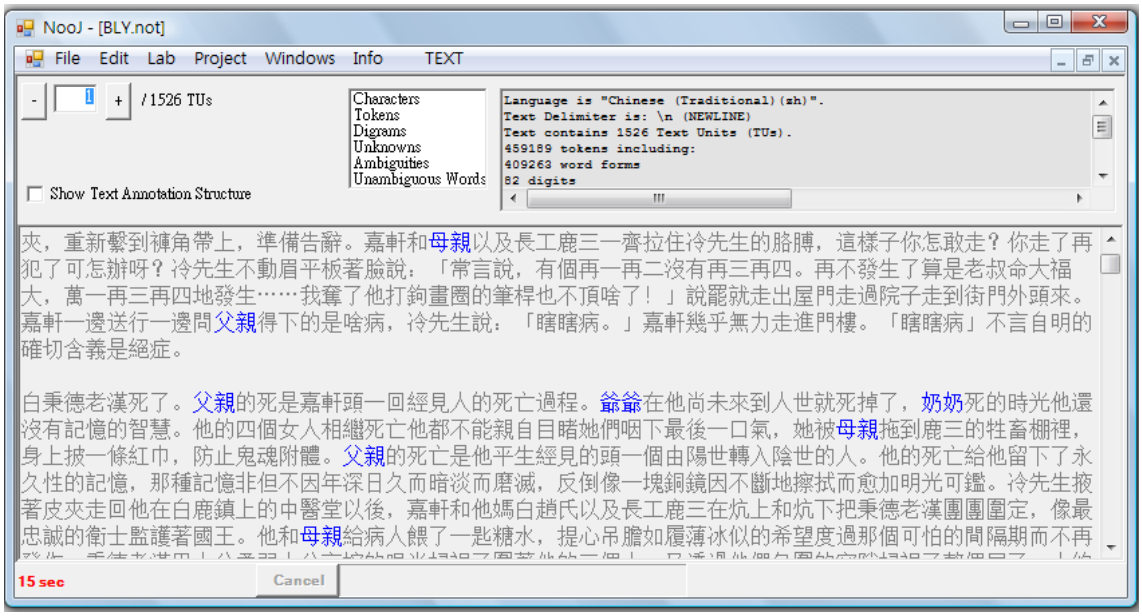


Figure 64 : Coloriage de termes représentant la famille dans *Plateau du cerf blanc*

Nous avons étudié ce thème de la famille dans les trente-neuf textes littéraires avec le résultat suivant :

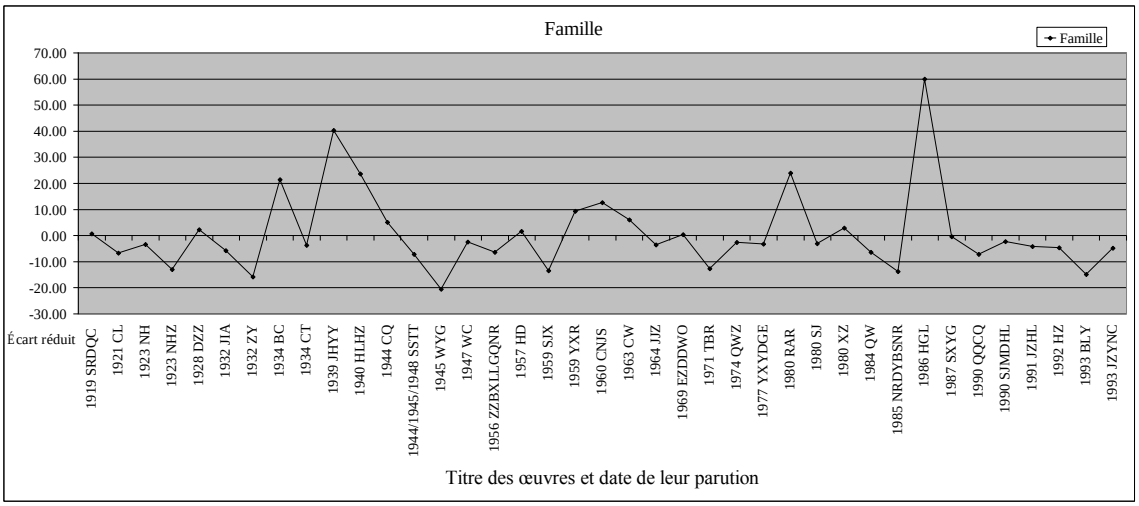


Figure 65 : Étude du thème de la famille dans les trente-neuf œuvres littéraires

Ce thème est largement traité par certains auteurs tels que Lin Yutang [*Un Moment à Pékin*, 1939], Dai Houying [*Oh l'homme, l'homme*, 1980] ou Mo Yan [*Le Clan du sorgho*, 1986].

8.2.3 Thème de la guerre

La guerre occupe une place importante dans la vie de nos sociétés qu'il s'agisse des deux Guerres Mondiales, de la guerre sino-japonaise, ou de la guerre civile. Les termes qui représentent ce thème ont reçu l'étiquette <War> dans le dictionnaire ***Thematic study***. Son extraction se présente comme suit :

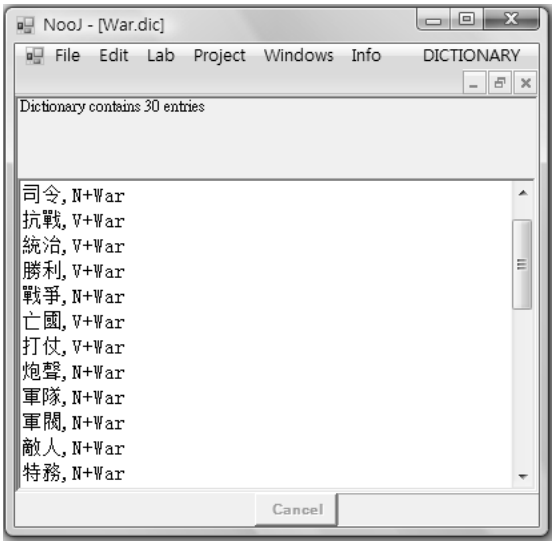


Figure 66 : Extraction de termes représentatifs de la guerre

Nous avons utilisé ce dictionnaire ***Thematic study*** pour reconnaître les termes utilisés par Lin Yutang dans *Un Moment à Pékin* [1939]. Les termes trouvés sont coloriés comme il apparaît dans la présentation suivante :

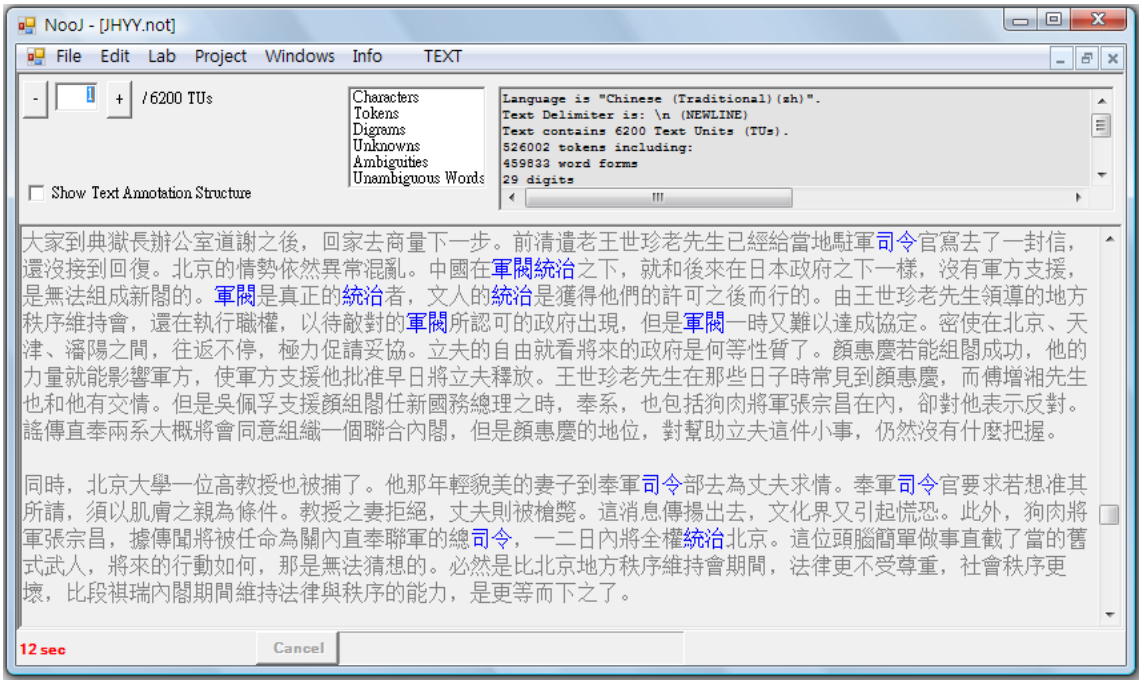


Figure 67 : Coloriage de termes représentant la guerre dans *Un Moment à Pékin*

Une fois posé le nombre de termes utilisés par chaque auteur, nous avons étudié l'importance du thème de la guerre dans leurs textes. L'importance de ce thème chez chaque auteur se trouve représentée par le graphique ci-dessous :

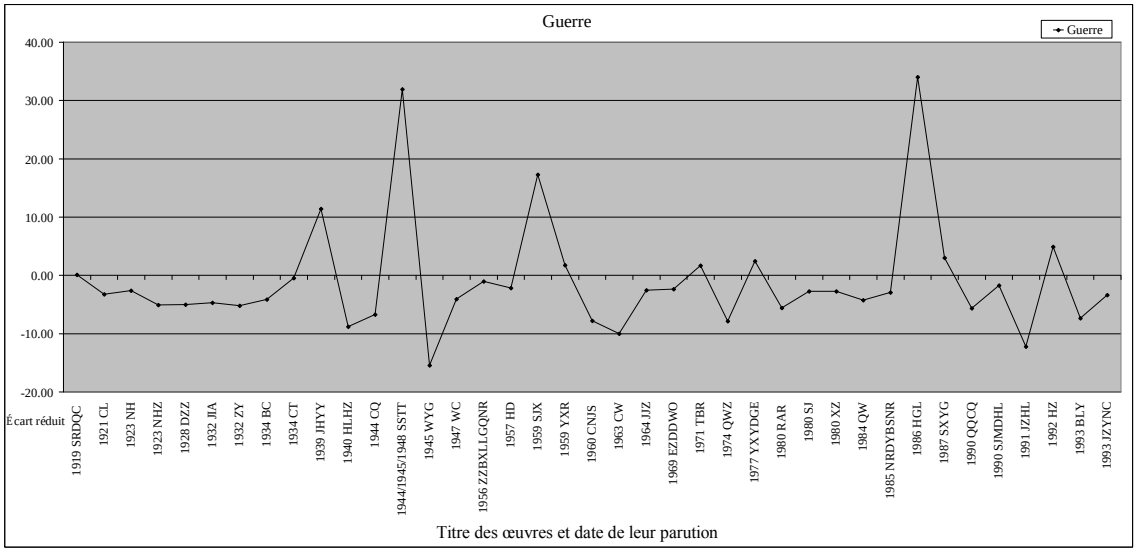


Figure 68 : Étude du thème de la guerre dans les trente-neuf œuvres littéraires

L'examen de ce graphique permet de constater que Lao She évoque aussi souvent la guerre dans *Quatre générations sous un même toit* [1944/1945/1948] que Mo Yan dans *Le Clan du sorgho* [1986].

8.2.4 Thème des réflexions personnelles

Les textes littéraires présentent également des réflexions personnelles sur divers sujets. Les termes qui présentent ce thème sont étiquetés <PerRef> dans le dictionnaire ***Thematic study***. L'extraction de ces termes se présente comme suit :

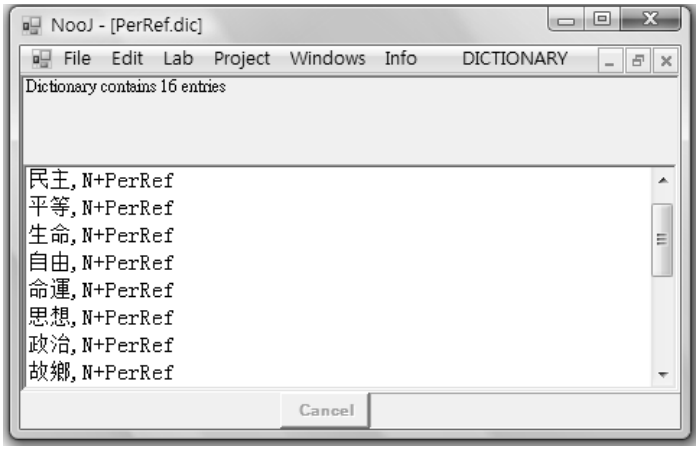


Figure 69 : Extraction des termes représentatifs des réflexions personnelles

En appliquant ce dictionnaire, les termes représentant les réflexions personnelles sont coloriés comme suit, dans l'exemple *Famille* de Ba Jin [1932] :

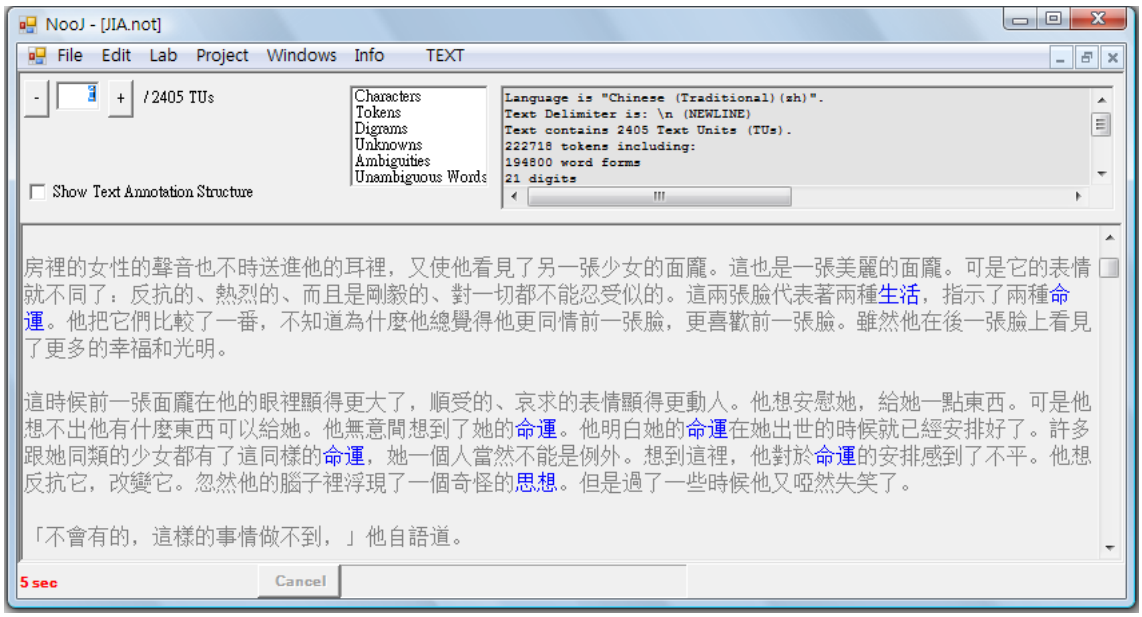


Figure 70 : Coloriage de termes introduisant des réflexions personnelles dans *Famille*

Nous avons d'abord effectué une enquête statistique pour connaître le nombre de termes utilisés dans chaque texte. Nous avons ensuite apprécié l'importance de ce thème chez les auteurs, avec le résultat suivant :

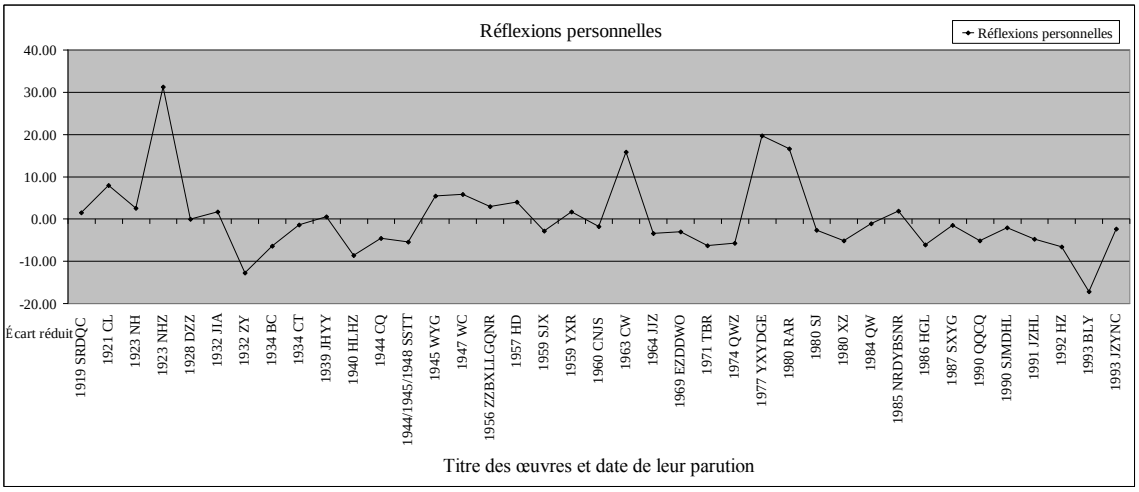


Figure 71 : Étude du thème des réflexions personnelles dans les trente-neuf œuvres littéraires

Ce thème revêt une importance particulière dans *L'Instituteur* de Ye Shengtao [1923], *Outside the Window* de Qiong Yao [1963] ou *The Orphan of Asia* de Wu Zhuoliu [1977].

8.2.5 Thème de la révolution culturelle

La révolution culturelle est un des thèmes les plus souvent traités dans les œuvres littéraires. Les termes représentatifs de ce thème sont étiquetés <CulRev> dans le dictionnaire ***Thematic study***. L'extraction de ce dictionnaire se présente ainsi :

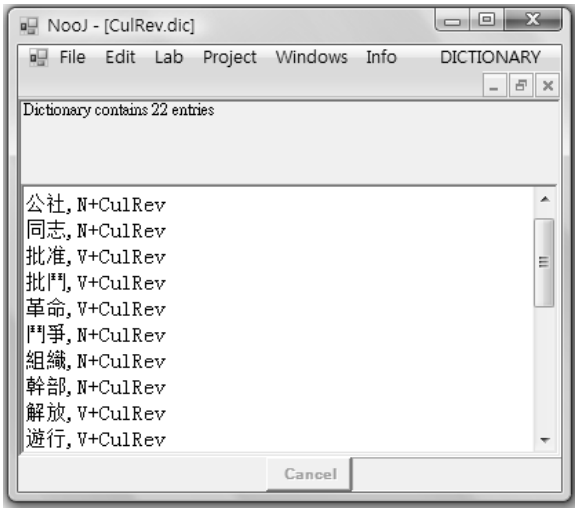


Figure 72 : Extraction de termes représentatifs de la révolution culturelle

En appliquant ce dictionnaire au *Vivre* de Yu Hua [1992], nous avons obtenu le texte colorié comme suit :

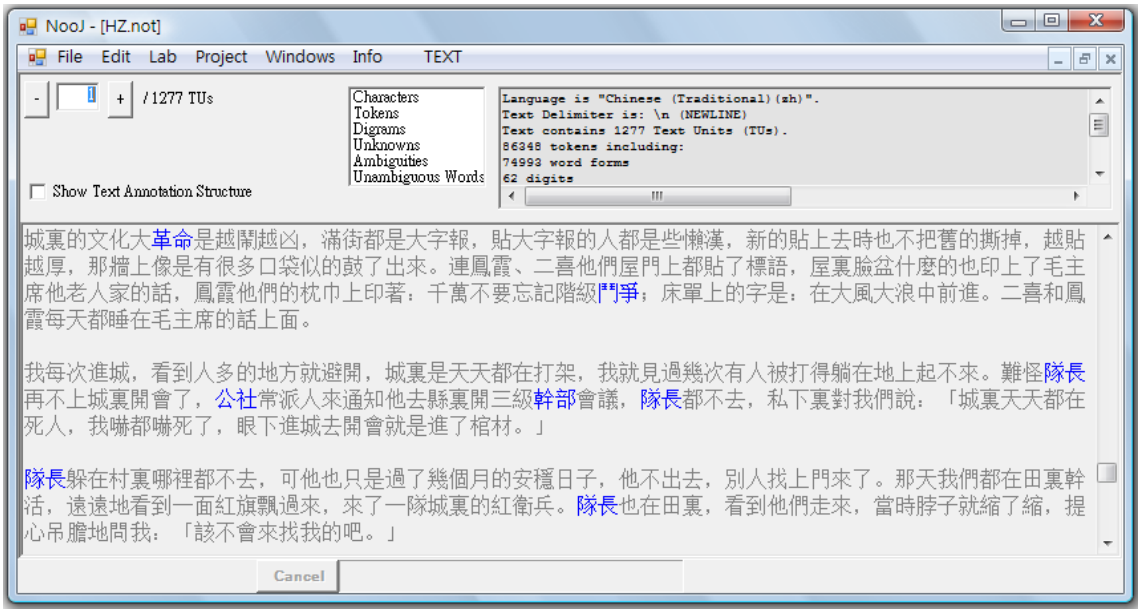


Figure 73 : Coloriage de termes représentant la révolution culturelle dans *Vivre*

Nous avons d'abord obtenu une statistique sur le nombre de termes utilisés chez les différents auteurs. Ensuite, nous avons étudié l'importance de ce thème dans leurs œuvres. Le résultat de cette étude est présenté par le graphique suivant :

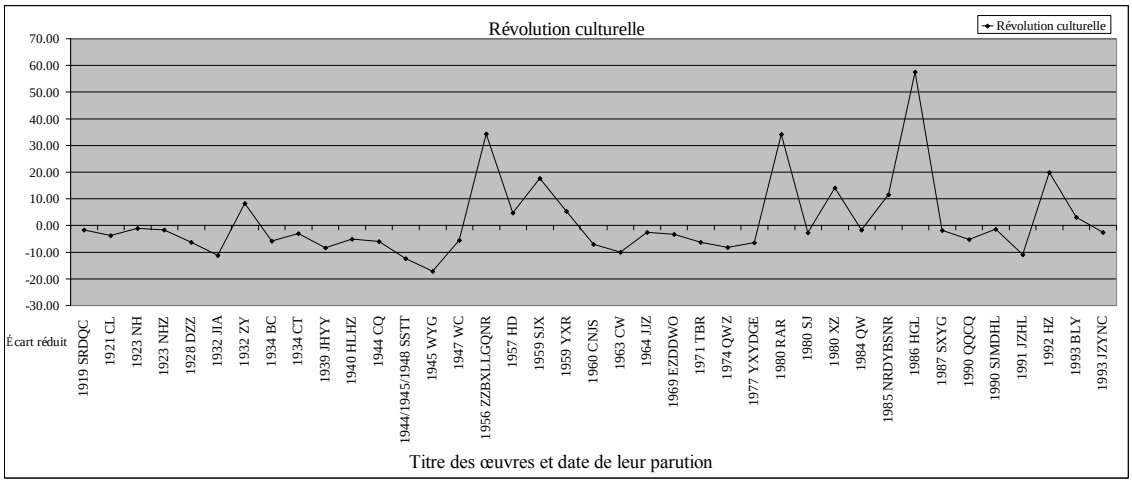


Figure 74 : Étude du thème de la révolution culturelle dans les trente-neuf œuvres littéraires

On constate que ce thème est plus présent chez les auteurs tels que Wang Meng [*Un Jeune homme arrivé récemment au département de l'organisation*, 1956], Dai Houying [*Oh l'homme, l'homme*, 1980] ou Mo Yan [*Le Clan du sorgho*, 1986].

8.2.6 Cinq thèmes dans *Quatre générations sous un même toit*

Les cinq thèmes peuvent être étudiés dans un seul texte. Par exemple le roman de Lao She *Quatre générations sous un même toit*. Les termes représentatifs des cinq thèmes peuvent être reconnus à l'aide du dictionnaire ***Thematic study***. Le texte apparait dans différentes couleurs selon les différents thèmes :

Figure 75 : Emploi de couleurs pour les termes représentant les cinq thèmes dans *Quatre générations sous un même toit*

Ainsi dans le roman de Lao She, *Quatre générations sous un même toit*, les thèmes abordés et leur proportion sont représentés par le schéma suivant :

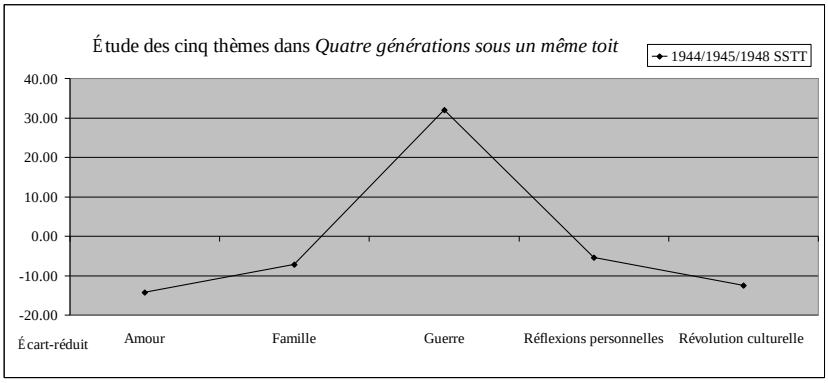


Figure 76 : Étude des cinq thèmes dans *Quatre générations sous un même toit*

8.2.7 Étude thématique selon le lieu d'origine de l'auteur

Il est possible de classer les textes d'un point de vue géographique, selon le pays d'origine de l'écrivain. Les textes sont classés en deux groupes : vingt-trois textes (soit 58,97 %) ont été écrits par des auteurs vivant sur le continent chinois, seize (soit 41,03 %) par des taïwanais.

L'étude de l'évolution des thèmes à l'intérieur de ces deux groupes d'œuvres est traduite dans le graphique ci-dessous :

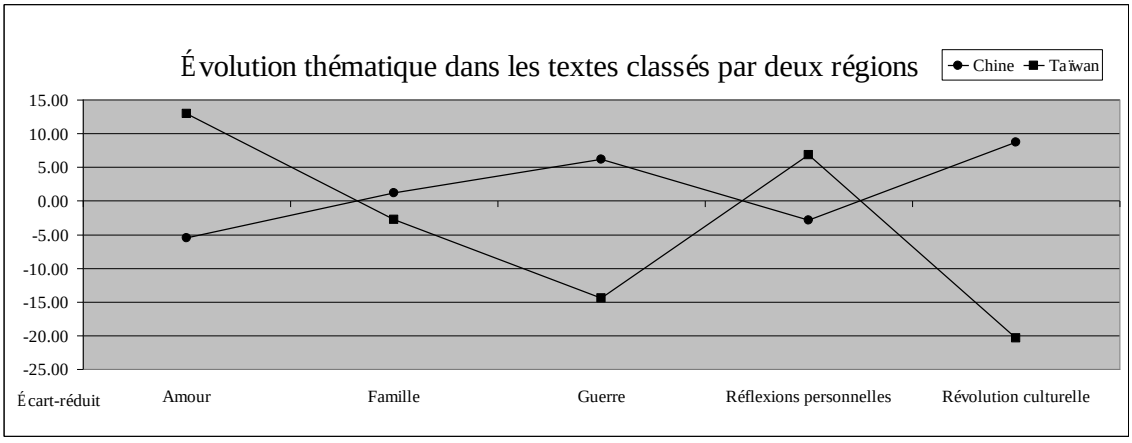


Figure 77 : Évolution thématique dans les textes classés selon le lieu d'origine de l'auteur

8.3 Conclusion

Nous avons utilisé les ressources linguistiques du module chinois de ***NooJ*** pour effectuer une analyse thématique de notre corpus. Cette analyse est fondée sur une étude des termes représentatifs qui forment un vocabulaire caractéristique. Le développement des thèmes induit la reprise de ce vocabulaire. La reconnaissance thématique s'appuie donc sur la distribution des termes représentatifs et sur une réflexion quant à leur fréquence. Cette reconnaissance permet de catégoriser les textes par thèmes. La catégorisation des textes met en évidence l'évolution de divers thèmes dans les textes littéraires publiés au XX^{e} siècle.

Conclusion

Nos recherches ont eu pour objectif de développer un ensemble de ressources linguistiques formalisées permettant d'analyser automatiquement des textes écrits en chinois moderne avec des caractères traditionnels. Pour ce faire, nous avons dû passer par les étapes suivantes :

1) Constitution du corpus

Pour obtenir des références lexicales et syntaxiques en chinois moderne, nous avons constitué un corpus dans lequel les textes sélectionnés ont été répartis en deux groupes : un groupe littéraire et un groupe journalistique. Ces données ont été récupérées à partir d'Internet. Les textes littéraires sont au nombre de trente-neuf et ont été écrits entre les années 1910 et les années 1990. Le volume de ces textes représente environ 7 300 000 caractères. Le groupe des textes journalistiques est constitué d'articles extraits de l'*United Daily News* publiés entre le 1er juillet 2007 et le 30 juin 2008. Ces textes comptent approximativement 7 000 000 de caractères. Par ailleurs, ce groupe contient aussi les articles de presse du *Quotidien du Peuple* du 1er au 31 mars 2008 comptant environ 3 600 000 caractères. Nous avons utilisé ce corpus pour construire les dictionnaires électroniques. Il nous a permis en outre de proposer des exemples de la langue chinoise à partir desquels nous avons pu décrire de façon formelle le vocabulaire.

2) Construction des dictionnaires électroniques

Le problème qui se pose avant tout traitement automatique du chinois est celui de la reconnaissance des unités lexicales. Pour le résoudre, nous avons construit, dans un premier temps, six dictionnaires électroniques ***ChDic*** (Chinese Dictionary), ***DicBkTitl*** (Dictionary of Book Titles), ***DicChSurn*** (Dictionary of Chinese Surnames), ***DicExpres*** (Dictionary of Expressions), ***GeoDic*** (Geographical Dictionary) et ***DicProNam*** (Dictionary of Proper Names). Le dictionnaire électronique ***ChDic*** comprend 79 046 entrées qui sont des unités lexicales générales telles qu'affixes, semi-affixes, mots simples, mots composés, locutions,

etc. Le dictionnaire électronique ***DicBkTitl*** propose 93 titres d'œuvres. Le dictionnaire électronique ***DicChSurn*** reconnaît 1 156 noms de famille chinois. Le dictionnaire électronique ***GeoDic*** regroupe 2 418 noms de lieu (pays, capitales, ports, etc.). Le dictionnaire électronique ***DicExpres*** regroupe 591 proverbes et expressions à double volet avec leurs variantes phrastiques. Le dictionnaire électronique ***DicProNam*** comprend 171 noms propres, notamment d'hommes politiques, de poètes, d'auteurs, etc. Ces dictionnaires permettent de couvrir l'ensemble du vocabulaire de notre corpus.

Dans ces six dictionnaires, les entrées sont associées à des informations linguistiques telles que les catégories (par exemple, **N** pour nom ; **V** pour verbe), les distributions (par exemple, **Hum** pour les noms humains), les quantifieurs appropriés aux différents noms, par exemple, <Ben> pour 書 *shū* 'livre' ; <Zhang> pour 椅子 *yǐzi* 'chaise'. Nous avons aussi associé aux entrées traits des étiquettes de réduplication qui permettent d'obtenir leurs formes réduplicatives. Les traits utilisés sont par exemple, <Ren>, <Gancui> ou <Lengqing>.

3) Développement des grammaires

Nous avons commencé par la description de phénomènes locaux. Cette description s'est effectuée avec des grammaires locales qui visent à reconnaître la composition numérique, les expressions de temps, les combinaisons des appellations de personnes. Par ailleurs, nous avons aussi construit des grammaires qui permettent de reconnaître cinq types différents de groupes nominaux noyaux, et de les étiqueter dans notre corpus.

4) Application : analyse thématique

Nous avons appliqué le module chinois à l'analyse thématique de trente-neuf textes littéraires. Cette analyse nous a permis d'obtenir des mots-clefs dans les données littéraires. Puis nous avons associé ces mots-clefs à des termes représentatifs. Ensuite, nous avons analysé les occurrences des termes représentatifs en les regroupant en diverses classes. Le classement des termes a permis de catégoriser chaque texte selon un ou plusieurs thèmes. Nous avons pu classer les textes par

thèmes. Grâce à cette catégorisation, nous avons pu connaître l'évolution thématique dans les trente-neuf textes littéraires. Cette évolution qui respecte un développement chronologique est représentée sous la forme statistique écart-réduit. Les résultats de cette application sont présentés dans la section 8.2.

Perspectives

Les résultats que nous avons obtenus permettent d'envisager des travaux futurs portant sur une description plus complexe et plus fine du chinois moderne. Nous nous proposons d'affiner les informations présentées dans les dictionnaires électroniques. Ces informations sont utilisées lors du développement des grammaires et de la mise en œuvre des requêtes. Ce perfectionnement permettra de diminuer les bruits [*cf.* 7.3.1.2] dans les différentes recherches. En outre, il serait intéressant d'étudier des groupes nominaux de structures plus complexes. Nous avons vu que la structure des groupes nominaux se fait plus complexe à mesure qu'augmente le nombre de leurs composants. Nous nous proposons donc de développer des grammaires qui visent à reconnaître des groupes nominaux de forme longue. Enfin, il nous semble aussi intéressant d'étudier l'évolution thématique en analysant non seulement les termes représentatifs mais aussi les groupes de mots. Ainsi, nous nous proposons d'effectuer des recherches sur des expressions particulières, ce qui permettrait d'exploiter plus en détail les différences inhérentes à chaque texte.

Annexe

Annexe 1 : Liste des textes littéraires[97]

Titres des œuvres	Noms des auteurs	Titres des textes numériques	Dates
斯人獨憔悴 *sīrén dú qiáocuì* ‘*Cette Personne vit dans sa souffrance solitaire*’	冰心 Bing Xin	SRDQC	1919
沉淪 *chénlún* ‘*Naufrag*’	郁達夫 Yu Dafu	CL	1921
吶喊 *nàhǎn* ‘*Cris d’appel*’	魯迅 Lu Xun	NH	1923
倪煥之 *NíHuànzhī* ‘*L’Instituteur*’	葉聖陶 Ye Shengtao	NHZ	1923
地之子 *dìzhī zǐ* ‘*Son of the Earth*’	臺靜農 Tai Jingnong	DZZ	1928
家 *jiā* ‘*Famille*’	巴金 Ba Jin	JIA	1932
子夜 *zǐyè* ‘*Minuit*’	茅盾 Mao Dun	ZY	1932
邊城 *biānchéng* ‘*Le Passeur du Chadong*’	沈從文 Shen Congwen	BC	1934
春桃 *Chūntáo* ‘*Spring Peach*’	許地山 Xu Dishan	CT	1934

[97] Nous donnons en français les titres donnés par les traducteurs [*cf*. Bady, 1993 et Lévy, 2000]. Il est à noter qu’il peut exister en français plusieurs titres pour un même titre en chinois. S’il n’existe pas de traduction française, nous proposons les titres anglais.

京華煙雲 *jīnghuá yānyún* ‘*Un Moment à Pékin*’	林語堂 Lin Yutang	JHYY	1939
呼蘭河傳 *hūlánhé zhuàn* ‘*Les Contes de la rivière Hulan*’	蕭紅 Xiao Hong	HLHZ	1940
傳奇 *chuánqí* ‘*Romances*’	張愛玲 Zhang Ailing	CQ	1944
四世同堂 *sìshìtóngtáng* ‘*Quatre générations sous un même toit*’	老舍 Lao She	SSTT	1944/1945/1948
未央歌 *wèiyānggē* ‘*A Novel Written in Modern China*’	鹿橋 Lu Qiao	WYG	1945
圍城 *wéichéng* ‘*La Forteresse assiégée*’	錢鍾書 Qian Zhongshu	WC	1947
組織部新來了個青年人 *zǔzhībù xīnláile gè qīngniánrén* ‘*Un Jeune homme arrivé récemment au département de l’organisatio*n’	王蒙 Wang Meng	ZZBXLLGQNR	1956
紅豆 *hóngdòu* ‘*Haricot*’	宗璞 Zong Pu	HD	1957
三家巷 *sānjiāxiàng* ‘*Three Family Lane*’	歐陽山 Ouyang Shan	SJX	1959
原鄉人 *yuánxiāngrén* ‘*Native Men*’	鍾理和 Zhong Lihe	YXR	1959
城南舊事 *chéngnán jiùshì* ‘*My Memories of Old Beijing*’	林海音 Lin Haiyin	CNJS	1960
窗外 *chuāngwài* ‘*Outside the Window*’	瓊瑤 Qiong Yao	CW	1963
將軍族 *jiāngjūnzú* ‘*A Race of Generals*’	陳映真 Chen Yingzhen	JJZ	1964

兒子的大玩偶 *érzi de dà wánǒu* ‘*La Poupée du fils*’	黃春明 Huang Chunming	EZDDWO	1969
臺北人 *Táiběirén* ‘*Gens de Taipei*’	白先勇 Bai Xianyong	TBR	1971
棋王 *qíwáng* ‘The *Chess Champion*’	張系國 Zhang Xiguo	QWZ	1974
亞細亞的孤兒 *yàxìyà de gū'ér* ‘*The Orphan of Asia*’	吳濁流 Wu Zhuoliu	YXYDGE	1977
人啊！人 *rén a rén* ‘*Oh l'homme, l'homme*’	戴厚英 Dai Houying	RAR	1980
受戒 *shòujiè* ‘*Ordination*’	汪曾祺 Wang Zengqi	SJ	1980
洗澡 *xǐzǎo* ‘*Le Bain*’	楊絳 Yang Jiang	XZ	1980
棋王 *qíwáng* ‘*Le Roi des échecs*’	阿城 A Cheng	QW	1984
男人的一半是女人 *nánrén de yībàn shì nǚrén* ‘*La Moitié de l'homme, c'est la femme*’	張賢亮 Zhang Xianliang	NRDYBSNR	1985
紅高粱 *hónggāoliáng* ‘*Le Clan du sorgho*’	莫言 Mo Yan	HGL	1986
四喜憂國 *Sìxǐ yōuguó* ‘*Lucky Worries about His Country*’	張大春 Zhang Dachun	SXYG	1987
妻妾成群 *qīqiè chéngqún* ‘*Épouses et concubines*’	蘇童 Su Tong	QQCQ	1990
世紀末的華麗 *shìjì mò de huálì* ‘*Splendeur fin de siècle*’	朱天文 Zhu Tianwen	SJMDHL	1990
橘子紅了 *júzi hóng le* ‘*The Orange Is Red*’	琦君 Qi Jun	JZHL	1991

活著 *huózhe* ‘*Vivre*’	余華 Yu Hua	HZ	1992
白鹿原 *báilùyuán* ‘*Plateau du cerf blanc*’	陳忠實 Chen Zhongshi	BLY	1993
嫁妝一牛車 *jiàzhuang yī niúchē* ‘*Une charrette à bœuf en échange de son épouse*’	王禎和 Wang Zhenhe	JZYNC	1993

Bibliographie

Références chinoises[98]

Cao Wei 曹炜. (2004). Xiàndài hànyǔ cíhuì yánjiū / jiù 现代汉语词汇研究 ‘Recherches sur les mots chinois modernes’. Beijing 北京 : Beijing daxue chubanshe 北京大学出版社 (2003).

Chao Yuen Ren 赵元任. (1979). Hànyǔ kǒuyǔ yǔfǎ 汉语口语语法 ‘A Grammar of Spoken Chinese’. Beijing 北京 : Shangwu yinshuguan 商务印书馆 (1968).

Chen Chengze 陈承泽. (1983). Guówénfǎ cǎochuàng 国文法草创 ‘Première rédaction de la grammaire chinoise’. 2e éd. Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Chen Duxiu 陳獨秀. (1917). Wénxué gémìnglùn 文學革命論 ‘À propos de la révolution littéraire’. In *Xin Qingnian 新青年 ‘Nouvelle jeunesse’*. Beijing 北京. Vol. 2. No. 6. p. 1-4.

Chen Liwei et Yuan Qi (éd.) 陈力为与袁琦主编. (1995). Jìsuàn yǔyánxué jìnzhǎn yǔ yìngyòng 计算语言学进展与应用 ‘Advances and Applications on Computational Linguistics’. Beijing 北京 : Qinghua daxue chubanshe 清华大学出版社.

Chen Sihe 陈思和. (1999). Zhōngguó dāngdài wénxuéshǐ jiàochéng 中国当代文学史教程 ‘Cours sur la littérature chinoise contemporaine’. Shanghai 上海 : Fudan daxue chubanshe 复旦大学出版社.

Diao yanbin 刁晏斌. (2006). Xiàndài hànyǔshǐ gàilùn 现代汉语史概论 ‘Introduction générale à l’histoire du chinois moderne’. Coll. « Yǔyánxué jiàocái xìliè 语言学教材系列 ‘Série de documents d’enseignement linguistiques’ ». Beijing 北京 : Beijing daxue chubanshe 北京大学出版社.

Duan Huiming et Zhou Qiang 段慧明和周强. (1994). Xiàndài hànyǔ yǔliàokù jiāgōng zhōng de qiēcí yǔ cíxìng biāozhù chǔlǐ 现代汉语语料库加工中的切词与词性标注处理

[98] Si les auteurs ont déjà traduit le titre de leur livre ou de leur article en anglais, nous le gardons. Dans le cas contraire, nous avons traduit les titres de livres ou d’articles en français.

‘Segmentation et annotation des catégories dans le corpus du chinois moderne’. In *Zhōngguó jìsuànjībào 中国计算机报 ‘China Information World’*. No. 21. Beijing 北京. p. 85-87.

Fang Guangdao 方光焘. (1997). Fāng Guāngdào yǔyánxué lùnwénjí 方光焘语言学论文集 ‘Essais linguistiques de Fang Guangdao’. Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Feng Zhiwei et Xu Fuji 冯志伟和许福吉. (2001). Quèdìng qiēcí dānwèi de mǒuxiē yǔfǎ yīnsù 确定切词单位的某些语法因素 ‘Some Grammatical Considerations in the Determination of Unit in Word Segmentation’. In *Journal of Chinese Language and Computer*. Vol. 11. No. 2. Ohio : Foreign Language Publications (The Ohio State University). p. 127-136.

Feng Zhiwei et Yang Quan 冯志伟和杨泉. (2008). Miànxiàng zhōngwén xìnxī / xí chǔlǐ de “n + n + n” jiégòu jùfǎ gōngnéng qíyì wèntí yánjiū / jiù 面向中文信息处理的 “n + n + n” 结构句法功能歧义问题研究 ‘Disambiguity Study of Structure “n + n + n” for Chinese Information Processing’. In *Hànyǔ xuéxí 汉语学习 ‘Chinese Language Learning’*. No. 6. Beijing 北京. p. 37-47.

Feng Zhiwei 冯志伟. (1992). Jìsuàn yǔyánxué duì lǐlùn yǔyánxué de tiǎozhàn 计算语言学对理论语言学的挑战 ‘Le défi lancé par la linguistique computationnelle à la linguistique théorique’. In *Yǔyán wénzì yìngyòng 语言文字应用 ‘Applied Linguistics’*. No. 1. Beijing 北京. p. 84-97.

Feng Zhiwei 冯志伟. (2001). Quèdìng qiēcí dānwèi de mǒuxiē fēiyǔfǎ yīnsù 确定切词单位的某些非语法因素 ‘The Non-grammatical Factors to Determine the Segmentation Element’. In *Zhōngwén jìsuànqì yǔyánxué qī / qí kān 中文计算器语言学期刊 ‘International Journal of Computational Linguistics and Chinese Language Processing’*. Vol. 15. No. 5. Beijing 北京. p. 8-14 et 51.

Feng Zhiwei 冯志伟 (a). (2006). Dāngqián zìrán yǔyán chǔlǐ fāzhǎn de jǐ gè tèdiǎn 当前自然语言处理发展的几个特点 ‘Several Features of Current Development of Natural Language Processing’. In *Jìnán dàxué huáwén xuéyuàn xuébào 暨南大学华文学院学报*

‘Journal of College of Chinese Language and Culture of Jinan University’. No. 1. Jinan 暨南 : Jinan University 暨南大学. p. 34-40.

Feng Zhiwei 冯志伟 (b). (2006). Zìrán yǔyán chǔlǐ de lìshǐ yǔ xiànzhuàng 自然语言处理的历史与现状 ‘The Past and Present of Natural Language Processing’. In *Zhōngguó wàiyǔ 中国外语 ‘Foreign Languages in China’*. Vol. 5. No. 1. Beijing 北京. p. 14-22.

Fu Huaiqing 符淮青. (1996). Hànyǔ cíhuìxué shǐ 汉语词汇学史 ‘Histoire de la lexicologie chinoise’. Coll. « Hànyǔ fāzhǎnshǐ cóngshū 汉语发展史丛书 ‘Série sur le développement de la langue chinoise’ ». Hefei 合肥 : Anhui jiaoyu chubanshe 安徽教育出版社.

Fu Huaiqing 符淮青. (2005). Xiàndài hànyǔ cíhuì (zēngdìngběn) 现代汉语词汇 (增订本) ‘Lexique du chinois moderne (version révisée)’. Coll. « Yǔyánxué jiàocái xìliè 语言学教材系列 ‘Série de documents d’enseignement linguistiques’ ». 2e éd. Beijing 北京 : Beijing daxue chubanshe 北京大学出版社 (1985).

Gao Gengsheng 高更生. (2006). Xiànxíng hànzì guīfàn wèntí 现行汉字规范问题 ‘Questions sur la normalisation des caractères chinois actuels’. Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Gao Heng 高亨. (1997). Gǔzì tōng jiǎhuì diǎn 古字通假會典 ‘Dictionnaire de mots d’emprunt et d’agrégat logique’. Beijing 北京 : Qilu shushe 齊魯書社.

Ge Benyi 葛本仪. (2002). Xiàndài hànyǔ cíhuìxué 现代汉语词汇学 ‘Lexicologie du chinois moderne’. Jinan 济南 : Shandong renmin chubanshe 山东人民出版社 (2001).

Gong Hong (éd.) 龚宏主编. (1998). Zhōngguó wénxuéshǐhuà (xiàndàijuàn) 中国文学史话 (现代卷) ‘Littérature chinoise (volume consacré à la littérature moderne)’. Vol.10. Changchun 长春 : Jilin renmin chubanshe 吉林人民出版社.

Guo Rui 郭锐. (1999). Yǔwén cídiǎn de cíxìng biāozhù wèntí 语文词典的词性标注问题 ‘Problèmes de l’annotation des catégories de mots dans les dictionnaires’. In *Zhōngguó yǔwén 中国语文 ‘Studies of the Chinese Language’*. No. 269. Beijing 北京. p. 8-25.

Guo Rui 郭锐. (2000). Biǎoshù gōngnéng de zhuǎnhuà hé / hàn “De” zì de zuòyòng 表述功能的转化和“的”字的作用 ‘The Conversion of the Expressional Functions and an Analysis of the Particle de in Mandarin Chinese’. In *Dāngdài yǔyánxué 当代语言学 ‘Contemporary Linguistics’*. Vol. 2. No. 1. Beijing 北京. p. 37-52 et 62.

Guo Rui 郭锐. (2001). Cípín yǔ cí de gōngnéng de xiāngguānxìng 词频与词的功能的相关性 ‘Rapport relatif entre les fréquences lexicales et les fonctions des mots’. In *Yǔyán yánjiū / jiù 语言研究 ‘Linguistic Researches’*. No. 80. Beijing 北京. p. 1-9.

Guo Rui 郭锐. (2004). Xiàndài hànyǔ cílèi yánjiū / jiù 现代汉语词类研究 ‘Recherches sur les catégories en chinois moderne’. Vol. 3. Coll. « Zhōngguó yǔyánxué wénkù 中国语言学文库 ‘Données de la linguistique chinoise’ ». Beijing 北京 : Shangwu yinshuguan 商务印书馆 (2002).

Guo Yikun, Huang Xuanjing et Wu Lide 郭以昆、黄萱菁和吴立德. (1999). Dàguīmó wénběn jiǎnsuǒ de xiànzhuàng jí fāzhǎn 大规模文本检索的现状及发展 ‘Status Quo and Development on Large Scale Text Retrieval’. In *Jìsuànjī gōngchéng 计算机工程 ‘Computer Engineering’*. Vol. 25. No. 3. Beijing 北京. p. 3-4 et 7.

Hao Tianyong et Zhang Chunxia 郝天永和张春霞. (2005). Hànyǔ zìdòng fēncí de yánjiū / jiù xiànkuàng yǔ kùnnán 汉语自动分词的研究现况与困难 ‘The State of the Art and Difficulties in Automatic Chinese Word Segmentation’. In *Xìtǒng fǎngzhēn xuébào 系统仿真学报 ‘Journal of System Simulation’*. Vol. 17. No. 1. Beijing 北京. p. 138-143 et 147.

He Jie 何杰. (2001). Xiàndài hànyǔ liàngcí yánjiū / jiù (xiūdìngbǎn) 现代汉语量词研究 (修订版) ‘Recherches sur les quantifieurs en chinois moderne (version révisée)’. 2e éd. Beijing 北京 : Minzu chubanshe 民族出版社.

He Jiuying 何九盈. (2005). Zhōngguó xiàndài yǔyánxuéshǐ 中国现代语言学史 ‘Histoire de la linguistique du chinois moderne’. 3e éd. Guangzhou 广州 : Guangdong jiaoyu chubanshe 广东教育出版社.

He Yuanjian 何元建. (2007). Shēngchéng yǔyánxué bèijǐng xià de hànyǔ yǔfǎ jí fānyì yánjiū / jiù 生成语言学背景下的汉语语法及翻译研究 ‘Recherches sur la grammaire

chinoise et la traduction appliquées à la linguistique formelle et transformationnelle'. Coll. « Yǔyánxué qiányán cóngshū 语言学前沿丛书 'Série sur la linguistique' ». Beijing 北京 : Beijing daxue chubanshe 北京大学出版社.

Hong Zicheng 洪子诚. (2007). Zhōngguó dāngdài wénxuéshǐ (xiūdìngbǎn) 中国当代文学史 (修订版) 'Littérature chinoise moderne (version révisée)'. Beijing 北京 : Beijing daxue chubanshe 北京大学出版社.

Hu Shi 胡適. (1917). Wénxué gǎiliáng chúyì 文學改良雛議 'Humbles suggestions primordiales en vue d'une meilleure littérature'. In *Xin Qingnian 新青年 'Nouvelle jeunesse'*. Beijing 北京. Vol. 2. No. 5. p. 18-28.

Hu Yushu 胡裕树. (1982). Shìlùn hànyǔ jùshǒu de míngcíxìng chéngfèn 试论汉语句首的名词性成分 'Réflexions sur les éléments nominaux au début des phrases chinoises'. In *Yǔyán jiāoxué yǔ yánjiū / jiù 语言教学与研究 'Language Teaching and Linguistic Studies'*. No. 4. Beijing 北京. p. 13-20.

Hu Yushu 胡裕树. (1995). Xiàndài hànyǔ (chóngdìngběn) 现代汉语 (重订本) 'Chinois moderne (version revue et corrigée)'. Shanghai 上海 : Shanghai jiaoyu chubanshe 上海教育出版社.

Huang Changning et Zhao Jun 黄昌宁和赵军. (1998). Jīyú lìzi de jīběn míngcí duǎnyǔ shìbié zhōng cíyǔ fēnbù xiāngsìdù de yánjiū / jiù 基于例子的基本名词短语识别中词语分布相似度的研究 'A Research on Distributional Word Similarity Used for Example Base BaseNP Recognition'. In *Móshì shìbié yǔ réngōng zhìnéng 模式识别与人工智能 'Pattern Recognition and Artificial Intelligence'*. Vol. 11. No. 2. Hefei 合肥. p. 40-46.

Huang Changning et Zhao Jun 黄昌宁和赵军. (1999). Jīyú zhuǎnhuàn de hànyǔ jīběn míngcí duǎnyǔ shìbié móxíng 基于转换的汉语基本名词短语识别模型 'A Transformation — Based Model for Chinese Base NP Recognition'. In *Zhōngwén xìnxī / xí xuébào 中文信息学报 'Journal of Chinese Information Processing'*. Vol. 13. No. 2. Beijing 北京. p. 1-7 et 39.

Huang Changning et Zhou Qiang 黄昌宁和周强. (1994). Miànxiàng yǔliàokù biāozhù de hànyǔ yīcún tǐxì de tàntǎo 面向语料库标注的汉语依存体系的探讨 'Approach to the

Chinese Dependency Formalism for the Tagging of Corpus'. In *Zhōngwén xìnxī / xí xuébào 中文信息学报 'Journal of Chinese Information Processing'*. Vol. 8. No. 3. Beijing 北京. p. 35-52.

Huang Changning et Zhou Qiang 黄昌宁和周强 (a). (1998). Hànyǔ gàilǜxíng shàngxiàwén wúguān yǔfǎ de zìdòng tuīdǎo 汉语概率型上下文无关语法的自动推导 'An Inference Approach for Chinese Probabilistic Context-Free Grammar'. In *Jìsuànjī xuébào 计算机学报 'Chinese Journal of Computers'*. Vol. 21. No. 5. Beijing 北京. p. 385-392.

Huang Changning et Zhou Qiang 黄昌宁和周强 (b). (1998). Hànyǔ jùfǎ guīzé de zìdòng gòuzào fāngfǎ yánjiū / jiù 汉语句法规则的自动构造方法研究 'Research of the Automatic Construction Methods for Chinese Context-Free Grammar'. In *Zhōngwén xìnxī / xí xuébào 中文信息学报 'Journal of Chinese Information Processing'*. Vol. 12. No. 3. Beijing 北京. p. 1-7.

Huang Changning, Zhou Qiang et Sun Maosong 黄昌宁、周强和孙茂松. (2000). Hànyǔ zuì cháng míngcí duǎnyǔ de zìdòng shìbié 汉语最长名词短语的自动识别 'Automatically Identify Chinese Maximal Noun Phrases'. In *Ruǎnjiàn xuébào 软件学报 'Journal of Software'*. Vol. 11. No. 2. Beijing 北京. p. 195-201.

Huang Chu-Ren 黃居仁. (1988). Zàixī guóyǔ 「lǐngshǔ zhǔyǔ」 jiégòu gàihuà cízǔ jiégòu yǔfǎ (*GPSG*) jí cíhuì gōngnéng yǔfǎ (*LFG*) zhī bǐjiào yánjiū / jiù 再析國語「領屬主語」結構概化詞組結構語法 (*GPSG*) 及詞彙功能語法 (*LFG*) 之比較研究 'A Reanalysis of Mandarin Chinese 'Possessive Subjects' — A Comparative Study of GPSG and *LFG*'. In *Hàn xué yánjiū / jiù 漢學研究 'Sinology Studies'*. Vol. 6. No. 2. Taipei 臺北. p. 143-168.

Jia Ning et Zhang Quan 贾宁和张全. (2007). Jīyú zuì dà shāngmóxíng de zhōngwén xìngmíng shìbié 基于最大熵模型的中文姓名识别 'Identification of Chinese Names Based on Maximum Entropy Model'. In *Jìsuànjī gōngchéng 计算机工程 'Computer Engineering'*. Vol. 33. No. 9. Beijing 北京. p. 31-33.

Jin Guangjin 靳光瑾. (2001). Xiàndài hànyǔ dòngcí yǔyì jìsuàn lǐlùn 现代汉语动词语义计算理论 ‘Théorie du calcul sur le sens des mots en chinois moderne’. Beijing 北京 : Beijing daxue chubanshe 北京大学出版社.

Jin Zhaozi 金兆梓. (1922). Guówénfǎ zhī yánjiū / jiù 國文法之研究 ‘Recherches sur la grammaire chinoise’. Shanghai 上海 : Zhonghua shuju 中華書局.

Kang Shiyong 亢世勇. (2002). 《Xiàndài hànyǔ xīnyǔcí xìnxī / xí diànzǐ cídiǎn》 de yánjiū / jiù yǔ shíxiàn 《现代汉语新语词信息电子词典》的研究与实现 ‘Development and Study of the “Modern Chinese New Words Information Electronic Dictionary”’. In *Zhōngwén jìsuànqì yǔyánxué qī / qí kān 中文计算器语言学期刊 ‘International Journal of Computational Linguistics and Chinese Language Processing’*. Vol. 7. No. 2. Beijing 北京. p. 89-99.

Kang Shiyong 亢世勇. (2004). Miànxiàng xìnxī / xí chǔlǐ de xiàndài hànyǔ yǔfǎ yánjiū / jiù 面向信息处理的现代汉语语法研究 ‘Recherches sur la grammaire du chinois moderne en fonction du traitement des informations’. Coll. « Yǔyán wénzì lǐlùn yǔ yìngyòng yánjiū / jiù wénkù 语言文字理论与应用研究文库 ‘Données des théories linguistiques et de leurs applications’ ». Shanghai 上海 : Shanghai cishu chubanshe 上海辞书出版社.

Lang Jun, Li Zhenghua, Li sheng, Liu Ting et Qin Bing 郎君、李正华、李生、刘挺和秦兵. (2008). Zhōngwén rénchēng míngcí duǎnyǔ dān-fùshù zìdòng shìbié 中文人称名词短语单复数自动识别 ‘Number Type Recognition of Chinese Personal Noun Phrase’. In *Zìdònghuà xuébào 自动化学报 ‘Acta Automatica Sinica’*. Vol. 34. No. 8. Beijing 北京. p. 972-979.

Li Jianhua et Wang Xiaolong 李建华和王晓龙. (2000). Zhōngwén rénmíng zìdòng shìbié de yī zhǒng yǒuxiào fāngfǎ 中文人名自动识别的一种有效方法 ‘An Effective Method on Automatic Identification of Chinese Name’. In *Gāojìshù tōngxùn 高技术通讯 ‘Chinese High Technology Letters’*. No. 2. Beijing 北京. p. 46-49.

Li Jinxi 黎錦熙. (1925). Xīnzhù guóyǔ wénfǎ 新著國語文法 ‘New Chinese Grammar of the National Language’. 3e éd. Shanghai 上海 : Shangwu yinshuguan 商務印書館 (1924) .

Li Xingjian. 李行健. (1998). Xiàndài hànyǔ guīfàn zìdiǎn 现代汉语规范字典 ‘Dictionnaire de normalisation du chinois moderne’. Beijing 北京 : Yuwen chubanshe 语文出版社 ‘Language and Literature Press’.

Li Xingjian (éd.) 李行健主编. (2002). Xiàndài hànyǔ yìxíngcí guīfàn cídiǎn 现代汉语异形词规范词典 ‘Dictionnaire de normalisation des variantes lexicales en chinois moderne’. Shanghai 上海 : Shanghai cishu chubanshe 上海辞书出版社.

Li Xingjian et Yu Zhihong 李行健与于志鸿. (2005). Xiàndài hànyǔ yìxíngcí yánjiū / jiù 现代汉语异形词研究 ‘Recherches sur les variantes lexicales en chinois moderne’. Shanghai 上海 : Shanghai cishu chubanshe 上海辞书出版社.

Lin Xingguang 林杏光. (1999). Cíhuì yǔyì hé / hàn jìsuàn yǔyánxué 词汇语义和计算语言学 ‘Sens des mots et linguistique computationnelle’. Beijing 北京 : Yuwen chubanshe 语文出版社.

Ling Shaowen et al. 凌紹雯等纂修 (a). (1996). Xīnxiū Kāngxī zìdiǎn (shàng) 新修康熙字典 (上) ‘Nouveau dictionnaire de Kangxi (I)’. Taipei 臺北 : Qiye shuju 啟業書局.

Ling Shaowen et al. 凌紹雯等纂修 (b). (1996). Xīnxiū Kāngxī zìdiǎn (xià) 新修康熙字典 (下) ‘Nouveau dictionnaire de Kangxi (II)’. Taipei 臺北 : Qiye shuju 啟業書局.

Liu Kaiying 刘开瑛. (2000). Zhōngwén wénkù zìdòng fēncí hé / hàn biāozhù 中文文库自动分词和标注 ‘Segmentation automatique et annotation d’un corpus chinois’. Coll. « Yǔyán yǔ jìsuànqì cóngshū 语言与计算器丛书 ‘Série sur la langue et l’informatique’ ». Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Lü Dexin, Zhou Qiaoli, Zhu Jiantao et Cai Dongfeng 吕德新、周俏丽、朱江涛和蔡东风. (2006). Jīyú qǐfāshì xìnxī / xí de zhōngwén xìngmíng shìbié fāngfǎ 基于启发式信息的中文姓名识别方法 ‘Chinese name identification method on heuristic information’. In *Shěnyáng hángkōng gōngyè xuéyuàn xuébào 沈阳航空工业学院学报 ‘Journal of Shenyang Institute of Aeronautical Engineering’*. Vol. 23. No. 3. Shenyang 沈阳. p. 35-37.

Lu Jianming 陆俭明. (1993). Hànyǔ jùzi de tèdiǎn 汉语句子的特点 ‘Particularités des phrases chinoises’. In *Hànyǔ xuéxí 汉语学习 ‘Language Learning’*. No. 73. Beijing 北京. p. 1-5.

Lu Jianming 陆俭明. (1994). Guānyú cí de jiānlèi wèntí 关于词的兼类问题 ‘Problèmes concernant les mots appartenant à plusieurs catégories’. In *Zhōngguó yǔwén 中国语文 ‘Studies of the Chinese Language’*. Beijing 北京. p. 28-34.

Lu Jianming 陆俭明. (2006). Bāshí niándài Zhōngguó yǔfǎ yánjiū / jiù (chóngpáiběn) 八十年代中国语法研究 (重排本) ‘Recherches sur la grammaire chinoise des années 80 (version revue et corrigée)’. Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Lü Shuxiang 吕叔湘. (1979). *Hànyǔ yǔfǎ fēnxī wèntí* 汉语语法分析问题 ‘Questions d’analyse grammaticales du chinois’. Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Lü Shuxiang 吕叔湘. (2002). Lǚ Shū / Shú xiāng quánjí 吕叔湘全集 ‘Collection de Lǚ Shū / Shú xiāng’, 19 vols. Vol 2 : Hànyǔ yǔfǎ lùnwénjí zēngdìngběn 汉语语法论文集增订本 ‘Collection d’essais sur la grammaire chinoise, version révisée’. Liaoning 辽宁 : Liaoning wanyou tushu faxing youxian gongsi 辽宁万有图书发行有限公司.

Luo Zhufeng (éd.). 羅竹風主編. (1994). Hànyǔ dàcídiǎn 漢語大詞典 ‘Grand dictionnaire chinois’. Taipei 臺北 : Taiwen donghua shuju 臺灣東華書局.

Ma Jianzhong 馬建忠. (1923). Măshì wéntōng 馬氏文通 ‘Chinese Grammar’. 16e éd. Shanghai 上海 : Shangwu yinshuguan 商務印書館 (1898).

Qian Liqun, Wen Rumin et Wu Fuhui 钱理群、温儒敏、吴福辉. (1998). Zhōngguó xiàndài wénxué sānshí nián (xiūdìngběn) 中国现代文学三十年 (修订本) ‘Trente ans de littérature chinoise moderne (version révisée)’. Beijing 北京 : Beijing daxue chubanshe 北京大学出版社.

Shen Jiaxuan 沈家煊. (2005). Xiàndài hànyǔ yǔfǎ de gōngnéng, yǔyòng, rènzhī yánjiū / jiù 现代汉语语法的功能、语用、认知研究 ‘Recherches sur les fonctions, les applications et les connaissances de la grammaire du chinois moderne’. Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Song Chunyang 宋春阳. (2005). Miànxiàng xìnxī / xí chǔlǐ de xiàndài hànyǔ “míng + míng” luóji / jí yǔyì yánjiū / jiù 面向信息处理的现代汉语 “名+名” 逻辑语义研究 ‘Recherches sur le sens de la forme N+N en chinois moderne en fonction du traitement des

informations'. Shanghai 上海 : Shanghai shiji chuban jituan xuelin chubanshe 上海世纪出版集团学林出版社.

Sun Hongkai, Hu Zengyi et Huang Xing (éd.) 孙宏开, 胡增益和黄行主编. (2007). Zhōngguó de yǔyán 中国的语言 'The Languages of China'. Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Sun Yinxin 孙银新. (2003). Xiàndài hànyǔ císù yánjiū / jiù 现代汉语词素研究 'Recherches sur les morphèmes en chinois moderne'. Coll. « Zhōnghuá xuérén cóngshū 中华学人丛书 'Série sur les connaissances de la langue chinoise' ». Beijing 北京 : Zhongguo wenshi chubanshe 中国文史出版社.

Wang Jue 王珏. (2001). Xiàndài hànyǔ míngcí yánjiū / jiù 现代汉语名词研究 'Recherches sur les mots nominaux en chinois moderne'. Shanghai 上海 : Huadong shifan daxue chubanshe 华东师范大学出版社.

Wang Li 王力. (1985). Zhōngguó xiàndài yǔfǎ 中国现代语法 'Grammaire du chinois moderne'. Beijing 北京 : Shangwu yinshuguan 商务印书馆 (1954).

Wang Li 王力. (2000). Yǔyánxué lùnwénjí 语言学论文集 'Recueil d'articles sur la linguistique'. Beijing 北京 : Shangwu yinshuguan 商务印书馆.

Wang Qilong 王启龙. (2003). Xiàndài hànyǔ xíngróngcí jìliàng yánjiū / jiù 现代汉语形容词计量研究 'Recherches quantitatives sur les adjectifs en chinois moderne'. Coll. « Zhōng-qīngnián yǔyánxuézhě wéncóng 中青年语言学者文丛 'Série des œuvres de chercheurs confirmés et de jeunes en linguistique' ». Beijing 北京 : Beijing yuyan daxue chubanshe 北京语言大学出版社.

Wen Duanzheng et Shen Huiyun (éd.) 温端政与沈慧云主编. (2004). Tōngyòng guànyòngyǔ cídiǎn 通用惯用语词典 'Dictionnaire des locutions fréquemment employées'. Coll. « Tōngyòng yǔyán wénzì xìliè gōngjùshū 通用语言文字系列工具书 'Dictionnaires de langue et d'écriture' ». Beijing 北京 : Yuyan chubanshe 语言出版社.

Weng Fuliang et Wang Yeyi 翁富良与王野翊. (1998). Jìsuàn yǔyánxué dǎolùn 计算语言学导论 'Introduction to Computational Linguistics'. Coll. « Dāngdài yǔyánxué lǐlùn cóngshū 当代语言学理论丛书 'Contemporary Linguistic Theory Series' ». Huang

Zhengde et Xu debao (éd.) 黄正德与许德宝主编. Beijing 北京 : Zhongguo shehui xueke chubanshe 中国社会学科出版社.

Xie Jialun 謝佳倫. (1998). Yíchuán yǎnsuànfǎ yìngyòng yú zhōngwén duàncí zhī yánjiū / jiù 遺傳演算法應用於中文斷詞之研究 ‘A Genetic Approach to Chinese Text Segmentation’. Zhongli 中壢 : Guoli Zhongyang daxue 國立中央大學.

Xin Ju 辛菊. (2005). Xiàndài hànyǔ yǔfǎ xiūcí yánjiū / jiù 现代汉语语法修辞研究 ‘Recherches sur les styles grammaticaux et la rhétorique en chinois moderne’. Taiyuan 太原 : Shuhai chubanshe 书海出版社.

Xing Fuyi 邢福义. (1989). Cílèi pànbié sì yàodiǎn 词类判别四要点 ‘Quatre principes applicables à la détermination des catégories des mots’. In *Yǔyán jiāoxué yǔ yánjiū / jiù 语言教学与研究 ‘Language Teaching and Linguistic Studies’*. No. 3. Beijing 北京. p. 60-68.

Xiong Dunshen (éd.) 熊鈍生主編. (1986). Cíhǎi 辭海 ‘Lexique chinois’. Taipei 臺北 : Taiwen Zhonghua shuju 臺灣中華書局.

Xu Shen 許慎. (2005). Shuōwén jiězì zhù 說文解字注 ‘Shuowen jiezi annonté’. Taipei 臺北 : Dingyuan wenhua shiye youxian gongsi 頂淵文化事業有限公司.

Yang Chun 杨春. (2004). Xiàndài hànyǔ zhōng de yìxíngcí 现代汉语中的异形词 ‘Variantes lexicales en chinois moderne’. Beijing 北京 : Huaxia chubanshe 华夏出版社.

Yang Xinzhang 杨信彰. (2006). Míngcíhuà zài yǔtǐ zhōng de zuòyòng — jīyú xiǎoxíng yǔliàokù de yī xiàng fēnxī 名词化在语体中的作用—基于小型语料库的一项分析 ‘A Corpus-based Approach to the Role of Nominalization in Styles’. In *Wàiyǔ diànhuà jiāoxué 外语电化教学 ‘Computer-assisted Foreign Language Education in China’*. No. 108. Beijing 北京. p. 1-7.

Yang Xipeng 杨锡彭. (2003). Hànyǔ yǔsù lùn 汉语语素论 ‘Morphèmes chinois’. Nanjing 南京 : Nanjing daxue chubanshe 南京大学出版社.

Yuan Yulin 袁毓林. (2001). Jìsuàn yǔyánxué de lǐlùn fāngfǎ hé / hàn yánjiū / jiù qǔxiàng 计算语言学的理论方法和研究取向 ‘Theoretical Methodology and Study Orientation of Computational Linguistics’. In *Zhōngguó shèhuì kēxué 中国社会科学 ‘Social Sciences in China’*. No. 4. Beijing 北京. p. 157-168 et 206.

Yuan Yulin 袁毓林. (2002). Míngcí dàibiǎo dòngcí duǎnyǔ hé / hàn dàicí suǒzhǐ de bōdòng 名词代表动词短语和代词所指的波动 ‘Nouns Standing for VP and Fluctuation of Pronouns’ Referent’. In *Zhōngguó yǔwén 中国语文 ‘Studies of the Chinese Language’*. No. 287. Beijing 北京. p. 99-110 et 190.

Zhan Weidong 詹卫东. (2000). Miànxiàng zhōngwén xìnxī / xí chǔlǐ de xiàndài hànyǔ duǎnyǔ jiégòu guīzé yánjiū / jiù 面向中文信息处理的现代汉语短语结构规则研究 ‘A Study of Constructing Rules of Phrases in Contemporary Chinese for Chinese Information Processing’. Coll. « Zhōngwén xìnxī / xí chǔlǐ cóngshū 中文信息处理丛书 ‘Série sur le traitement informatique du chinois’ ». Beijing 北京 : Qinghua daxue chubanshe et Guangxi kexue jishu chubanshe 清华大学出版社与广西科学技术出版社.

Zhang Qiong 章琼. (2004). Xiàndài hànyǔ tōngyòngzì duìyìng yìtǐzì zhěnglǐ 现代汉语通用字对应异体字整理 ‘Présentation de caractères d’usage courant et de leurs variantes en chinois moderne’. Chengdu 成都 : Sichuan chuban jituan bashu shushe 四川出版集团巴蜀书社.

Zhou Jian (éd.) 周荐主编. (2007). 20 shìjì Zhōngguó cíhuìxué 20 世纪中国词汇学 ‘Lexicologie du chinois du XX^{e} siècle’. Coll. « 20 shìjì Zhōngguó yǔyánxué cóngshū 20 世纪中国语言学丛书 ‘Série sur la langue chinoise du XX^{e} siècle’ ». Xing Fuyi (éd.) 邢福义总主编. Beijing 北京 : Zhongguo renmin daxue chubanshe 中国人民大学出版社.

Zhou Qiang et Yu Shiwen 周强和俞士汶. (1993). Yī zhǒng qiēfēn hé / hàn cíxìng biāozhù xiāng rónghé de hànyǔ yǔliàokù duōjí chǔlǐ fāngfǎ 一种切分和词性标注相融合的汉语语料库多级处理方法 ‘Méthode de traitement des différents niveaux de langue par application conjointe de la segmentation et de l’annotation des catégories dans le corpus chinois’. In *Chen Liwei et Yuan Qi (éd.), 陈力为与袁琦主编， Jìsuàn yǔyánxué yánjiū / jiù yǔ yìngyòng 计算语言学研究与应用 ‘Advances and Applications on Computational Linguistics’*. Beijing 北京 : Beijing yuyan xueyuan chubanshe 北京语言学院出版社. p. 126-131.

Zhou Qiang et Yu Shiwen 周强和俞士汶. (1996). Hànyǔ duǎnyǔ biāozhù biāojìjí de quèdìng 汉语短语标注标记集的确定 ‘Definition of the Tagset for Annotating Chinese

Phrase'. In *Zhōngwén xìnxī / xí xuébào 中文信息学报 'Journal of Chinese Information Processing'*.Vol. 10. No. 4. Beijing 北京. p. 1-11.

Zhou Qiang et Zhang Yuqi 周强和张昱琪. (2002). Hànyǔ jīběn duǎnyǔ de zìdòng shìbié 汉语基本短语的自动识别 'Automatic Identification of Chinese Base Phrases'. In *Zhōngwén xìnxī / xí xuébào 中文信息学报 'Journal of Chinese Information Processing'*. Vol. 16. No. 6. Beijing 北京. p. 1-8.

Zhou Qiang 周强 (a). (1995). Jīyú yǔliàokù hé / hàn miànxiàng tǒngjìxué de zìrán yǔyán chǔlǐ jìshù jièshào 基于语料库和面向统计学的自然语言处理技术介绍 'Introduction to the Corpus-Based, Statistics-Oriented Natural Language Processing Techniques'. In *Jìsuànjī kēxué 计算机科学 'Computer Science'*. Vol. 22. No. 4. Beijing 北京. p. 36-40.

Zhou Qiang 周强 (b). (1995). Guīzé hé / hàn tǒngjì xiāng jiéhé de hànyǔ cílèi biāozhù fāngfǎ 规则和统计相结合的汉语词类标注方法 'Approche d'une combinaison de la description des règles linguistiques et statistiques en vue d'une catégorisation des mots'. In *Zhōngwén xìnxī / xí xuébào 中文信息学报 'Journal of Chinese Information Processing'*.Vol. 9. No. 3. Beijing 北京. p. 1-10.

Zhou Qiang 周强. (1996). Hànyǔ yǔliàokù de duǎnyǔ zìdòng huàfēn hé / hàn biāozhù yánjiū / jiù 汉语语料库的短语自动划分和标注研究 'Phrase Bracketing and Annotating on Chinese Language Corpus' (Thèse). Beijing 北京 : Peking University 北京大学.

Zhou Qiang 周强. (1997). Hànyǔ duǎnyǔ de zìdòng huàfēn hé / hàn biāozhù 汉语短语的自动划分和标注 'Automatically Bracket and Tag Chinese Phrases'. In *Zhōngwén xìnxī / xí xuébào 中文信息学报 'Journal of Chinese Information Processing'*. Vol. 11. No. 1. Beijing 北京. p. 1-10.

Zhou Qiang 周强. (1998). Hànyǔ jùfǎ guīzé de zìdòng huòqǔ jí qí yìngyòng 汉语句法规则的自动获取及其应用 'Automatic Syntactic Knowledge Acquisition for the Chinese Language and its Applications' (Rapport de Post-Doc). Beijing 北京 : Tsinghua University 清华大学.

Zhou Qiang, Zhan Weidong et Ren Haibo 周强、詹卫东和任海波. (2001). Gòujiàn dàguīmó hànyǔ yǔkuàikù 构建大规模汉语语块库 'Build a large scale Chinese Functional

Chunk Bank'. In *Huang Changning, Zhang Pu (éd.), 黄昌宁、张普主编, Zìrán yǔyán lǐjiě yǔ jīqì / qì fānyì 自然语言理解与机器翻译 'Compréhension de la langue naturelle et traduction automatique'*. Beijing 北京 : Qinghua daxue chubanshe 清华大学出版社. p. 102-107.

Zhu Dexi 朱德熙. (1985). Xiàndài shūmiàn hànyǔ lǐ de xūhuà dòngcí hé / hàn míngdòngcí 现代书面汉语里的虚化动词和名动词 'Des verbes délexicalisés et des verbes nominalisés en chinois moderne à l'écrit'. In *Dì-yī jiè guójì hànyǔ jiāoxué tǎolùn huìyì wénxuǎn 第一届国际汉语教学讨论会议文选 'Première conférence internationale sur l'enseignement du chinois (une anthologie)'*. Beijing 北京. p. 10-16.

Zhu Dexi 朱德熙. (2001). Xiàndài hànyǔ cílèi yǔfǎ yánjiū / jiù 现 代汉语词类语法研究 'Recherches sur les catégories en chinois moderne et sur leurs fonctions grammaticales'. Coll. « Shāngwù yìnshūguǎn wénkù 商务印书馆文库 'The Commercial Press Library' ». Beijing 北京 : Shangwu yinshuguan 商务印书馆 (1980).

Zhu Dexi 朱德熙. (2004). Yǔfǎ jiǎngyì 语法讲义 'Cours de grammaire'. Beijing 北京 : Shangwu yinshuguan 商务印书馆 (1982).

Zhu Donglin, Ding Fan et Zhu Xiaojin (éd.) 朱棟霖、丁帆、朱曉進主編. (2000). Èrshí shìjì Zhōngguó wénxuéshǐ (shàngcè) 二十世紀中國文學史 (上冊) 'Littérature chinoise du XX^{e} siècle (I)'. Taïpei 臺北 : Wen shizhe chubanshe 文史哲出版社.

Zhu Donglin, Ding Fan et Zhu Xiaojin (éd.) 朱棟霖、丁帆、朱曉進主編. (2000). Èrshí shìjì Zhōngguó wénxuéshǐ (xiàcè) 二十世紀中國文學史 (下冊) 'Littérature chinoise du XX^{e} siècle (II)'. Taïpei 臺北 : Wen shizhe chubanshe 文史哲出版社.

Références anglaises et françaises

Abeillé Anne. (1998). Grammaires génératives et grammaires d'unification. In *Langages*. Vol. 32. No. 129. Paris : Armand Colin. p. 24-36.

Abeillé Anne. (2007). Les grammaires d'unification. Coll. « Langues et syntaxe ». Paris : Lavoisier.

Allen James. (1987). Natural Language Understanding. California : The Benjamin/Cummings Publishing Company, INC.

Bady Paul. (1993). La littérature chinoise moderne. Coll. « Que sais-je ? ». Paris : Presses Universitaires de France.

Beesley Kenneth R. and Karttunen Lauri. (2003). Finite-State Morphology. California : CSLI Publications Stanford.

Bresnan Joan and Kaplan John Ronald. (1982). Lexical Functional Grammar: A Formal System for Grammatical Representation. In *J. Bresnan (dir.), The Mental Representation of Grammatical Relations*. Cambridge : MIT Press. p. 173-281.

Busemann Stephan. (1991). Using Pattern-Action Rules for the Generation of GPSG Structures From MT-Oriented Semantics. In *Mylopoulos John and Reiter Raymond (eds.), IJCAI-91 : Proceedings of the twelfth International Joint Conference on Artificial Intelligence*. Sydney : Morgan Kaufmann Publishers Inc. p. 1003-1009.

Chen Keh-jiann, Huang Chu-Ren and Chang Li-Li. 1997. Segmentation Standard for Chinese Natural Language Processing. In *Zhōngwén jìsuànjī yǔyánxué* qī / qí kān *中文計算機語言學期刊 'International Journal of Computational Linguistics and Chinese Language Processing'*. Vol. 2. No. 2. Taipei 臺北. p. 47-62.

Cori Marcel et Jacqueline Léon. (2002). La constitution du TAL. Étude historique des dénominations et des concepts. In *Traitement Automatique des Langues*. Vol. 43. No. 3. Paris. p. 21-55.

Dalbera Jean-Philippe. (2002). Le corpus entre données, analyse et théorie. In *Corpus*. No. 1. Nice. p. 89-104.

Denoual Etienne. (2006). Méthodes en caractères pour le traitement automatique des langues (Thèse). Grenoble : Université Joseph Fourier.

De Saussure Ferdinand. (1971). Cours de linguistique générale. Paris : Payot.

Huang Chu-Ren and Chang Ru-Yng. (2004). Categorical ambiguity and information content : A Corpus-based study of Chinese. In *Tao Hongyin (ed.) 陶红印主编, Special Issue on Corpora, Language Use, and Grammar Hànyǔ yǔyán hé / hàn jìsuàn xuébào 汉语语言和计算学报 'Journal of Chinese Language and Computing'*. Vol. 14. No. 2.

Zhōngwén yǔ dōngfāng yǔyán xìnxī / xí chǔlǐ xuéhuì 中文与东方语言信息处理学会 Singapore : Chinese and Oriental Languages Information Processing Society. p. 157-165.

Huang, Chu-Ren. 1986. Coordination Schemas and Chinese NP Coordination in GPSG. In *Cahiers de Linguistique Asie Orientale*. Vol. 15. No. 1. Paris. p. 107-127.

Kabore Raphaël. (1998). La réduplication. Coll. « Faits de langue ». Vol. 6. No. 11. p. 359-376.

Kaplan Ronald, Karttunen Lauri, Kay Martin, Pollard Carl, Sag Ivan A., Shieber Stuart, and Zaenen Annie. (1986). Unification and Grammatical Theory. In *Proceedings of the Fifth Annual Meeting of the West Coast Conference on Formal Linguistics*. Stanford : Special Libraries Association and CSLI Publications. p. 238-254.

Kleene Stephen Cole. (1951). Representation of Events in Nerve Nets and Finite Automata (Research Memorandum of Project Rand). Santa Minca : Rand Corporation.

Lee Hsiang-Ping, Huang Chu-Ren, Chen Keh-jiann, Chen Chun-Ling, Chen Yong-Xiang and Weng Cui-Xia. (2005). The Sinica Sense Management System: Design and Implementation. In *Huang Chu-Ren and Ji Donghong (ed.), CLSW-5 : Zhōngwén jìsuànjī yǔyánxué qī / qí kān 中文計算機語言學期刊 'International Journal of Computational Linguistics and Chinese Language Processing'*. Vol. 10. No. 4. Taipei 臺北. p. 417-430.

Lévy André. (2000). Dictionnaire de littérature chinoise. Coll. « Quadrige dicos poche ». Paris : Presses Universitaires de France.

Li Charles N. and Thompson Sandra A. (1989). Hànyǔ yǔfǎ 漢語語法 Mandarin Chinese: A Functional Reference Grammar. Berkeley, Los Angeles and London : University of California Press.

Lin Huei-Chi et Silberztein Max. (2007). Ressources lexicales chinoises pour le TALN. In *Actes de TALN 2007 RECITAL 2007 : XIV*e *Conférence sur le Traitement Automatique des Langues Naturelles*. Toulouse : Institut de Recherche en Informatique de Toulouse. p. 183-192.

Lin Huei-Chi. (2008). Treatment of Chinese Orthographical and Lexical Variants with NooJ. In *Proceedings of the 2007 International NooJ Conference*. Newcastle : Cambridge Scholars Publishing. p. 139-148.

Liu Haitao and Wang Lulu. (2007). A Description of Chinese NPs using Head-Driven Phrase Structure Grammar. In *Stefan Müller (ed.), HPSG-2007: The Proceedings of the fourteenth International Conference on Head-Driven Phrase Structure Grammar*. California (Stanford) : CSLI Publications. p. 287-305.

Packard Jerome L. (2006). The Morphology of Chinese: A Linguistic and Cognitive Approach. New York : Cambridge Universitu Press (2000).

Pollard Carl and Sag Ivan A. (1991). An Integrated Theory of Complement Control. In *Language* Washington : Linguistic Society of America. Vol. 67. No. 1. p. 63-113.

Rizzolo Olivier. (2004). Šatrovački : la construction et l'exploitation d'un corpus de verlan serbo-croate. In *Corpus*. No. 3. Nice. p. 261-309.

Silberztein Max. (1993). Les groupes nominaux productifs et les noms composés lexicalisés. In *Linguisticæ Investigationes*. Vol. 17. No. 2. Amsterdam : John Benjamins Publishing Company. p. 405-426.

Silberztein Max. (2003). NooJ Manual. Downloadable from www.NooJ4nlp.net.

Silberztein Max. (2004). NooJ : an oriented object approach. In *INTEX pour la linguistique et le traitement automatique des langues*. Les Cahiers de la Maison des Sciences de l'Homme Ledoux. Presses Universitaires de Franche-Comté, Besançon.

Silberztein Max. (2007). An Alternative Approach to Tagging. Invited Paper In *Proceedings of 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007: Natrual Language Processing and Information System*. Coll. « LNCS series (4592) ». Berlin / Heidelberg : Springer-Verlag. p. 1-11.

Silberztein Max. (2008). Complex Annotations with NooJ. In *Proceedings of the 2007 International NooJ Conference*. Newcastle : Cambridge Scholars Publishing. p. 214-227.

Sun Chaofen. (2006). Chinese : A Linguistic Introduction. New York : Cambridge University Press.

Turing Alan Mathison. (1950). Computing machinery and intelligence. In *Mind*. Vol. 59. No. 236. New York. p. 433-460.

Yang-Drocoucrt Zhitang. (2007). Parlons chinois. Paris : L'Harmattan.

Zhou Qiang and Yu Shiwen. (1994). Blending Segmentation with Tagging in Chinese Language Corpus Processing. In *COLING-94 : Proceedings of the Fifteenth Conference on*

Computational Linguistics. 2 vols. New Jersey (Morristwon) : Association for Computational Linguistics. p. 1274-1278 (Vol. 2).

Zhou Qiang and Yu Shiwen. (1997). Annotating the Contemporary Chinese Corpus. In *International Journal of Corpus Linguistics*. 2 vols. Amsterdam : John Benjamins Publishing Company. p. 239-258 (Vol. 2).

Références Internet

Běijing dàxué xìnxī / xí kēxué jìshù xuéyuàn 北京大學信息科學技術學院 'School of electronic engineering and computer science, Pekin University' : http://eecs.pku.edu.cn/eecswww/xszx/jsyy.html

Hā'ěrbīn gōngyè dàxué xìnxī / xí jiǎnsuǒ yánjiū / jiù shì 哈尔滨工业大学信息检索研究室 'HIT Center for Information Retrieval' : http://ir.hit.edu.cn

The Lancaster Corpus of Mandarin Chinese (LCMC) : http://www.lancs.ac.uk/fass/projects/corpus/LCMC/

The Penn Treebank Project : http://www.cis.upenn.edu/~treebank/

Unicode Consortium : www.Unicode.org

Yǔyán wénzì wǎng 語言文字網 'Site Langue et lexique' : http://www.yywzw.com/index.aspx

Zhōngguó kēxuéyuàn jìsuàn jìshù yánjiū / jiù suǒ — Zìrán yǔyán chǔlǐ yánjiū / jiù zǔ 中国科学院计算技术研究所—自然语言处理研究组 'Natural Language Processing' : http://mtgroup.ict.ac.cn/index.php

Zhōngwén zìrán yǔyán chǔlǐ kāifàng píngtái 中文自然语言处理开放平台 'CNLP Platform' : http://www.nlp.org.cn/

Zhōngyāng yánjiū / jiù yuàn — Xiàndài hànyǔ biāojì yǔliàokù 中央研究院—現代漢語標記語料庫 'Academia Sinica Balanced Corpus of Modern Chines' : http://dbo.sinica.edu.tw/SinicaCorpus/

Titre : Un module NooJ pour le traitement automatique du chinois : formalisation du vocabulaire et des têtes des groupes nominaux.

Cette étude présente le développement du module d'analyse automatique du chinois qui permet de reconnaître dans les textes les unités lexicales en chinois moderne puis les groupes nominaux noyaux. Pour atteindre ces deux objectifs principaux, nous devions résoudre les problèmes suivants :

1) identifier les unités lexicales en chinois moderne ;
2) déterminer leurs catégories ;
3) décrire la structure de syntaxe locale et des groupes nominaux noyaux.

C'est ainsi que nous avons été amenée à constituer d'abord un corpus regroupant des textes littéraires et journalistiques publiés au XX^{e} siècle. Ces textes sont écrits en chinois moderne avec des caractères traditionnels. Grâce à ces données textuelles, nous avons pu recueillir des informations linguistiques telles qu'unités lexicales, structures syntagmatiques ou règles grammaticales. Ensuite, nous avons construit des dictionnaires électroniques dans lesquels chaque unité lexicale est représentée par une entrée, à laquelle sont associées des informations linguistiques telles que catégories lexicales, classes de distribution sémantique ou descriptions formelles de certaines formes lexicales. À ce stade, nous avons cherché à identifier les unités lexicales du lexique chinois et leurs catégories en les recensant. Grâce à cette liste, l'analyseur lexical peut traiter des unités lexicales de différents types, en bloc, sans les découper en composants. Ainsi, on traite les unités lexicales suivantes comme des unités atomiques :

理髮 *lǐfà / fǎ* <arranger-cheveux> 'faire la coiffure'
放假 *fàngjià* <distribuer-vacance> 'être en vacances'
刀子口 *dāozikǒu* <couteau-bouche> 'parole cruelle'
研究員 *yánjiū / jiù yuán* <effectuer des recherches-K> 'chercheur'
翻譯系統 *fānyì xìtǒng* <traduire-système> 'système de traduction'
浪漫主義 *làngmàn zhǔyì* <romantique- -isme> 'romantisme'

Puis, nous avons décrit de manière formelle un certain nombre de syntagmes locaux, ainsi que cinq types de groupes nominaux noyaux. Enfin, nous avons utilisé le module chinois ainsi développé pour étudier l'évolution thématique dans les textes littéraires.

Mots-clefs : Traitement Automatique des Langues Naturelles. Formalisation du chinois. Dictionnaire électronique du chinois. Description syntaxique des groupes nominaux chinois.

Title: A NooJ Module for the Automatic Processing of Chinese: Formalising the Chinese Vocabulary and Noun Phrases.

This study presents the development of a module for the automatic parsing of Chinese that will allow to recognize automatically lexical units in modern Chinese, as well as central Noun Phrases in texts. In order to reach these two principle objectives, we solved the following problems:

1) identify lexical units in modern Chinese;
2) determine their categories;
3) describe certain local syntactic structures as well as the structure of central Noun Phrases.

Firstly we constructed a corpus regrouping literary and journalistic texts published in the XXth century. These texts are written in modern Chinese with traditional characters. Thanks to textual data, we could collect linguistic information such as lexical units, syntagmatic structures or grammatical rules. Then, we constructed several electronic dictionaries in which each entry represents a lexeme, with which is associated linguistic information such as its lexical category, its semantic distributional class or certain formal properties. At this stage, we tried to identify the lexical units of Chinese lexicon and their categories in order to list them. Thanks to this list, an automatic lexical analyzer can process various types of lexical units in bloc, without deconstructing them in components. For instance, the lexical parser processes the following lexical units as atomic units:

理髮 *lǐfà / fǎ* <operate-hair> 'have a haircut'
放假 *fàngjià* <distribute-vacation> 'have vacation'
刀子口 *dāozikǒu* <knife-mouth> 'straight talk'
研究員 *yánjiū / jiù yuán* <research-K> 'researcher'
翻譯系統 *fānyì xìtǒng* <translate-system> 'translation system'
浪漫主義 *làngmàn zhǔyì* <romantic- -ism> 'romanticism'

Then, we described formally certain local syntagms and five types of central Noun Phrases. Finally, we used this Chinese module to study thematic evolution in literary texts.

Keywords: Natural Language Processing. Automatic Processing of Chinese. Electronic Dictionaries for Chinese. Syntactic Description of Chinese Noun Phrases.

Titre : Un module NooJ pour le traitement automatique du chinois : formalisation du vocabulaire et des têtes des groupes nominaux.

Cette étude présente le développement du module d'analyse automatique du chinois qui permet de reconnaître dans les textes les unités lexicales en chinois moderne puis les groupes nominaux noyaux. Pour atteindre ces deux objectifs principaux, nous devions résoudre les problèmes suivants :

1) identifier les unités lexicales en chinois moderne ;
2) déterminer leurs catégories ;
3) décrire la structure de syntaxe locale et des groupes nominaux noyaux.

C'est ainsi que nous avons été amenée à constituer d'abord un corpus regroupant des textes littéraires et journalistiques publiés au XX^{e} siècle. Ces textes sont écrits en chinois moderne avec des caractères traditionnels. Grâce à ces données textuelles, nous avons pu recueillir des informations linguistiques telles qu'unités lexicales, structures syntagmatiques ou règles grammaticales. Ensuite, nous avons construit des dictionnaires électroniques dans lesquels chaque unité lexicale est représentée par une entrée, à laquelle sont associées des informations linguistiques telles que catégories lexicales, classes de distribution sémantique ou descriptions formelles de certaines formes lexicales. À ce stade, nous avons cherché à identifier les unités lexicales du lexique chinois et leurs catégories en les recensant. Grâce à cette liste, l'analyseur lexical peut traiter des unités lexicales de différents types, en bloc, sans les découper en composants. Ainsi, on traite les unités lexicales suivantes comme des unités atomiques :

理髮 *lǐfǎ* <arranger-cheveux> 'faire la coiffure'
放假 *fàngjià* <distribuer-vacance> 'être en vacances'
刀子口 *dāozikǒu* <couteau-bouche> 'parole cruelle'
研究員 *yánjiū / jiù yuán* <effectuer des recherches-K> 'chercheur'
翻譯系統 *fānyì xìtǒng* <traduire-système> 'système de traduction'
浪漫主義 *làngmàn zhǔyì* <romantique- -isme> 'romantisme'

Puis, nous avons décrit de manière formelle un certain nombre de syntagmes locaux, ainsi que cinq types de groupes nominaux noyaux. Enfin, nous avons utilisé le module chinois ainsi développé pour étudier l'évolution thématique dans les textes littéraires.

Mots-clefs : Traitement Automatique des Langues Naturelles. Formalisation du chinois. Dictionnaire électronique du chinois. Description syntaxique des groupes nominaux chinois.

Title: A NooJ Module for the Automatic Processing of Chinese: Formalising the Chinese Vocabulary and Noun Phrases.

This study presents the development of a module for the automatic parsing of Chinese that will allow to recognize automatically lexical units in modern Chinese, as well as central Noun Phrases in texts. In order to reach these two principle objectives, we solved the following problems:

1) identify lexical units in modern Chinese;
2) determine their categories;
3) describe certain local syntactic structures as well as the structure of central Noun Phrases.

Firstly we constructed a corpus regrouping literary and journalistic texts published in the XXth century. These texts are written in modern Chinese with traditional characters. Thanks to textual data, we could collect linguistic information such as lexical units, syntagmatic structures or grammatical rules. Then, we constructed several electronic dictionaries in which each entry represents a lexeme, with which is associated linguistic information such as its lexical category, its semantic distributional class or certain formal properties. At this stage, we tried to identify the lexical units of Chinese lexicon and their categories in order to list them. Thanks to this list, an automatic lexical analyzer can process various types of lexical units in bloc, without deconstructing them in components. For instance, the lexical parser processes the following lexical units as atomic units:

理髮 *lǐfà / fǎ* <operate-hair> ‘have a haircut’
放假 *fàngjià* <distribute-vacation> ‘have vacation’
刀子口 *dāozikǒu* <knife-mouth> ‘straight talk’
研究員 *yánjiū / jiù yuán* <research-K> ‘researcher’
翻譯系統 *fānyì xìtǒng* <translate-system> ‘translation system’
浪漫主義 *làngmàn zhǔyì* <romantic- -ism> ‘romanticism’

Then, we described formally certain local syntagms and five types of central Noun Phrases. Finally, we used this Chinese module to study thematic evolution in literary texts.

Keywords: Natural Language Processing. Automatic Processing of Chinese. Electronic Dictionaries for Chinese. Syntactic Description of Chinese Noun Phrases.